LA MÉMOIRE DES SÉQUENCES FIGÉES : UNE TROISIÈME ARTICULATION OU LA RÉHABILITATION DU CULTUREL DANS LE LINGUISTIQUE

Salah MEJRI

Université de Tunis I, Tunis, Tunisie

D'habitude, quand on parle de mémoire, on pense surtout à des faits de nature abstraite, n'ayant pas forcément d'appui matériel immédiat, c'est-à-dire en l'absence de tout contrat sémiotique établi directement entre un support concret et un ensemble d'évocations plus ou moins nettes. Si nous parlons de contrat sémiotique, c'est parce qu'à l'origine de cette relation, il n'existe pas de consubstantialité entre le support et son contenu : le même support peut être en rapport avec plusieurs contenus différents. Aux antipodes de ce genre de relation se trouve celle qui fonde le signe linguistique simple, lequel fonctionne selon un mode relativement arbitraire, du moins synchroniquement. Tel ne semble pas être le cas des unités lexicales construites, particulièrement les séquences polylexicales, c'est-à-dire les séquences figées.

1. LES PARTICULARITÉS DES SF

Elles sont situées à mi-chemin entre le fonctionnement symbolique fondé sur le principe de la prégnance forte et celui des unités fondamentalement arbitraires. Comparées à ce dernier type de signes, elles se distinguent par des caractéristiques saillantes qui en conditionnent le fonctionnement :

- elles ont un signifiant lourd qui crée une dissymétrie par rapport au signifié et qui modifie en même temps les rôles respectifs des facettes de ce signe linguistique. Comparez, par exemple, *mourir* et *boire le bouillon d'onze heures* ou *casser sa pipe*; *s'enfuir* et *prendre la poudre d'escampette*; etc.;
- elles se trouvent investies d'un fonctionnement qui rappelle en même temps son origine discursive (une sorte de fonction évidentielle) et qui en fait une unité lexicale ayant un sens global propre et un signifié globalisé.

Or ce signifiant globalisé demeure remotivable, quelle que soit la soudure des constituants de la séquence. Pourquoi ?

- parce que sa nature polylexicale ne sert pas de signifiant étanche au sens, c'est-à-dire segmentable seulement en unités de la deuxième articulation (les phonèmes) ou des unités de prononciation (des syllabes);
- parce que sa polylexicalité a pour corollaire systématique une nouvelle fonction : puisque la SF emprunte son signifiant à d'autres unités du lexique et qu'elle prend naissance dans une réalisation discursive, elle porte nécessairement en elle sa propre mémoire :

• une mémoire syntaxique qui se vérifie surtout au niveau des archaïsmes : *Advienne que pourra; Plus fait douceur que violence;* etc.;
• une mémoire lexicale qui conserve plusieurs unités qui ne doivent leur existence actuelle qu'aux expressions qui les véhiculent : *au fur et à mesure;*
• une mémoire sémantique qui participe à la structuration sémantique du lexique telle qu'elle apparaît dans les articles de certains dictionnaires, par exemple.

En d'autres termes, la SF est susceptible de deux décodages : l'un, le plus courant, est fondé sur son fonctionnement en tant que signe global, et l'autre, toujours possible, qui substitue à cette approche synthétique une autre analytique remotivant les constituants aussi bien au niveau de leurs signifiés que de leurs signifiants pris isolément :

• «Ils s'en allaient, têtus, sur les six bottes, continuer une enquête qui *ne tournait pas rond* comme leurs képis» (H. Bazin, *L'huile sur le feu,* Grasset 1954 : 146).
• «Plus tard, il [=Picasso] fit son portrait. Jacob lui lut ses vers et Picasso — qui *parlait français comme un peintre espagnol* — lui déclara qu'il était le seul poète français de l'époque» (*Le Monde* du 13-8-1994 : 11).

2. LES ARTICULATIONS DANS LES SF

Ces emplois sont loin d'être exceptionnels (cf. le langage journalistique, certains textes de romanciers ou poètes modernes et contemporains, etc.) et illustrent la coexistence de deux types de découpage :

- l'un est fondé sur la base du signe global (signifiant et signifié) : c'est le découpage conforme aux SF considérées comme des unités lexicales ayant un signifiant vu dans sa globalité et un signifié global correspondant :

«Il est hors de conteste, répondit M. de Norpois, que la déposition du colonel devenait nécessaire pour peu que le gouvernement pensât qu'il pouvait bien y *avoir* là *anguille sous roche*» (M. Proust, *À la recherche du temps perdu,* II : 240).

- l'autre est fondé sur l'autonomie des constituants avec ou sans respect de l'intégrité du signifiant :
- dans l'exemple de *ne pas tourner en rond,* de H. Bazin, la SF, bien qu'employée intégralement, sa globalité sémantique n'est pas respectée;
- dans le suivant, il est porté atteinte à la solidarité des constituants :

«C'est bien ce que je pensais, tu *caches sous roche* une mauvaise excuse» (R. Queneau, *Le vol d'Icare,* Gallimard 1968 : 94).

Donc unité et pluralité donnent lieu à des segments s'articulant soit à un premier niveau, c'est-à-dire celui du sens global, soit à un second, c'est-à-dire celui qui correspond aux unités lexicales servant de constituants à la SF; ce qui signifie qu'au niveau de la première articulation, on a affaire à une articulation impliquant des unités douées de sens à deux niveaux : une première articulation globale et une première articulation plurielle.

Cette articulation à deux étages provient d'une segmentation objective imposée au décodeur dans les situations de remotivation; ce qui en fait une donnée empiriquement vérifiable.

Remarque : on pourrait postuler un dédoublement au niveau de la deuxième articulation mais cela ne représente, bien qu'il existe dans le cas de la désintégration du signifiant (cf. les exemples de *parler français comme une vache espagnole* et *anguille sous roche*), ni un fait systématique ni une opération impliquant uniquement le signifiant de la séquence(cf. les jeux de mots sur les paronymes). C'est un phénomène qui demeure, somme toute, accidentel et très limité.

André Martinet précise dans *La linguistique synchronique* (1970 : 33) que l'importance de la double articulation ne réside pas seulement dans son caractère très économique, mais aussi dans le fait que la langue, «en confiant le soin de former ses signifiants à des unités sans face signifiée, les phonèmes, elle les protège contre les atteintes du sens». C'est pour cela que tout parallèle dans le fonctionnement entre les deux articulations ne nous semble pas fondé. Dans un cas, le sens est la raison d'être des unités; dans l'autre, l'absence de sens est la condition *sine qua non* du fonctionnement des phonèmes en tant qu'éléments constitutifs des signifiants.

Devant l'importance des manipulations auxquelles le signifiant de la SF pourrait donner lieu, on ne peut pas s'empêcher de relever la systémicité de la double segmentation opérée au niveau de la première articulation :

1- Cette systémicité provient du fait que toute séquence discursive est susceptible de faire l'objet d'une fixité totale ou partielle; les séquences figées ne sont pas nécessairement des syntagmes bien formés. Il arrive dans certains cas que les éléments fixes soient :

- des éléments de nature fonctionnelle : *à la* + nom ou adjectif ethnique; *tout comme,* etc. «un délirant univers *à la* Aldous Huxley où tout est travail, ordre et propreté» (*Le monde diplomatique*, Juin 1992 : 1).
- des phrases entières : *Les carottes sont cuites; le coeur n'y est pas*; etc.
- des formules du genre : *à vos souhaits; marché conclu*; etc.

Certains écrivains n'hésitent pas à agglutiner à l'écrit certaines séquences libres pour en faire une unité, imitant en cela les mécanismes du figement :

> «À l'étranger, il faut toujours montrer un visage serein, *un pays-où-tout-se-passe-bien.* Le linge sale ne se livre pas aux yeux des étrangers» (T.B. Jelloun, *La nuit de l'erreur,* Seuil 1997 : 102).

2- Cette systémicité se traduit aussi par l'impact que le figement a sur toutes les dimensions du système :

- lexicalement, il nourrit régulièrement le lexique d'unités toujours plus expressives (cf. les argots), de séquences de plus en plus nouvelles (cf. le mouvement néologique), et d'unités phraséologiques fixant les dimensions culturelles de la communauté dans la langue (cf. par exemple les parémies);

- syntaxiquement, il dote la langue de nouveaux outils exprimant les diverses relations syntaxiques (cf. les morphèmes discontinus, les locutions prépositionnelles et conjonctives, l'expression de l'intensité, les périphrases verbales, les verbes supports, etc.);

- morphologiquement, il détermine la formation polylexicale d'un très grand nombre d'unités (cf. les locutions verbales, les structures des séquences nominales, etc.);

- sémantiquement, il participe à la structuration de la polysémie des unités lexicales (cf. par exemple, la manière dont sont construits les articles de dictionnaires comme le *G.L.L.F.* où l'organisation des sémèmes se fait sur la base des emplois figés), et fixe des contenus culturels dans le sémantisme lexical.

3- Cette systémicité est fondamentale pour :

- la survie du système puisqu'elle permet, en plus des points évoqués plus haut, d'établir une sorte de pont entre la langue et la parole (ou le discours) faisant de cette dernière un outil de renouvellement et de réaménagement successifs et constants du système : par le biais du discours, le système se réorganise et se réadapte continuellement;

- la redynamisation du système : en plus des dimensions syntaxiques et sémantiques déjà évoquées, nous rappelons l'activité terminologique qui met en relief la fonction référentielle du langage et où les SF jouent un rôle déterminant (cf. ce que dit Benvéniste de la synapsie);

- l'économie générale du système : en plus des avantages certains de la double articulation du langage, elle ajoute un niveau intermédiaire de segmentation faisant d'unités initialement autonomes des formants ayant nécessairement le statut d'éléments phoniques étanches au sens, du moins dans l'interprétation globale de la séquence. C'est dans ce cadre que nous parlons du figement en tant que recyclage de la parole usée. Le système, ne se contentant plus des ressources des deux articulations, produit des SF dès qu'il est mis en usage et qu'il acquiert le dynamisme vital aux langues utilisées. C'est ce qui fait du figement un phénomène propre aux langues naturelles. Même si l'orthographe garde les traces de l'autonomie initiale, dans la conscience linguistique des locuteurs le signifiant lexical est appréhendé d'une manière globale. En d'autres termes, l'analyse des SF, appliquée à une séquence comme *noyer le poisson,* donne lieu à un découpage comme suit :

a / [nwajelɔ̃pwasõ]	Sa global
«entretenir la confusion pour duper ou lasser son adversaire»	Sé global

où il n'y a pas la moindre correspondance entre les constituants et la synthèse sémantique. Ainsi les unités douées de sens abandonnent-elles leur statut morphématique au profit du fonctionnement phonématique de leurs propres constituants; cet abandon n'étant jamais définitif, la réactivation de la segmentation morphématique reste toujours possible comme c'est le cas dans l'exemple suivant :

b / «Et toi, le coupa brusquement l'inspecteur Ali, tu essaies de *noyer le poisson* alors que l'eau est son élément.» (D. Chraïbi, *L'inspecteur Ali à Trinity college,* 1996 : 90)

où le premier découpage reste valable dans une lecture qui se limiterait à la première proposition. Dès qu'on entame la proposition adversative avec ce qu'elle contient comme reprise sémique *(noyer, poisson / eau)* et renvoi anaphorique au constituant nominal de la séquence verbale (noyer *le poisson /son* élément), s'opère une nouvelle construction sémantique qui s'ajoute à la synthèse sémantique et où les constituants de départ reprennent leur autonomie grâce à la segmentation supplémentaire évoquée plus haut. L'aspect contradictoire de cette expression dégagé à la suite de ce commentaire autonymique, ne serait pas possible sans le repli mémoriel inscrit dans la structure même de la séquence. Pour schématiser, on pourrait dire que les SF, quels qu'en soient la structure et l'emploi dans le discours, sont susceptibles d'une analyse traduisant à la fois le dédoublement et le parallélisme entre les types de Sa et les deux types de Sé correspondants :

signifiant global	et	signifiant pluriel (polylexical)
signifié de synthèse		signifié analytique

Nous disons bien dédoublement non polysémie parce que nous croyons que le traitement des SF en termes de polysémie ne tiendrait pas compte de sa dimension polylexicale et l'assimilerait à une unité unilexicale. La polysémie, croyons-nous savoir, exige un cadre formel restreint (base, affixe, catégorie syntaxique...) et une filiation dans ce cadre, même s'il y a variation du comportement syntaxique en rapport avec le changement sémantique. Michel Bréal a bien perçu cette particularité dans son *Essai de sémantique* (1897) quand il a établi un lien direct entre la polysémie et la notion de mot. L'invariance du sigifiant limité au cadre du mot est ce qui favorise la superposition des sémèmes entre lesquels on peut dégager des relations sémantiques diverses (cf., par ex., le traitement de la polysémie dans R. Martin, *Pour une logique du sens,* 1992).

Appliquer l'approche polysémique à la SF serait très réducteur et fausserait dès le départ les données de l'analyse :

a. Le premier problème serait celui du rattachement de la polysémie à un support signifiant précis. Puisque la SF est par définition polylexicale, rattacher ce sens à l'un de ses constituants nous conduirait à faire abstraction des autres éléments alors qu'ils participent à la synthèse sémantique de la séquence de la même manière (cf. les analyses des locutions verbales de Hervé Curat 1982 et 1984 et de A.M. Schmid 1989 et 1991).

b. Le deuxième problème serait celui de la nature de la segmentation du signifiant de la séquence : l'invariance du signifiant du mot simple repose clairement sur une segmentation conforme aux deux articulations courantes. Traiter la SF de la même manière conduirait à faire abstraction de l'articulation supplémentaire décrite plus haut.

c. Le troisième problème serait celui du rapport existant entre les unités de départ et la nouvelle signification qui est fondé dans la SF, contrairement à ce qui se passe dans les unités simples où il y a consubstantialité entre le Sa et les Sés anciens ou nouveaux, sur une sorte de dédoublement : les unités de départ prêtent leurs signifiants à la séquence sans

que cette opération les prive de recouvrer leur fonctionnement libre. Elles ont donc, dans le cadre de la SF, une double attache, celle qui les relie au signifiant global dans lequel elles perdent momentanément leur identité morphématique (deuxième articulation) et celle qui assure la réactivation de l'identité temporairement perdue (l'articulation supplémentaire de la première articulation). Ne pas tenir compte de la particularité de ces liens, c'est trahir le fonctionnement réel de ce type de signes linguistiques qui participent à la fois du mot (ce qui fait leur unité) et du syntagme ou de la phrase (ce qui leur donne leur identité). Rappelons que si on parle de consubtantialité dans le cas de l'unité simple, c'est parce que la relation entre la forme et le contenu est fondé sur des liens directs; ce qui n'est pas le cas de la SF où il y a nécessairement la médiation des rapports établis entre les constituants (cf. la notion de catène chez H. Frei 1962).

Tout cela nous conduit à la conclusion que la polylexicalité est à la SF ce que la polysémie est à l'unité simple. Se rattache à ce constat un certain nombre de faits :

1- La polylexicalité fournit à la SF une prégnance symbolique plus grande que celle du signe simple ou du signe construit obtenu par dérivation :

Ex : *chien* [ʃjɛ̃] «Mammifère domestique...»
chiot [ʃjo] «Jeune chien»
terre-neuve [tɛRnoev] «Gros chien à tête large..., originaire de...»

Si la prégnance a un degré minimal dans les mots simples, elle traduit la régularité de la langue avec les séries affixées. Mais avec les SF, le signifiant polylexical, en plus du contrat sémiotique primaire qui l'associe au signifié global, se trouve chargé d'une marque symbolique beaucoup plus grande, provenant d'une forte prégnance laissée par la séquence libre employée comme signifiant.

2- C'est cet abandon des constituants libres de leur fonction sémiotique primaire, selon laquelle ils fonctionnent comme des signes ayant des signifiants et des signifiés propres, qui crée une rupture dans le contrat sémiotique, rupture à laquelle s'ajoute un nouveau contrat scellant la nouvelle relation Sa / Sé. La rupture explique le retrait des valeurs initiales des constituants; l'ajout de la nouvelle signification fonde l'épaisseur idiomatique.

3- À cette fonction symbolique on peut ajouter une fonction étymologique assumée par le Sa de la SF. Parmi toutes les unités lexicales, les SF sont les unités qui déclinent immédiatement leur origine. Leurs signifiants indiquent directement leur nature discursive, leur construction syntaxique et les items lexicaux qui leur servent d'habillement.

4- L'aboutissement logique de l'accumulation de telles particularités, c'est le renforcement de la fonction mémorielle qui prend appui sur tout le parcours connu par la SF, de la formation jusqu'à la lexicalisation la plus totale. Y interviennent évidemment tous les faits linguistiques pertinents : la structuration sémantique, les mécanismes référentiels, l'intégration des éléments pragmatiques dans le sens de la SF, les contraintes d'emploi, auxquels on peut ajouter toutes les indications extralinguistiques qui s'y rattachent.

Un exemple : *mettre (avoir) la puce à l'oreille* qui a servi de titre au recueil de Duneton (Stock 1978).

C'est une séquence qui, comparée à son équivalent «éveiller l'attention, la méfiance», sa paraphrase lexicographique, se distingue par cette épaisseur due à la surcharge mémorielle. Se profilent derrière ce signifié global :

- un signifiant qui, malgré sa polylexicalité, fonctionne globalement;
- une dénomination oblique par le biais de laquelle des unités abandonnent leur fonction référentielle de départ (il n'est question dans cette séquence ni de *puce* ni d'*oreille*);
- un mécanisme métaphorique qui justifie le rapprochement entre le sens propre et le sens figuré, chargé d'allusions érotiques (cf. Duneton, 58-63, A. Rey et Chantreau, 992-993 et le *G.L.L.F.*), puis une extension de sens qui lui donne une grande généralisation d'emploi;
- un transfert de domaine qui fait qu'on dénomme une réalité au moyen des termes d'une autre réalité (insecte, partie du corps (soucis,...)).

3. LE CULTUREL

Décrire les mécanismes sémantiques en action dans de telles séquences est certes très éclairant mais ne rend pas directement compte de la «langue comme réceptables de croyances communes [...] révél[ant] beaucoup sur l'imaginaire collectif». «Le lexique [ne] porte-[t-il pas] en lui la marque de croyances profondément enracinées ?» (R. Martin, 1987 : 9).

Cette dimension culturelle est assumée par la fonction mémorielle des unités lexicales, qu'elles soient simples ou complexes (cf. les emprunts, les catachrèses, etc.).

Nous pensons que cette fonction du signe linguistique, jusque-là marginalisée, est très importante. Sans être propre aux SF, — les emprunts, les métaphores et les métonymies lexicalisées, les diverses formes de troncation, les dérivés sont autant de formes concernées par ce phénomène —, cette fonction mémorielle est dans les unités polylexicales à la fois systématique, systémique et évidente :

- systématique, parce qu'aucune SF ne peut exister sans trace de mémoire; son signifiant étant par définition chargé de cette fonction. Sa polylexicalité, de par son origine discursive, témoigne du parcours conceptuel suivi par l'unité polylexicale de la séquence libre avant sa soudure en un signifiant global jusqu'à la globalisation sémantique finale (cf. Mejri 1997)
- systémique, en ce sens qu'elle implique toutes les dimensions du système linguistique : toute séquence porte en elle les traces du fonctionnement du système lors de sa formation (cf. les archaïsmes, par ex.), et qu'elle ne se limite pas au lexique; elle implique aussi la syntaxe (cf. la formation des outils syntaxiques : déterminants complexes, locutions conjonctives et prépositives, adverbes...) et la sémantique (les problèmes relatifs à la globalisation sémantique, à la conceptualisation, aux mécanismes tropiques, etc.)
- évidente, parce qu'elle relève, en synchronie, de la compétence des locuteurs : le signifiant de la SF, en tant que donnée de base, rappelle par certains de ses aspects le signifiant onomatopéique; d'où toutes les actualisations (remotivations) possibles prenant

appui sur le seul signifiant (cf. l'ex. de *noyer le poisson* déjà cité), sans le recours à la moindre donnée étymologique.

Le statut d'articulation supplémentaire permet de rendre compte de tous les transferts qui finissent par se fixer dans les SF qui restent relativement peu étudiés bien qu'ils posent énormément de problèmes dans des domaines aussi variés que l'enseignement, la traduction ou le traitement automatique du langage.

Nous avons essayé de montrer ailleurs (Mejri 1977) que la différence entre les langues réside entre autres dans le choix des domaines de transfert, c'est-à-dire dans le choix des signifiants propres à des domaines comme les parties du corps, les animaux ou des activités variées servant à la dénomination d'autres domaines comme les fruits, les légumes, les pierres précieuses, les outils, les attitudes, etc. Les mécanismes sont identiques d'une langue à une autre mais c'est la nature du domaine-cible qui fait la différence.

Le mot *doigt* par exemple entre dans des expressions françaises renvoyant à l'action (*ne rien faire de ses doigts, ne pas bouger le petit doigt*), au nombre (*compter sur le bout des doigts*), à la perfection (*connaître sur le bout du doigt*), etc. Dans le dialectal tunisien, il est plutôt question d'inégalité ضوابعك مش قد قد [swabʕikmuʃqadqad] litt. «Tes doigts ne sont pas égaux ou de même niveau», de vol يدور فى ضوابعو [jdawarfiswabʕu:] litt. «tourner les doigts», d'attitude كان ضبعك عسل ماتلحسوش الكل [kansubʕikʕsalma:talahsuʃilkul] litt. «Si ton doigt est de miel, ne le lèche pas entièrement», etc.

Cette charge culturelle trouve aussi son expression dans une sorte de marques positives et négatives dont les constituants sont porteurs et dont la combinaison fournit la charge sémantique finale de la séquence.Cela donne aux expressions une structure qui fonctionne selon une logique élémentaire naturelle, traduisant la vision que la communauté culturelle a des choses.

Exemples :

perdre la vie → «mourir»
- + - ⇨ Sans mécanismes
trouver la mort → «mourir» tropiques
+ - -
Casser sa pipe → «mourir»
- + - ⇨ Avec mécanismes
Manger les mauves par les racines → «mourir»
+ + - - tropiques

Le même mécanisme est en oeuvre dans le dialectal tunisien; ce qui varie, c'est le choix des domaines.

On peut dire autant des outils syntaxiques (cf. la description des relations de lieu en français, Vandeloise 1986 et Mejri 1997), des énoncés proverbiaux (Mejri 1997), etc.

Ainsi le culturel n'est-il plus considéré comme une surcharge qui se greffe sur le langage; il est, au contraire, partie intégrante des mécanismes du système. Avec l'usage de la langue, il se dépose régulièrement et crée avec le temps des fixations fondées sur des mécanismes de prototypie et de transferts multiples tropiques ou autres. C'est la raison pour laquelle on a toujours associé langue et identité culturelle, langue et intelligence : en réalité, tout passe par les filtres décrits plus haut.

4. CONCLUSION

Les SF semblent être un exemple qui illustre très bien cette réhabilitation du culturel en en faisant une articulation dont l'étude permettrait entre autres :

- de préciser davantage cette notion de nouvelle segmentation en en décrivant d'une manière plus fine les mécanismes profonds;
- de dégager dans le cadre de la même langue son impact sur la structuration hyperonymique du lexique en distinguant les domaines structurants des domaines structurés;
- d'exploiter toutes ces données dans des études interlinguistiques pour établir les correspondances possibles entre les langues et pour en tirer profit dans la traduction;
- de prévoir dans le traitement informatique cette segmentation supplémentaire qui, moyennant certains critères syntaxiques et sémantiques, peut aider à faciliter les opérations de reconnaissance.

RÉFÉRENCES

CURAT, H. (1982) : *La locution verbale en français moderne,* Québec, Presses de l'Université Laval.

GROSS, G. (1996) : *Les expressions en français; noms composés et autres locutions,* Ophrys, 162 p.

MEJRI, S. (1994) : «Séquences figées et expression de l'intensité», *Cahiers de lexicologie,* n° 65, 1994-2, pp. 111-122.

MEJRI, S. (1996) : «Binarisme, dualité et séquences figées», *Les formes du sens, Mélanges Robert Martin,* Duculot, pp. 249-256.

MEJRI, S. (1997a) : «Défigement et jeux de mots», *Études linguistiques,* vol. 3, *Mélanges Abdelkader Méhiri*, pp. 75-92.

MEJRI, S. (1997b) : *Le figement lexical. Descriptions linguistiques et structuration sémantique,* Publications de la faculté des lettres La Manouba, Tunisie.

OPPOSÉS DE LANGUE, OPPOSÉS DE DISCOURS : LES LAPSUS ANTONYMIQUES

Pierre J. L. ARNAUD

CRTT, Université Lumière Lyon-2, Lyon, France

La mémoire des mots, ce peut être la manière dont ceux-ci sont représentés dans notre esprit de manière à être compris ou bien à être disponibles en production. La mémoire des mots, ce peut aussi être la trace que chaque mot porte des relations qu'il entretient avec les autres mots de la langue. Les deux mémoires sont impliquées entre autres dans un phénomène banal en production, le lapsus, dont une sous-catégorie comprend le lapsus antonymique. Dire le contraire de ce qu'on voulait dire n'a rien d'exceptionnel, et il arrive qu'on se soit simplement mal exprimé, mais dans la plupart des cas, il s'agit d'un lapsus dans lequel un mot-cible est remplacé par son antonyme. Dans ce qui suit, on examine les erreurs antonymiques présentes dans un corpus de lapsus français, dans la perspective d'en tirer des conclusions sur la nature de l'antonymie et sur les mécanismes de production que ces erreurs mettent en évidence.

L'ANTONYMIE EN LINGUISTIQUE

Il existe une abondante littérature lexicologique-sémantique sur l'antonymie, parmi laquelle les chapitres correspondants des traités de Leech (1974), Lyons (1977) et Cruse (1986) font figure de classiques. On se limitera dans ce qui suit à ce qui est nécessaire pour l'analyse des données.

En s'inspirant de Cruse, mais en modifiant sa terminologie (Cruse utilise le terme *opposites* comme hyperonyme, alors qu'on utilise ici *antonymes*, terme qu'il réserve à la catégorie ici dénommée *polaires*), on peut distinguer trois catégories d'antonymes : 1) les polaires, qui sont gradables (chaud/froid), 2) les complémentaires (mort/vivant), et 3) les inverses (maître/élève). L'antonymie n'est pas une catégorie simple, et certaines paires antonymiques semblent plus fortement opposées que d'autres. Ainsi, selon Cruse, *work/play* et *town/country* sont des antonymes relativement faibles en raison du manque d'une échelle unidimensionnelle claire qui sous-tendrait leur opposition. De la même façon, deux antonymes sont d'autant meilleurs que l'opposition épuise une plus grande proportion de leur sens : *giant/dwarf* sont moins bons que *large/small*. *Tea/coffee* et *gas/electricity* sont ressentis comme des antonymes par certains locuteurs, mais dans des contextes où ils représentent un choix binaire. L'antonymie peut donc être considérée comme une relation prototypique, ce qui s'accorde bien avec les propositions de Chaffin (1992) pour qui les relations sémantiques ne doivent pas être considérées comme des primitifs, mais sont en réalité semblables aux concepts en ce qu'on peut y démontrer

l'existence de gradients de typicité, que des comparaisons de similarité sont possibles, et que des phénomènes d'instanciation par le contexte y apparaissent.

En attaquant les problèmes sous l'angle de la linguistique de corpus, Mettinger (1994) a analysé le comportement en discours des antonymes dans un ensemble de romans en anglais. Selon cet auteur, la recherche linguistique antérieure sur les antonymes était fortement théorique et fondée sur une approche déductive et logique tendant à négliger les données réelles. Mettinger distingue quatre catégories d'antonymie : l'adversativité, l'opposition sémantique systémique, l'opposition sémantique non systémique, et le contraste. L'adversativité dépend du monde réel ou mental. L'opposition sémantique systémique, qui repose sur la structure sémantique du lexique, est stable et relativement indépendante du contexte et on peut dire d'une paire d'occurrences qu'il s'agit d'opposés systémiques si on peut trouver un archilexème commun. L'opposition sémantique non systémique correspond à des paires de mots qui ne sont opposés que par l'intermédiaire des connaissances encyclopédiques. L'un des exemples de Mettinger est *life/literature*, qui sont opposés dans un passage de critique littéraire. Les *contrastes* apparaissent souvent dans un nombre réduit de cadres syntagmatiques tels que *X ou Y, X et Y.*

La linguistique de corpus est également représentée par Fellbaum (1995), qui a examiné les cooccurrences d'antonymes nominaux et verbaux dans le corpus de Brown (Kucera et Francis 1967), comme Justeson et Katz (1991) l'avaient fait pour les adjectifs. Fellbaum montre que les antonymes verbaux et nominaux sont présents dans le corpus en cooccurrence intra-phrase de façon hautement significative. Il apparaît également que des concepts opposés peuvent être coprésents dans des phrases sous la forme de mots de classes différentes. Comme chez Mettinger, un certain nombre de cadres syntaxiques sont mis en évidence comme, pour le nom, *X as well as Y; from X to Y; now X, now Y.*

L'ANTONYMIE EN PSYCHOLINGUISTIQUE

Une constatation classique (cf. Clark, 1970; Hörmann, 1972) est que dans les expériences d'associations verbales, un adjectif qui a un antonyme direct (*grand/petit*) fait toujours apparaître celui-ci comme réponse primaire, et que ce genre d'association est celui pour lequel les fréquences des réponses primaires sont les plus élevées. Il existerait donc un très fort lien associatif entre antonymes. Dans une représentation décompositionnelle du sens, il n'y aurait qu'un trait de différence entre deux antonymes, qu'une simple inversion de polarité séparerait (on sait qu'en réalité les antonymes prototypiques sont gradables, ce qui pose des problèmes pour une représentation à traits binaires). Dans une perspective associationniste classique, ce lien mental proviendrait des rencontres d'occurrences antonymiques qui seraient souvent voisines en discours (Charles et Miller, 1989), et les liens associatifs existent entre unités lexicales, et non pas entre concepts (Gross et Miller, 1990). Selon ces auteurs, le lexique adjectival, contrairement au lexique des noms, serait structuré par l'antonymie, par des liens directs dans le cas des antonymes directs, et par l'intermédiaire des polaires dans le cas des antonymes indirects (*grand/minuscule*). Charles, Reed et Derryberry (1994) ont obtenu des résultats expérimentaux à l'aide de tâches chronométrées de décision de similarité qu'ils interprètent comme démontrant que l'antonymie directe peut reposer sur des liens associatifs, alors que l'antonymie indirecte et la synonymie reposent essentiellement sur des relations conceptuelles.

Si, on l'a vu plus haut, des recherches ont montré qu'il est exact que les antonymes cooccurrent souvent (Mettinger, 1994; Fellbaum, 1995), et si l'on peut constater que de nombreuses phraséologies comportent des antonymes (*ça ne me fait ni chaud ni froid, sans queue ni tête, petits et grands, été comme hiver, c'est le jour et la nuit*), l'idée que les liens associatifs proviendraient de la cooccurrence en discours a toutefois été critiquée, car on peut suggérer que les antonymes cooccurrent précisément parce qu'ils sont étroitement liés sémantiquement et ainsi inverser l'argument (Levelt, 1989; Murphy et Andrew, 1993).

MODÈLES DE LA LEXICALISATION ET LIEU DE LA SUBSTITUTION

La plupart des modèles psycholinguistiques actuels de la lexicalisation, c'est-à-dire de la production de mots en discours, comportent trois niveaux de représentation des mots, un niveau conceptuel, un niveau d'unités dites *lemmes* qui sont des unités linguistiques répondant à des spécifications conceptuelles et contenant ou renvoyant aux caractéristiques grammaticales des mots, et enfin un niveau de *lexèmes*, terme qui désigne dans ce cadre la représentation phonologique de la forme des mots. Les différences essentielles entre modèles résident dans la présence ou non de rétroaction. Dans les modèles à stades distincts (cf. Levelt *et al.*, 1991), un lemme est activé depuis le niveau conceptuel, et ensuite ce lemme envoie de l'activation vers la couche lexémique. Les lapsus par substitution proviennent d'une compétition entre unités de même niveau. Dans les modèles interactifs ou connexionnistes (cf. Stemberger, 1985; Dell, 1986; Harley, 1993), un ensemble de traits conceptuels envoie de l'activation à des degrés variables vers plusieurs lemmes qui correspondent à tous ou à certains des traits conceptuels; les lemmes activés envoient à leur tour de l'activation à des degrés variables vers des lexèmes en fonction des phonèmes que ces derniers ont en commun. Il se produit une réverbération de l'activation entre niveaux, et les lapsus se produisent lorsqu'une unité-erreur atteint par sommation un niveau d'activation supérieur à l'unité-cible et suffisant pour remporter la compétition.

Les modèles à deux stades séparés ne parviennent pas à rendre compte de l'existence statistiquement significative de lapsus mixtes sémantiques-formels (Harley, 1984), que les modèles interactifs expliquent parfaitement, et c'est à ces derniers qu'on s'intéressera dans ce qui suit. Par ailleurs, les lapsus par substitutions ne peuvent suffire à eux seuls à prouver la nécessité d'une couche lemmatique, malgré des indices qui tendent dans cette direction comme la concordance de genre entre cible et erreur dans des langues comme le français et l'allemand. Si l'on s'en tient à la seule lexicalisation, ce sont en réalité les lapsus d'une catégorie marginale, celle des mélanges sémantiques, qui démontrent la nécessité de lemmes dans un modèle de la production (cf. Arnaud, en préparation). La réflexion sur la nature des lemmes n'a cependant guère avancé depuis leur introduction par Kempen et Huijbers en 1983.

LES DONNÉES

La collection de lapsus français, dont les lapsus antonymiques étudiés ici constituent un sous-ensemble, comporte 2 400 erreurs rassemblées extensivement par l'auteur (pour la méthodologie, cf. Arnaud, 1997). Sur ce total, 1 087 lapsus constituent des substitutions de mots entiers, dont 332 sont des substitutions sémantiques pures ou mixtes, c'est-à-dire des substitutions où mot-cible et mot-erreur présentent une ressemblance de sens subjective. Une substitution mixte manifeste une ressemblance sur

deux ou exceptionnellement trois aspects, sémantique et formel par exemple. Les substitutions mixtes ne seront pas distinguées des pures dans ce qui suit.

Certains lapsus antonymiques ne sont détectables que par la contradiction logique ou encyclopédique qu'ils entraînent :

(1) (le commentateur explique que, cette année, le Président de la République reste debout pendant tout le défilé du 14 juillet) l'année passée il s'asseyait au passage des drapeaux et des étendards

(2) a indiqué que la frontière orientale de la Pologne était tout indiquée — c'était la ligne Oder-Neisse

(3) ce qui est marrant, c'est le scoop Mitterrand est de droite (l'énoncé est ironique et veut se moquer d'un "non-scoop", la conversation tournant autour d'un ouvrage qui "découvre" que Mitterrand était un homme de gauche; la cible était donc *gauche*)

Il est fatal qu'en situation de collecte extensive, certains de ces lapsus les moins évidents passent inaperçus, et les statistiques sur les antonymes devront être considérées comme plus indicatives que définitives (même si, dans l'ensemble, on peut considérer les données sur les substitutions de mots comme fiables, cf. Arnaud, 1997).

Parmi les lapsus sémantiques, 91 concernent des substitutions antonymiques simples, dans lesquelles erreur et cible sont des opposés monolexémiques; ce nombre n'inclut pas 22 erreurs par antonymes de discours (voir plus bas). Le pourcentage de lapsus antonymiques parmi les substitutions sémantiques est de 27,41 %, et est comparable à celui publié par Hotopf (1982) qui était de 31,25 % sur ses données anglaises.

De nombreux lapsus impliquent des antonymes lexicaux des trois catégories classiques mentionnées plus haut.

(4) objectivement, j'ai tort|j'ai raison

(5) i faut que je range la vaisselle sale|euh, propre

(6) c'est toi qui as fermé|ouvert hier en rentrant ?

(7) i fait pas froid|i fait pas chaud dans la maison

(8) et tes cactus, tu les rentres l'été ?

(9) comment ils parviennent à se mettre en position haute|pardon, en position basse

(10) nous sommes dans une situation telle que si nous n'avons pas une hausse rapide des taux d'intérêt|une baisse rapide des taux d'intérêt,

(11) si vous saviez ce que ça me compliquerait|ce que ça me simplifierait l'existence

D'autres substitutions impliquent des antonymes sans doute moins prototypiques, mais qui, *dans le contexte*, sont parfaitement perceptibles comme tels.

(12) j'ai quelques renseignements à vous donner (< demander)

(13) c'est vrai que quand on publie trop les tirages baissent|augmentent, pardon

(14) ce n'est pas parce que je suis le gendre|que je suis le beau-père, pardon, de M. Pierre Botton,

(15) un des seuls musiciens /fr/létrangers invités aux Concerts Spirituels à Paris

(16) fixant la liste des industries qui seront nationalisées|privatisées, pardon

Un autre sous-ensemble comprend principalement des suites N+Adj avec substitution de l'adjectif. Ce qui distingue ces erreurs des précédentes est le fait que les syntagmes impliqués sont probablement préfabriqués (*chunked*) et qu'on peut voir là une substitution de syntagmes et non d'adjectifs seuls. De nombreux cas impliquent notamment *dernier/prochain* :

(17) c'est pour ça que j'ai pris de la très bonne viande la prochaine fois|non, la dernière fois

(18) les principes qu'on avait mis au point l'année prochaine

On peut regrouper dans une autre sous-catégorie des substitutions pour lesquelles la paire cible/erreur ne serait probablement pas perçue comme antonymique par des informateurs (la phrase-test *X est l'opposé de Y* apparaît comme anormale ou douteuse, par exemple). Il s'agit des désignations de référents qui sont en contraste dans le discours, voire dans la situation d'énonciation (catégories *contraste* et *adversativité* de Mettinger, 1994).

(19) je sais que les clavecinistes et les puristes vont s'étonner qu'on puisse jouer Rameau au clavecin

L'information commentée dans ce passage n'aurait rien d'étonnant, mais la discussion porte sur le fait que la personne interviewée est un pianiste qui aime jouer Rameau, et la cible était *piano*. *Piano* et *clavecin* ne sont pas des antonymes, mais des cohyponymes, et c'est la situation d'énonciation qui les met clairement en contraste. De la même façon, *supérieur* et *secondaire* s'opposent dans une réunion d'universitaires où l'on parle d'horaires :

(20) mon mari est enseignant dans le supérieur|dans le secondaire, je veux dire

Garrett (1992) rapporte des cas semblables, où, par exemple, *deaf* se substitue à *blind* ou *syntactic* à *semantic*. Les membres de ces paires ne sont pas antonymiques, mais l'intention communicative était "sourd seulement, pas aveugle", "seulement syntaxique, donc pas sémantique".

Ce genre de lapsus est particulièrement fréquent entre adjectifs de nationalité.

(21) on n'aura plus une Université de Lausanne et une Université de Genève, mais une seule université alémanique

Alémanique se substitue à *romane* avec lequel il est en contraste implicite dans le discours, dans lequel il n'a à aucun moment été question explicitement des universités de la Suisse germanophone.

(22) (il est question des noms en Mac- dont il est parfois impossible de dire s'ils sont écossais ou irlandais; un contre-exemple indubitablement irlandais vient d'être cité) oui, ça c'est écossais
(NB il s'agit bien d'un lapsus, que le locuteur a reconnu comme tel)

(23) - Vous allez manifester à Rome. Est-ce que ce n'est pas une immixtion dans les affaires françaises ?
- Vous voulez dire italiennes ?
- Italiennes.

(24) (pendant un épisode de la guerre des pêches entre la France et l'Espagne) le tribunal maritime vient d'imposer une amende très élevée à ce chalutier donc françaislespagnol

(25) il pense en effet que les Amérilque les Allemands ont fait beaucoup de mal (de Henry Morgenthau, secrétaire américain au trésor en 1944, et de ses projets pour l'Allemagne).

CATEGORIES GRAMMATICALES

Le tableau 1 reproduit les données de catégories grammaticales pour les erreurs sémantiques. La différence entre lapsus sémantiques non antonymiques et lapsus antonymiques est très significative (χ^2 = 132,06 — 3 d.d.l. — $p < 0,001$), ce qui confirme la constatation de Garrett (1992) sur la part dominante de l'antonymie dans les substitutions adjectivales anglaises. Ceci peut constituer un argument en faveur de l'idée que la structure sémantique du lexique adjectival pourrait reposer sur l'antonymie (voir plus haut), mais pas permettre de décider si la substitution est d'origine conceptuelle ou linguistique. En outre, le fait que noms et verbes soient concernés, à un degré certes moindre que les adjectifs, mais néanmoins réel, est à rapprocher des données de Fellbaum (1995).

	non-antonymiques		antonymiques	
	n	%	n	%
Noms	202	87,07	18	19,78
Verbes	20	8,62	17	18,68
Adjectifs	8	3,44	42	46,15
Adverbes	2	0,86	14	15,38
total	232	100	91	100

Tableau 1 : Substitutions sémantiques et catégories grammaticales

LIEU DES SUBSTITUTIONS ANTONYMIQUES

Où se produisent les substitutions antonymiques ? Cette question, notons-le, n'est pas la même que : Comment se produisent les substitutions ? En effet, que le mécanisme soit intra ou inter-niveaux, et quelle que soit la cause ultime qui amène une unité-erreur à un degré d'activation supérieur à celui de l'unité-cible, fluctuations aléatoires autour de la valeur de repos, rémanence d'une activation antérieure, activation provenant de causes centrales autres que le message[1] comme une contamination perceptive, voire causes de type freudien inaccessibles à la démonstration scientifique, la question de la forme et de l'organisation des représentations est capitale.

Comme on l'a vu, les antonymes sont fortement liés associativement, et les substitutions d'antonymes "classiques" (erreurs 4 à 11) pourraient provenir de mécanismes associatifs où, lorsqu'une unité reçoit de l'activation, elle en envoie une forte dose à ses associés les plus proches. Si on accepte que les associations existent entre unités linguistiques (voir plus haut) et non entre concepts, les substitutions se produiraient entre lemmes, voire entre lexèmes, et la proximité sémantique erreur/cible ne serait alors au fond qu'un épiphénomène, puisque si le sens a pu jouer un rôle dans l'établissement de l'association, il n'intervient pas dans le déclenchement de l'erreur. Quelles raisons peuvent permettre de postuler une origine associative aux lapsus ? Leur caractère involontaire, irrépressible, rappelle les phénomènes d'amorçage (McNamara, 1992, 1994). On considère souvent que les effets d'amorçage qui apparaissent dans divers paradigmes expérimentaux sont tellement rapides qu'ils ne permettent pas un accès lexical complet, c'est-à-dire jusqu'à la représentation conceptuelle (Le Ny, 1989), et Shelton et Martin (1992) ont obtenu expérimentalement des effets d'amorçage automatique pour des associés, mais pas pour des mots proches sémantiquement mais non associés. Cependant, des expériences récentes d'amorçage par amorces masquées, donc ne pouvant parvenir à la conscience des sujets, avec tâches de décision lexicale et de dénomination (Perea et Gotor, 1997) ont démontré des effets extrêmement rapides non seulement pour des paires de mots associés, mais aussi pour des paires à lien sémantique mais non associées. Il est donc possible que l'amorçage automatique se produise aussi pour des non-associés, et ceci réduit la nécessité de postuler un lieu linguistique, c'est-à-dire infra-conceptuel, spécialement pour les substitutions d'antonymes classiques.

Les substitutions d'antonymes classiques ne sont pas les seules, on l'a vu, et l'examen des autres catégories est indispensable. *Clavecin* et *piano* (erreur 19) et les adjectifs de nationalité (21 à 25), on l'a vu également, sont des cohyponymes, mais sont opposés par le discours. Les circonstances d'énonciation sont telles qu'il y a une forte opposition, même implicite. Rappelons-nous qu'il n'est à aucun moment question de la Suisse alémanique dans le discours précédant l'apparition de l'erreur 21; par ailleurs, le discours précédant l'erreur 23 ne comportait pas d'occurrence de *France* ou de *français*, et il n'est donc pas possible dans ces deux cas d'invoquer un état d'activation rémanente (au moins à court terme) des unités linguistiques correspondantes. L'explication causale pourrait être que l'opposition de deux concepts les met tous deux en un état d'activation élevé et que, quel que soit le mécanisme, compétition intra-niveau ou résonance inter-niveaux entre unités partageant des traits, le concept-erreur a peu de mal à remporter la compétition et donc à être lexicalisé à la place de l'autre.

[1] Ce terme désigne le contenu de communication pré-verbal.

Il existe d'autres cas de lapsus qui font penser à un phénomène général de permutation de contraires. C'est ainsi qu'on rencontre des oppositions d'antonymes morphologiques, dont on peut se demander s'ils sont lexicaux, c'est-à-dire préfabriqués, ou générés au coup par coup, ou encore si leur représentation n'est pas à mi-chemin entre ces deux cas :

(26) pendant ce temps le petit truc i gèleli s'dégèle

On trouve également des cas d'inversion de polarité non plus lexicale, mais grammaticale :

(27) i faut pas être réalisteli faut être réaliste, pardon

(28) J'veux dire, i z'ont horreur des granulés pour hamsters. Ils les détestent paslils les détestent.

(29) j'suis pas sûre qu't'as bien remué la salade; parce que t'as mis beaucoup trop de tempslpas assez de temps, j'veux dire

On ne peut exclure que de tels lapsus aient leur origine dans un module grammatical, donc linguistique, mais y voir la conséquence d'un changement de polarité conceptuel présente l'avantage d'une explication unique pour les lapsus antonymiques, dont on a vu qu'ils constituent une catégorie remarquable par le nombre et la variété.

Que la gestion mentale des oppositions présente souvent des difficultés peut être illustré par le fait suivant, observé lors du colloque où cet article a été présenté. Une participante, qui faisait une communication sur les nominalisations, expliqua d'abord que, historiquement, certains verbes précédaient les noms correspondants. Puis, projetant un transparent, elle en fit un commentaire qui pouvait sembler aller en sens contraire. Quelques instants plus tard, elle s'interrompit dans son discours, pour finir par se demander à voix haute si elle n'avait pas dit le contraire de ce qu'elle avait voulu dire — à ce stade, l'assistance était de toute façon irrémédiablement perdue ! S'il n'y a pas eu là de lapsus par substitution de mots proprement dit, on n'en a pas moins eu affaire à un phénomène très proche de certains des lapsus étudiés ci-dessus.

Ces considérations ne peuvent que rester spéculatives, car, dans la masse de données que nous fournissent les lapsus substitutionnels, très peu constituent en fait des preuves permettant de trancher entre les détails de plusieurs modèles. Une raison en est que les lapsus naturels ne sont pas susceptibles de manipulations expérimentales. Une autre raison est que la nature des lemmes est encore insuffisamment spécifiée dans les modèles, notamment en ce qui concerne leur contenu sémantique. Or, la question de savoir ce qui du sens des mots est linguistique et ce qui est extra- (supra-) linguistique n'est pas sans intérêt...

RÉFÉRENCES

ARNAUD, P. J. L. (1997) : "Les ratés de la dénomination individuelle : typologie des lapsus par substitution de mots", C. Boisson et Ph. Thoiron (dir), *Autour de la dénomination*, Lyon, P.U.L., pp. 307-331.

ARNAUD, P. J. L. (en préparation) : "Target-error resemblance in word-substitution speech errors and the mental lexicon".

CHAFFIN, R. (1992) : "The concept of a semantic relation", A. Lehrer & E. F. Kittay (eds), *Frames, Fields, and Contrasts: New Essays in Semantic and Lexical Organization*, Hillsdale, N.J., Erlbaum, pp. 253-288.

CHARLES, W. G. & G. A. MILLER (1989) : "Contexts of antonymous adjectives", *Applied Psycholinguistics*, 10, pp. 357-375.

CHARLES, W. G., REED, M. A. & D. DERRYBERRY (1994) : "Conceptual and associative processing in antonymy and synonymy", *Applied Psycholinguistics*, 15, pp. 329-354.

CLARK, H. H. (1970) : "Word associations and linguistic theory", J. Lyons (ed.) *New Horizons in Linguistics*, Harmondsworth, Penguin, pp. 271-286.

CRUSE, D. A. (1986) : *Lexical Semantics*, Cambridge, C.U.P.

DELL, G. (1986) : "A spreading-activation theory of retrieval in sentence production", *Psychological Review*, 93, pp. 283-321.

FELLBAUM, C. (1995) : "Co-occurrence and antonymy", *International Journal of Lexicography*, 8, pp. 281-303.

GARRETT, M. F. (1992) : "Lexical retrieval processes: semantic field effects", A. Lehrer & E. F. Kittay (eds), *Frames, Fields, and Contents: New Essays in Lexical and Semantic Organization*, Hillsdale, N.J., Erlbaum, pp. 1-15.

GROSS, D. & K. J. MILLER (1990) : "Adjectives in WordNet", *International Journal of Lexicography*, 3/4, pp. 265-277.

HARLEY, T. A. (1984) : "A critique of top-down independent levels models of speech production: evidence from non-plan-internal speech errors", *Cognitive Science*, 8, pp. 191-219.

HARLEY, T. A. (1993) : "Phonological activation of semantic competitors during lexical access in speech production", *Language and Cognitive Processes*, 8, pp. 291-309.

HÖRMANN, H. (1971) : *Introduction à la psycholinguistique*, Paris, Larousse.

HOTOPF, W. H. N. (1982) : "Semantic similarity in whole-word slips of the tongue", V. A. Fromkin (ed.), *Errors in Linguistic Performance*, New York, Academic Press, pp. 97-109.

JUSTESON, J. S. & S. M. KATZ (1991) : "Co-occurrences of antonymous adjectives and their contexts", *Computational Linguistics*, 17, pp. 1-19.

KEMPEN, G., & P. HUIJBERS (1983) :."The lexicalization process in sentence production and naming: indirect election of words", *Cognition*, 14, pp. 185-209.

KUCERA, H. & W. N. FRANCIS (1967) : *Computational Analysis of Present-day American English*, Providence, Brown U.P.

LEECH, G. (1974) : *Semantics*, Harmondsworth, Penguin.

LE NY, J.-F. (1989) : "Accès au lexique et compréhension du langage : la ligne de démarcation sémantique", *Lexique* (Lille), pp. 65-85.

LEVELT, W. J. M. (1989) : *Speaking: from Intention to Articulation*, Cambridge (Mass.), M.I.T. Press.

LEVELT, W. J. M., SCHRIEFERS, H., VORBERG, D., MEYER, A., PECHMANN, T. & J. HAVINGA (1991) : "Normal and Deviant lexical processing: A reply to Dell and O'Seaghdha", *Psychological Review*, 98, pp. 615-618.

LYONS, J. (1977) : *Semantics, Vol. 1*, Cambridge, C.U.P.

MCNAMARA, T. P. (1992) : "Theories of priming: I. Associative distance and lag", *Journal of Experimental Psychology: Learning, Memory and Cognition*, 18, pp. 1173-1190.

MCNAMARA, T. P. (1994) : "Theories of priming: II. Types of primes", *Journal of Experimental Psychology : Learning, Memory and Cognition*, 20, pp. 507-520.

METTINGER, A. (1994) : *Aspects of Semantic Opposition in English*, Oxford, Clarendon Press.

MURPHY, G. L. & J. M. ANDREW (1993) : "The conceptual basis of antonymy and synonymy in adjectives", *Journal of Memory and Language*, 32, pp. 301-319.

PEREA, M. & A. GOTOR (1997) : "Associative and semantic priming effects occur at very short stimulus-onset asynchronies in lexical decision and naming", *Cognition*, 62, pp. 223-240.

SHELTON, J. R. & R. C. MARTIN (1992) : "How semantic is automatic semantic priming ?", *Journal of Experimental Psychology: Learning, Memory and Cognition*, 18, pp. 1191-1210.

STEMBERGER, J. P. (1985) : "An interactive activation model of language production", A. W. Ellis (ed.), *Progress in the Psychology of Language, Vol. 1*, London, Erlbaum, pp. 143-186.

LES PROVERBES : DES DÉNOMINATIONS D'UN TYPE «TRÈS TRÈS SPÉCIAL»

Georges KLEIBER

Université des sciences humaines, Strasbourg 2 et Scolia, France

La nuit, tous les proverbes sont gris

INTRODUCTION

Un premier constat pour commencer : les proverbes ont actuellement la cote chez les sémanticiens et pragmaticiens. En témoigne une abondante littérature récente[1] qui essaie de cerner la définiton et le fonctionnement sémantico-pragmatique des proverbes. Un second constat, toujours pour commencer : cette littérature présente un désaccord profond sur la manière de les envisager. Deux tendances contraires s'y font jour. Une tendance «optimiste» qui pense qu'il est possible de définir linguistiquement le proverbe et une tendance plutôt «pessimiste» qui renoue en quelque sorte, sans s'y arrêter toutefois — heureusement —, avec le défaitisme définitoire d'un A. Taylor (1931)[2]. Les premiers considèrent que les proverbes constituent une catégorie linguistique suffisamment homogène pour être décrite de façon unitaire, avec des propriétés et des tests de reconnaissance linguistiques spécifiques[3], alors que les seconds[4] trouvent qu'une telle attitude est excessive : les proverbes forment une catégorie plutôt hétérogène[5], à laquelle

---

1 Voir depuis notre tentative de 1989 (reprise dans Kleiber, 1994), les articles d'Anscombre (1989, 1990, 1994 et 1995), Arnaud (1991, 1992), Arnaud et Moon (1993), Somolinos Rodriguez (1993), Franken (1995), Michaux (1995 et 1996), Forest (1996) et Gouvard (1996 et à paraître).

2 «*The definition of a proverb is too difficult to repay the undertaking*», Taylor (1931 : 3) cité par Schulze-Busacker (1984 : 135).

3 Se laissent placer sous cette bannière les travaux de Kleiber (1989), Anscombre (1994), et Arnaud (1991 et 1992).

4 Empressons-nous de souligner (i) que les travaux que nous rangeons dans ce deuxième courant, à savoir essentiellement ceux de Franken (1995), Michaux (1995 et 1996) et de Gouvard (1996 et à paraître), ne se cantonnent nullement dans une position négative stérile, mais présentent des propositions nouvelles stimulantes, notamment celles qui se rattachent à la poéticité du proverbe, et tracent des perspectives destinées à faire avancer positivement les choses et (ii) qu'il ne s'agit évidemment pas d'un jugement de valeur dépréciatif de notre part, bien au contraire.

5 «Mieux vaut se faire à l'idée, comme le suggèrent également Rodegem (1984), Michaux (1995) et Franken (1995), que les énoncés que l'on range sous l'étiquette de *proverbes* sont loin

les traits définitoires et les tests mis en relief par les premiers ne s'appliquent qu'imparfaitement.

Je me propose dans ce travail de reprendre le débat et d'examiner en détails, non pas le problème dans son entier, mais une des pièces litigieuses du dossier, celle qui concerne un des plus importants traits définitoires généralement attribué aux proverbes, celui d'être un jugement collectif et non un jugement individuel. La tradition veut que les proverbes ne soient pas la voix d'un particulier, mais véhiculent l'expression de la «sagesse populaire», *vox populi*, ou encore «sagesse des nations». Les partisans de la première tendance ont essayé, de diverses manières et dans des cadres théoriques différents, d'expliciter ce trait et d'en saisir les manifestations linguistiques au travers d'une batterie de tests identificatoires. Leur application, par les tenants du courant opposé, à des proverbes particuliers diversifiés est à la source d'une remise en cause assez radicale — et inattendue — à la fois du trait lui-même et des critères linguistiques s'y rapportant. L'enjeu, on le voit, est de taille, puisqu'il porte sur un aspect central de la conception et du rôle des proverbes : un proverbe peut-il ou non constituer un jugement individuel ?

Notre examen se fera en trois parties. Nous exposerons tout d'abord la position classique du proverbe-jugement collectif en rappelant et en revisitant les principaux critères formels qui viennent l'étayer. La deuxième partie donnera la parole aux détracteurs et présentera les principaux arguments avancés pour limiter ou invalider l'efficacité des critères définitoires évalués. Dans la troisième enfin, nous prendrons parti pour la première position à partir de l'hypothèse définitoire que nous avons fournie en 1989, à savoir que les proverbes sont des dénominations d'un type «très très spécial». En même temps qu'elle apportera une réponse à la question centrale posée, notre analyse permettra, chemin faisant, de mieux délimiter les proverbes par rapport aux phrases génériques d'une part et par rapport aux expressions idiomatiques phrastiques d'autre part. Elle aura l'avantage, en outre, d'expliquer *in fine* — ce qui n'est pas fait dans les traitements habituels[6] et qui représente donc de ce point de vue-là un progrès certain — l'origine du facteur collectif.

1. LES PROVERBES : DES JUGEMENTS COLLECTIFS

On retrouve à peu près dans toutes les définitions lexicographiques des proverbes qu'il s'agit de vérités ou de jugements qui sont communs à tout un groupe social. Vérités *traditionnelles* ou encore appelées *populaires* (Ollier, 1976), parce qu'elles font partie d'un stock ou «trésor de conseils empiriques accumulés au fil du temps par la sagesse populaire» (Anscombre, 1994 : 99). «*Proverbs in general, souligne Norrick (1985 : 40), are traditional by virtue of their beings items of folklore. In this sense, [...], traditionality amounts to common use in a linguistic community or in one of its lectal groups over a period of time, say more than one generation*».

Deux conséquences linguistiques en découlent :

- celui qui emploie un proverbe n'en est pas l'auteur;

---

de former une classe homogène qui pourrait recevoir une description linguistique homogène, mais constituent en fait un ensemble de *sous-classes*» (Gouvard, 1996 : 54).

6 De nombreux auteurs ont repris le trait de dénomination que nous avons postulé en 1989, sans toujours mesurer à leur juste aune les conséquences qu'il entraîne et les données qu'il permet d'expliquer.

- l'auteur d'un proverbe n'est pas un particulier, mais «quelque chose comme une conscience linguistique collective» (Anscombre, 1994 : 100).

1.1 Le locuteur d'un proverbe n'est pas l'auteur du proverbe

La première mérite d'être explicitée. Généralement, on entend essentiellement par là que le jugement parémique exprimé dans le proverbe ne peut être porté au crédit de celui qui emploie le proverbe. Un locuteur qui dit :

Qui trop embrasse mal étreint

n'est ainsi pas le responsable du contenu exprimé par le proverbe. Autrement dit, ce n'est pas lui qui «pense» ou qui est d'avis que si l'on en fait de trop on risque de mal le faire, même si c'est lui qui énonce effectivement le proverbe en question. En termes de polyphonie, s'il est le *locuteur* du proverbe, il n'est pas, par contre, «l'*énonciateur* du principe qui y est attaché» (Anscombre, 1994 : 100). La vérité générale exprimée par le proverbe a une autre source ou autre voix que celle du locuteur qui l'emploie. Dans le cadre de la théorie de la pertinence (Sperber et Wilson, 1989), on parle d'énoncé *échoïque* pour rendre compte du fait que le locuteur produit un énoncé qui n'est qu'un écho de propos ou de pensées d'autrui[7].

Ce qui est souvent ignoré ou reste implicite, même si les manifestations de ce phénomène sont signalées ailleurs, au niveau de la fixité ou rigidité relative du proverbe, c'est qu'il y a un deuxième sens à dire que le locuteur n'est pas l'auteur du proverbe. Il n'est pas non plus le responsable de la forme du proverbe, c'est-à-dire du choix des mots, de leur combinaison, du processus métaphorique choisi s'il y en a un, etc. Ni le traitement polyphonique ni l'analyse en termes d'énoncé échoïque ne rendent explicitement compte de cet aspect-là des choses, puisqu'ils n'exigent pas qu'il y ait eu énoncé antécédent effectif et donc répétition fidèle de paroles prononcées. Il faut uniquement que ce soit la pensée ou le contenu qui ne soit pas propre au locuteur, mais provienne d'autrui. On en tient une preuve toute relative dans la perte du caractère proverbial lorsqu'on paraphrase un proverbe, c'est-à-dire lorsqu'on explicite son sens. Quoiqu'on continue de respecter le contenu du proverbe ainsi explicité et que la responsabilité du locuteur ne se trouve donc pas engagée vis-à-vis du principe exprimé, puisque celui-ci reste grosso modo le même que celui du proverbe, l'expression produite perd dans l'opération son statut de proverbe. C'est ainsi que si on paraphrase :

La langue va où la dent fait mal
À chaque pot son couvercle[8]
Loin des yeux, loin du coeur

par des énoncés explicatifs tels que :

---

[7] De façon plus précise, comme le développe Gouvard (1996 : 56-57) en citant Sperber et Wilson (1989 : 357), un énoncé est échoïque lorsqu'on l'interprète de façon échoïque, c'est-à-dire lorsqu'on lui attribue une interprétation qui «doit sa pertinence au fait que le locuteur se fait à sa façon l'écho des propos ou des pensées d'autrui», dans le but de véhiculer une information qui lui est propre.

[8] Proverbes cités par Gouvard (1996).

On parle plus volontiers de ses peines
Chaque femme ou chaque homme finit par trouver l'homme ou la femme qui lui convient
L'éloignement de deux êtres qui s'aiment fait diminuer leur amour

on observe que ceux-ci y ont laissé des plumes ... proverbiales. Surtout les deux derniers, parce que le premier, étant donné précisément sa forme, peut encore passer pour un *écho* de forme et de contenu, c'est-à-dire pour un proverbe.

Il n'y a rien d'étonnant, étant donné l'oubli de la forme dont nous venons de parler, que les tests destinés à illustrer que le locuteur d'un proverbe n'est pas l'auteur du proverbe concernent avant tout le contenu notionnel du proverbe. Anscombre (1994 : 100), s'appuyant sur l'étude de Ducrot (1975) sur *je trouve que*, note que les proverbes ne se combinent guère avec cette expression performative d'opinion individuelle, lorsqu'il «s'agit d'exprimer une adhésion générale au principe exprimé par le proverbe» :

**Je trouve que la fortune sourit aux audacieux*
**Je trouve que qui va à la chasse perd sa place*
**Je trouve que petite pluie abat grand vent*
**Je trouve que prudence est mère de sûreté*[9]

Pour les mêmes raisons, il est difficile d'avoir les combinaisons *je crois que + proverbe, je pense que + proverbe, selon moi + proverbe, à mon avis + proverbe*, etc. (Gouvard, 1996 et Anscombre, 1994). Elles posent toutes l'identité entre le locuteur et l'énonciateur de l'opinion, alors que le proverbe postule au contraire la dissociation de ces deux rôles :

?Je crois que qui trop embrasse mal étreint
?Je pense que qui trop embrasse mal étreint
?Selon moi, qui trop embrasse mal étreint
?À mon avis, qui trop embrasse mal étreint

Les adverbes d'énonciation *franchement* et *visiblement* donnent lieu au même constat (Anscombre, 1994 : 101) :

? Visiblement, qui ne risque rien n'a rien
??Franchement, pas de nouvelles bonnes nouvelles

Le cas d'*estimer* est traité différemment. Comme il «admet la reprise d'un jugement dont le locuteur n'est pas l'auteur» (Anscombre, 1994 : 100), on s'attend à ce que les proverbes puissent lui servir de complément. Et pour Anscombre le résultat est effectivement meilleur, même s'il reconnaît dans une note et par les points d'interrogation dont il préfixe les deux premiers exemples de ci-dessous que la combinaison est généralement jugée imparfaite[10] :

?J'estime que la fortune sourit aux audacieux
?J'estime que qui va à la chasse perd sa place

---

9 La stellarisation est celle d'Anscombre.

10 D'un autre côté, il souligne qu'elle est «parfois même acceptée sans problème par les sujets parlants» (1994 : 100). Nous reviendrons sur ce point dans notre dernière partie.

J'estime que prudence est mère de sûreté
J'estime que le soleil luit pour tout le monde

Il faut bien souligner que le test de *je trouve que* + *proverbe* n'est pertinent que si la portée du jugement s'exerce sur la validité du contenu général exprimé par le proverbe. Si le locuteur n'est pas l'auteur du proverbe, il est par contre «l'auteur» de son emploi : c'est lui qui est responsable de l'énonciation du proverbe et du principe qui y est attaché, et qui, selon le type d'emplois[11], «endosse la responsabilité de déclarer ce principe applicable *hic et nunc*» (Anscombre, 1994 : 100). Si donc les performatifs d'opinion du type *je trouve que* et autres adverbes d'énonciation incompatibles en portée générale avec le proverbe ont pour objet l'application du proverbe à une situation particulière, l'interdiction d'une combinaison avec un proverbe se fait moins forte, comme le montre Anscombre (1994 : 101) avec les exemples suivants :

Je trouve que, pour une fois, à quelque chose malheur est bon
Je trouve que, au vu des circonstances, pas de nouvelles, bonnes nouvelles
Je trouve que, dans le cas qui nous occupe, le vin étant tiré, il faut le boire
Visiblement, de nos jours, qui ne risque rien, n'a rien
Franchement, au vu des circonstances, pas de nouvelles, bonnes nouvelles
À mon avis, dans ton cas, prudence est mère de sûreté

auxquels on peut ajouter :

Je crois que, avec Fred, loin des yeux, loin du coeur
Je pense que, dans cette affaire, qui trop embrasse mal étreint
Selon moi, étant donné la situation, prudence est mère de sûreté

Dans ce cas, le jugement exprimé par le locuteur-énonciateur ne dit pas que ce qu'exprime le proverbe est vrai, c'est-à-dire que le contenu notionnel est valide en général, mais que le proverbe est vrai dans telle ou telle situation particulière envisagée, c'est-à-dire que le proverbe s'applique à ou se trouve vérifié par telle ou telle situation particulière envisagée.

1.2 Jugement individuel et jugement collectif

Les tests employés jusqu'à présent permettent uniquement de conclure que ce n'est pas le locuteur d'un proverbe qui en est l'auteur. Ils n'autorisent pas, du moins directement, à affirmer qu'il ne s'agit pas d'un jugement individuel[12]. L'énonciateur, c'est-à-dire le responsable du proverbe, pourrait en effet être un autre particulier. La plupart des commentateurs rejettent cette possibilité : un particulier ne peut être tenu pour responsable d'un proverbe. Celui-ci est le fait d'un énonciateur collectif, de la *vox populi* et non de tel ou tel individu particulier. La polyphonie inhérente au proverbe met donc aux prises un particulier, le locuteur, qui énonce le proverbe, et un énonciateur collectif (cf. supra la *conscience linguistique collective* d'Anscombre). L'énoncé échoïque que représente le proverbe «n'est attribuable à aucune source précise, mais au peuple dans son ensemble»

[11] Nous ne pouvons dans le cadre de ce travail développer ce point comme il faudrait. Signalons simplement qu'un locuteur peut employer un proverbe de façon «non spécifique».
[12] Gouvard (1996) est un des rares à faire soigneusement cette distinction.

(Sperber et Wilson, 1989 : 358)[13], ce qui conduit Gouvard à une redéfinition «écho-écot» du proverbe comme étant en somme *l'écho d'un écho :*

«Un énoncé proverbial est un énoncé dont l'interprétation échoïque implique nécessairement que l'énoncé dont le locuteur se fait l'écho n'est lui-même interprétable que sous une forme échoïque.» (Gouvard, 1996 : 57)

Ce trait permet classiquement de séparer les proverbes des sentences, maximes, aphorismes, apophtegmes, slogans, etc., c'est-à-dire de tous les énoncés sentencieux qui ont un «père» particulier (identifiable ou non). On signale tout aussi traditionnellement que de tels énoncés individuels peuvent devenir proverbes[14]. Ce n'est pas parce que l'on connaît l'origine de tel ou tel proverbe particulier que le proverbe en perd pour autant son statut de vérité «collective». La connaissance diachronique que les proverbes :

À petite cloche, grand son
Honni soit qui mal y pense

sont à l'origine les devises de la maison de Grandson et de l'ordre de la Jarretière[15] ne les empêche pas synchroniquement d'apparaître comme des jugements non plus individuels, mais collectifs, des vérités générales faisant partie du stock de principes généraux communs à toute une communauté (ou *peuple*, pour reprendre le terme de Sperber et Wilson).

De la même manière, la confection de proverbes, activité poético-ludique assez répandue, n'entre pas en contradiction avec le trait de vérité collective attribué au proverbe. Ces proverbes «fabriqués» ou *formes proverbiales*, comme les appelle Gouvard (1996) pour les séparer des proverbes attestés, sont certes des créations individuelles, mais leur «géniteur» les a fabriqués de telle sorte qu'ils passent pour ne pas avoir d'auteur particulier, mais relèvent de la voix anonyme collective des proverbes. S'il n'en allait pas ainsi, leur auteur aurait manqué son coup : on ne les reconnaîtrait pas pour ce qu'il veut qu'on les reconnaisse, à savoir des proverbes. Sentences de particuliers passés en proverbes comme néologismes proverbiaux fabriqués par des particuliers ne peuvent ainsi remettre en cause la portée de jugement collectif reconnu aux proverbes. Dans les deux cas, celui des phrases devenues proverbes comme celui des formes proverbiales, il faut évidemment que l'expression en question réponde aux autres attributs qui définissent le proverbe ou qui, du moins, font que l'on reconnaisse telle ou telle forme comme étant ou comme pouvant être un proverbe.

Le critère classique qu'on associe à la propriété de jugement collectif est l'expression métalinguistique à dire indéfini *comme on dit* avec un *on* révélateur (Kleiber, 1989a et 1994; Anscombre et Gouvard, 1996), qui s'oppose à l'expression à dire défini *comme (le) dit X* où X représente un individu particulier :

Et, comme on dit, qui trop embrasse mal étreint

---

13 Que nous citons d'après Gouvard (1996 : 57).

14 Faits cités par Gouvard (1996 : 57).

15 Cf. l'expression *passer en proverbe*.

Comme on dit, jamais à bon chien il ne vient un os (cité par Gouvard, 1996)
Comme dit Rabelais, fays ce que voudras
Comme le dit La Rochefoucauld, le refus des louanges est un désir d'être loué deux fois (cité par Anscombre, 1994)

Anscombre (1994 : 99) cite également des tournures telles que *On a bien raison de dire que ..., si j'en crois la sagesse populaire* (var. *la sagesse des nations*) :

On a bien raison de dire qu'on n'est jamais trahi que par les siens
Si j'en crois la sagesse populaire / la sagesse des nations, il ne faut jamais mettre la charrue avec les boeufs

On peut aussi citer ici la tournure *comme (le) dit le proverbe* :

Comme le dit le proverbe, un homme averti en vaut deux

dont la particularité, par rapport aux expressions métonymiques du type *la phrase dit / cette expression dit ...*, avec lesquelles on peut être tenté de l'assimiler, consiste à ce que la définitude du proverbe se trouve justifiée par l'apparition même du proverbe comme complément. On ne peut avoir dans les mêmes conditions :

Comme le dit la phrase / l'expression, cherche ton ciel et tu trouveras ton soleil

un tel énoncé nécessitant une justification préalable (c'est-à-dire extérieure à la mention qui suit) de la définitude de la phrase ou de l'expression. L'emploi de l'indéfini confirme cette différence. Face à :

Comme dit un proverbe, le bon est bon, mais le meilleur est meilleur

on ne peut énoncer :

?*Comme dit une phrase, les menteries ont courtes jambes*
?*Comme dit une phrase, les proverbes doivent être décortiqués*

sans provoquer une surprise légitime, même si la phrase qui suit est une ... phrase, proverbe ou non.

Le critère du *Comme dit X* vs *comme on dit* demande toutefois des précisions supplémentaires. Premièrement, il n'est pas exclu d'avoir *comme (le) dit X* avec un proverbe, ainsi que le signale fort justement Gouvard (1996 : 50) :

Comme (le) dit ma concierge, les mouches se reposent toujours sur les chevaux maigres
Comme (le) dit ma grand-mère, vieux mari et jeune femme, des cornes en campagne

mais on observera qu'un tel X ne passe pas pour être l'auteur ou le responsable du proverbe, mais plutôt un utilisateur ou un «habitué» du proverbe. Ma grand-mère ne se trouve pas présentée comme étant à l'origine du proverbe *petite pluie abat grand vent*, mais

le fait de l'employer fréquemment constitue une habitude qui la caractérise[16]. On le vérifiera en transformant l'habitualité du dire de *comme dit X* en épisodicité, c'est-à-dire en dire-occurrence (Kleiber, 1987). Les énoncés obtenus :

?*Comme a dit ma concierge, les mouches se reposent toujours sur les chevaux maigres*
?*Comme a dit ma grand-mère, vieux mari et jeune femme, des cornes en campagne*

passent mal la rampe, dans l'interprétation où X est le responsable du proverbe.

Cette mise au point n'est pas sans conséquence pour le critère de la tournure *comme on dit* elle-même. Elle montre en effet que cette tournure n'est nullement une preuve directe du caractère collectif du jugement exprimé par le proverbe, mais représente une manifestation de l'utilisation commune ou partagée par les locuteurs qui en est faite. Elle rejoint un autre trait du proverbe, celui de la collocation avec l'adjectif *connu* (cf. *un proverbe (bien) connu*). Nous aurons l'occasion d'y revenir ci-dessous lorsque nous aborderons la question de la dénomination. Pour le moment, on remarquera en faveur de notre rectification la similitude aspectuo-temporelle avec la structure *comme dit X* : de même que l'on ne peut avoir en interprétation épisodique *comme a dit X + Proverbe*, de même on ne peut avoir en lecture événementielle *comme on a dit + Proverbe* :

?*Comme on a dit, petite pluie abat grand vent.*

Il est significatif de constater que la même contrainte vaut pour l'expression *comme dit le proverbe*[17] :

?*Comme l'a dit le proverbe, rien ne sert de courir il faut partir à point.*

Ce n'est pas pour autant qu'il faut écarter ce critère : l'indéfinitude collective du *on* jointe à l'habitualité du dire permettent de conclure indirectement que son énonciation, lorsqu'elle a lieu, n'est plus le fait d'un particulier, puisque c'est déjà une habitude de le dire (*comme on dit ...*) ou puisque *le proverbe le dit* déjà. D'autre pat, il a une vertu supplémentaire, c'est celle de mettre l'accent, par l'intermédiaire du verbe *dire* du *comme*, sur le côté formel de la chose — non touché, on le rappelle, par les verbes d'opinion — en exigeant que la forme ait, d'une manière ou d'une autre, un caractère remarquable, saillant[18]

---

16 Nous nous séparons sur ce point un tout petit peu de Gouvard (1996 : 50) qui pense que l'individu particulier X est supposé véhiculer la sagesse exprimée par le proverbe, «parce que ce trait fait partie de sa représentation prototypique».

17 Petite différence entre *comme on dit + proverbe* et *comme dit le proverbe + proverbe*, on peut avoir pour la seconde tournure le pronom *le* en pronom anticipant : *?comme on le dit, petite pluie abat grand vent* vs *comme le dit le proverbe, petite pluie abat grand vent.*

18 Ce fait n'est pas particulier au proverbe. Il intervient aussi avec les sentences et autres expressions remarquables (voir aussi pour les expressions idiomatiques). Sans justifications contextuelles spéciales, toujours possibles pour assurer la saillance du dit, on a difficilement *Comme le dit* ou *Comme l'a dit La Rochefoucauld / Comme le dit/ l'a dit ma grand-mère, il fait froid le soir/ je mange du veau / j'ai vu passer deux chevaux*, etc.

Le critère du *comme on dit + proverbe* n'est donc pas totalement injustifié, mais, pour montrer qu'un proverbe n'est pas un jugement individuel, on aurait dû, en bonne logique, d'abord reprendre le test utilisé pour démontrer que le locuteur n'est pas l'auteur du proverbe en plaçant un autre individu que le locuteur comme le sujet du verbe d'opinion[19]. Si des tournures telles que *Paul trouve que / croit que / pense que + proverbe, Selon Paul, + proverbe* apparaissent mal formées, alors on peut logiquement en conclure — directement — que le proverbe n'est pas le fait d'un individu particulier et alors en tirer la conclusion qu'il ne peut s'agir que d'un jugement déjà acquis, qui appartient à une conscience collective, etc. L'essai semble probant, puisque des énoncés tels que :

?Paul trouve que qui va à la chasse perd sa place
?Paul croit que qui trop embrasse mal étreint
?Paul pense que qui trop embrasse mal étreint
?Selon Paul, qui trop embrasse mal étreint

suscitent une réaction de surprise, si l'on reconnaît à la phrase complément le statut de proverbe, le proverbe n'acceptant guère d'être remis ou replacé en somme sous la responsabilité d'un particulier. S'il s'agit d'une application particulière, les choses s'améliorent, pour les raisons signalées ci-dessus :

Paul trouve que, pour une fois, à quelque chose malheur est bon
Paul pense que, pour une fois, à quelque chose malheur est bon
Paul croit que, dans cette affaire, qui trop embrasse mal étreint
Selon Paul, étant donné la situation, prudence est mère de sûreté

mais le phénomène du discours oblique fait que ce n'est pas Paul qui est obligé d'avoir prononcé le proverbe, le locuteur pouvant en l'occurrence l'appliquer pour caractériser l'opinion de Paul. Cette possibilité n'a toutefois aucune conséquence sur le résultat que l'on peut tirer du critère : les proverbes apparaissent comme étant des jugements non individuels.

2. UNE REMISE EN CAUSE DU CARACTÈRE «NON INDIVIDUEL» DÉFINITOIRE DES PROVERBES

Ce résultat définitoire se trouve par contre sérieusement ébranlé si l'on arrive à montrer que le critère n'est pas totalement fiable et qu'il existe malgré tout des énoncés où le proverbe se présente comme étant le jugement émis par un individu. Michaux (1995 et 1996) montre tout d'abord qu'on peut avoir deux types de lecture applicative du proverbe, qui ne remettent pas en cause le caractère non individuel du proverbe. À côté de l'application particulière évoquée ci-dessus, qu'elle appelle lecture *métalinguistique locale*, comme en :

Je trouve que, pour une fois, à quelque chose malheur est bon

il est possible d'avoir encore une autre lecture applicative métalinguistique, une lecture *métalinguistique générale*, qui «porte sur l'opinion du locuteur quant à la validité en

---

[19] Si on a recours à *je trouve que*, c'est à cause de la spécificité de *je trouve que* lorsqu'il est performatif (Ducrot, 1975).

général du proverbe subordonné» et qui donc «projette la validité du proverbe sur l'ensemble des situations vérifiées par le principe général sous-jacent à ce proverbe» (Michaux, 1996). Un locuteur peut ainsi dire :

C'est bien dommage que chien qui aboie ne mord jamais

pour rendre compte «de sa tristesse de constater que dans la vie (c'est-à-dire telle qu'il la connaît), le topos correspondant à la parémie en question est généralement valide» (Michaux, 1996). Il en va de même dans :

Je crains fort que dans l'absolu les cordonniers sont les plus mal chaussés

où il signifie sa crainte de voir le proverbe être vrai «dans l'absolu». Dans ces deux cas, comme dans la situation d'application particulière, il y a une dissociation polyphonique entre locuteur et énonciateur : «le locuteur indique clairement qu'il n'est pas l'auteur du proverbe, même s'il trouve applicable *hic et nunc* (lecture métalinguistique locale) ou en général (lecture métalinguistique générale) le principe qui lui est attaché» (Michaux, 1996). Ce ne sont donc pas de telles interprétations qui peuvent être invoquées pour remettre en cause l'incompatibilité du proverbe avec les jugements individuels. Décisive, par contre, s'avère la combinaison dans laquelle le locuteur exprime son avis sur le contenu et, en même temps qu'il en est le locuteur, se fait donc l'auteur de la parémie exposée :

Je trouve qu'abondance de biens ne nuit pas
Je trouve que l'argent ne fait pas le bonheur.

À côté de l'interprétation métalinguistique toujours possible, ces deux énoncés donnent lieu à une lecture portant sur le contenu de la complétive : «le locuteur n'indique pas qu'il considère que le proverbe est un principe valide en général. Il reprend à son compte l'attribution d'un prédicat à un syntagme nominal» (Michaux, 1996).

Gouvard (1995 : 50) souligne également la possibilité d'avoir des constructions qui impliquent la responsabilité énonciative du locuteur :

Je trouve que l'habit ne fait pas le moine
Je crois qu'il vaut mieux prévenir que guérir
Je crois que l'excès en tout est un défaut

mais il n'en conclut pas que les proverbes peuvent être des jugements individuels. Il rejette uniquement comme test formel d'identification le critère de l'incompatibilité des proverbes avec les verbes d'opinion individuelle. Il conserve en effet le test de *comme on dit* et maintient dans son approche définitoire du proverbe que le locuteur n'est pas l'auteur du proverbe, puisque celui-ci se trouve défini comme étant l'écho d'un écho.

Michaux (1996), par contre, va jusqu'au bout : les combinaisons *je trouve que* + *proverbe* à lecture de contenu prouvent, selon elle, qu'un «proverbe peut, dans certaines conditions, être assimilé à un jugement individuel émis par le locuteur hors de toute

situation particulière»[20]. Et elle formule une explication intéressante pour rendre compte du fait que certains proverbes peuvent accepter la tournure *je trouve que* en lecture subjective et d'autres non. Les proverbes qui intrinsèquement répondent au critère sémantique de prédication originelle[21] exigé par *je trouve que* acceptent sans trop de peine de devenir la complétive de *je trouve que*. Si, par exemple :

Je trouve qu'il n'est point de sot métier

passe mieux la rampe que :

?*Je trouve que l'eau va à la rivière*

c'est parce que le jugement *Il n'est point de sot métier* est plus facilement un jugement formé sur une prédication originelle, cette base de références subjective déjà disponible, que l'assertion *L'eau va à la rivière*. Pour sauver *Je trouve que l'eau va à la rivière*, il ne reste plus que la lecture métalinguistique générale (ou locale pour *Je trouve que, pour une fois, l'eau va à la rivière*), qui justifie la présence de *je trouve que*.

Le même fonctionnement se laisse appliquer, comme le montre Michaux (1995 et 1996) aux autres verbes d'opinion et aux autres verbes recteurs d'une complétive, à la seule différence qu'il faut à chaque fois que le proverbe présente intrinsèquement les traits exigés par chaque type de verbe. Ainsi *J'estime que + proverbe, je crois que + proverbe, je pense que + proverbe,* etc., seront des combinaisons permises en interprétation non métalinguistique si l'assertion proverbiale répond aux conditions spécifiques exigées respectivement par *j'estime que, je crois que, je pense que,* etc. Cette façon de procéder permet à Michaux (1996) d'expliquer autrement qu'Anscombre pourquoi l'énoncé :

**Je trouve que prudence est mère de sûreté*

est mal formé en lecture de jugement sur le contenu du proverbe, alors qu'une semblable lecture ne pose pas de difficulté pour :

J'estime que prudence est mère de sûreté.

Elle a pour conséquence supplémentaire de conduire à une subdivision des proverbes en deux groupes, selon qu'ils peuvent donner lieu ou non à une lecture sur leur contenu parémique. Le premier groupe, que Michaux (1996) rapproche des phrases génériques typifiantes locales telles que :

Les chats sont affectueux

---

[20] Elle est pourtant à un moment donné de l'article d'un avis différent, lorsqu'elle souligne que l'affirmation d'Anscombre que «le proverbe n'est en rien une opinion personnelle» constitue un fait qui n'est pas critiquable.

[21] Nous ne discuterons pas de cette notion ici. Voir à ce sujet Ducrot (1975) et son analyse de *Je trouve que ...* et les commentaires et discussions de Michaux (1996). Disons simplement qu'il s'agit de rendre compte du fait que l'opinion exprimée par le locuteur avec *je trouve que* s'établit à partir de données préalables ou *prédication originelle* dont dispose déjà le locuteur.

rassemble «les proverbes dont la nature même peut conduire à un jugement individuel sur leur contenu». Le second, qui est à rapprocher des phrases génériques typifiantes *a priori* comme :

Les castors construisent des barrages

réunit les proverbes «qui, lorsqu'ils permettent l'accès à un jugement d'évaluation, ne peuvent le faire que *via* une interprétation métalinguistique». Résultat final : on ne peut plus définir les proverbes comme des jugements non individuels, puisque le premier groupe se définit précisément par la possibilité de donner lieu dans certaines situations à un jugement individuel.

La chose est gênante, puisqu'on se prive d'un des traits définitoires qui semblait le plus solide, mais elle est logique : si les proverbes peuvent devenir les compléments de verbes d'opinion personnelle, on ne peut plus les définir comme étant des jugements collectifs, des assertions qui ne sont pas des avis personnels, mais qui constituent des vérités faisant partie d'une communauté, appartenant à la «conscience linguistique collective». Les deux conséquences linguistiques qui découlent du caractère *traditionnel* ou *populaire* assigné habituellement aux proverbes, que nous avons mis en avant au début de ce travail, sont évidemment caduques aussi :

- celui qui emploie un proverbe peut en être également l'énonciateur ou l'auteur;
- l'auteur d'un proverbe peut être un particulier.

3. VERS UNE SOLUTION EN TERMES DE DÉNOMINATION

Faut-il aller aussi loin ? Nous ne le pensons pas, malgré les données linguistiques relevées en faveur d'une telle conclusion par Michaux et Gouvard. Il ne s'agit pas de nier ces combinaisons avec des verbes qui engagent la subjectivité du locuteur. On peut effectivement avoir des énoncés tels que :

Je trouve qu'abondance de biens ne nuit pas
Je trouve que l'argent ne fait pas le bonheur
Je trouve que l'habit ne fait pas le moine
Je crois qu'il vaut mieux prévenir que guérir
Je crois que l'excès en tout est un défaut
J'estime que prudence est mère de sûreté
Paul croit qu'il vaut mieux prévenir que guérir
Paul estime que l'argent ne fait pas le bonheur

où c'est le contenu prédicatif SN-SV qui se trouve à chaque fois mis en jeu par le locuteur. Mais on n'est pas pour autant forcé à abandonner l'idée que les proverbes ne sont pas des jugements individuels. Tout simplement, parce qu'une telle conclusion n'est valide que si et seulement si on considère que les prédications en question continuent d'être des proverbes. Or, une telle position n'est nullement obligatoire. Il n'est pas du tout nécessaire de souscrire au maintien du statut de proverbe lorsqu'un proverbe se trouve inséré dans de telles combinaisons.

Notre hypothèse est que dans ces énoncés nous n'avons en fait pas la combinaison d'un verbe d'opinion et d'un proverbe. Ce n'est pas *je trouve que + proverbe, Je crois que + proverbe, j'estime que + proverbe*, etc., mais *je trouve que + proposition, je crois que + proposition, j'estime que + proposition*, etc. Comment cela est-il possible, alors qu'il n'y a pas changement de forme entre le proverbe testé et l'énoncé avec le verbe d'opinion ? C'est bien cette identité de forme qui conduit à penser qu'un proverbe peut donner lieu à un jugement individuel. En fait, c'est bien un proverbe qui se trouve pris lorsqu'on applique le test de l'insertion comme complément d'un verbe d'opinion individuelle, mais cette insertion, si elle ne touche pas à sa forme, affecte son statut de proverbe. Ce qui se trouve changé lors de ce placement sous la responsabilité d'un particulier, c'est le caractère de dénomination du proverbe.

3.1 Des dénominations d'un type «très très spécial»

Il nous faut donc au préalable rappeler que les proverbes sont des dénominations d'un type «très très spécial». La particularité fondamentale d'un proverbe est d'être, comme nous l'avons montré ailleurs (1989a et 1994), à la fois, une dénomination[22], c'est-à-dire une unité codée, faisant partie du code linguistique, en ce qu'elle nomme une entité générale et non un particulier, et une phrase. Ce double aspect, antinomique, fonde son originalité sémiotique, que nous avions soulignée comme suit : «En tant que phrase, il ne devrait pas être signe (ou unité codée), puisque l'interprétation d'une phrase est une construction et non un donné préalable. En tant que dénomination, il est néanmoins une unité codée, c'est-à-dire un signe. Un signe-phrase donc, qui possède les vertus du signe sans perdre pour autant son caractère de phrase, de même que susbstantifs, verbes, adjectifs, etc., sont des dénominations qui conservent les attributs spécifiques des catégories grammaticales qu'ils représentent» (1994 : 214).

On notera un premier avantage de cette caractérisation en relation avec notre sujet : le statut de dénomination phrastique des proverbes permet d'expliquer directement pourquoi ils passent pour être des jugements collectifs et non individuels : c'est parce qu'ils sont des phrases-dénominations que la prédication exprimée ne peut être portée au crédit d'un particulier. Cette prédication fait partie du code linguistique commun et, en tant que telle, est une unité dont l'existence n'a plus à être postulée. Elle s'impose à tout locuteur comme toutes les autres dénominations. Observons que cela ne signifie pas qu'il doive lui croire personnellement à son contenu, mais uniquement qu'il est obligé d'accepter que ce contenu, c'est-à-dire le principe attaché au proverbe, est le sens conventionnellement assigné à la phrase proverbiale. De même que le sens d'une unité lexicale est un sens «collectif», de même le contenu, c'est-à-dire la prédication ou le jugement d'un proverbe, est «collectif» et non une affaire de particulier. Le caractère de *vox populi* des proverbes n'est donc qu'une conséquence de leur caractère de dénominations phrastiques.

Ce qui est remarquable, c'est qu'on arrive à rendre compte par la même occasion du fait que le locuteur d'un proverbe n'est pas non plus l'auteur de la forme du proverbe. Nous avons en effet souligné ci-dessus que la plupart des commentateurs oubliaient qu'il y avait un deuxième sens à affirmer qu'un locuteur n'est pas le responsable du proverbe qu'il utilise : il n'est pas non plus maître du choix des mots, de leur combinaison, du processus

---

22 Les arguments sont nombreux (Kleiber, 1989 et 1994) : (relative) fixité de la forme, apprentissage, présence dans les dictionnaires, (relative) opacité, etc.

métaphorique choisi, s'il y en a un, etc. L'accent est uniquement mis sur le contenu, que ce soit dans l'approche polyphonique ou dans le traitement échoïque. Or, le locuteur d'un proverbe n'est nullement maître de la forme du proverbe qu'il énonce et s'il en va ainsi, c'est bien parce que la phrase énoncée est une dénomination, c'est-à-dire l'association stable d'une forme et d'un sens. De même que l'usager d'un substantif ou d'un adjectif n'est pas le responsable de la forme du substantif ou de l'adjectif, de même il n'est pas responsable de la forme du proverbe qu'il emploie.

Ipso facto, on comprend beaucoup mieux aussi la portée du critère de *comme on dit*. S'il convient à tous les proverbes, il ne leur est pourtant pas spécifique, comme le pense Gouvard (1996). Il satisfait aussi aux locutions figées non proverbiales :

Et, comme on dit, les carottes sont cuites.

Et, comme nous l'avions annoncé ci-dessus, il n'est nullement une manifestation directe du caractère collectif du *jugement* exprimé par le proverbe. Comme il ne s'applique pas aux dénominations simples :

?J'ai vu un «chien», comme on dit

et qu'il ne convient pas non plus aux séquences non figées :

?Et, comme on dit, il est huit heures

s'il s'applique aux proverbes et aux autres expressions figées, c'est parce qu'il s'agit de dénominations polylexicales.

3.2 Déproverbialisation

On peut à présent revenir au problème que constitue l'occurrence d'un proverbe avec des expressions les plaçant sous la responsabilité d'un particulier. Si l'on accepte notre caractérisation des proverbes comme étant des dénominations phrastiques, on peut expliquer quel élément se trouve modifié par une éventuelle insertion dans une structure du type *verbe d'opinion individuelle + proposition*. On sait qu'un proverbe en tant que phrase figée peut être, tout comme les autres expressions figées[23], défigé par un jeu formel (Grésillon et Maingueneau, 1984; Franken, 1995). Ces défigements portent sur l'aspect formel de la phrase. Il est aussi possible de ne pas modifier la forme, mais de toucher à l'autre aspect du proverbe, à son caractère de dénomination. Nous parlerons dans ce cas de *déproverbialisation*. La déproverbialisation est l'opération qui fait faire perdre au proverbe son côté dénominatif, pour ne lui laisser que son aspect de phrase. Le proverbe n'apparaît plus alors comme une phrase déjà construite, fixée, dont le sens est donné par avance, c'est-à-dire dont la prédication exprimée est acquise *a priori*, mais redevient une phrase comme les autres, qui présente à validation la combinaison SN-SV (s'il s'agit de cette structure). Un proverbe déproverbialisé est un proverbe qui perd son statut d'unité codée pour redevenir une simple phrase, une phrase générique, puisqu'ainsi que nous l'avons montré (1989a et 1994), les proverbes sont des phrases génériques. C'est ce qui se passe, nous semble-t-il, lorsqu'on place un proverbe sous la dépendance d'une expression de

[23] Pour une synthèse sur les locutions figées, voir Gross (1996) et Mejri (1996).

jugement individuel. C'est ainsi que le proverbe *L'argent ne fait pas le bonheur* s'il est inséré dans une structure du type *je trouve que / je crois que ...* en lecture non métalinguistique se trouve être déproverbialisé : il se voit privé de son statut de dénomination pour ne conserver que le statut d'une phrase générique, par laquelle, le locuteur-auteur donne son avis personnel sur le rapport entre le fait d'avoir de l'argent et celui d'être heureux. On peut ainsi maintenir qu'un proverbe n'est pas l'expression d'un jugement particulier, tout en rendant compte par le biais de la déproverbialisation des emplois du type *Je trouve que l'argent ne fait pas le bonheur.*

Un argument de poids peut être invoqué : lorsque la forme figée, qui, on le rappelle, est l'élément essentiel du caractère dénominatif du proverbe, est trop éloigné, soit par la construction syntaxique, soit par un côté métaphorique, de la forme des phrases génériques, elle rend difficile l'insertion du proverbe dans une structure qui le subordonne à une opinion particulière. Pour une bonne et simple raison : une phrase dotée d'une telle forme se combine difficilement à une expression qui, elle, ne présente, au contraire, aucun caractère saillant. Le proverbe ne passe du coup pas l'étape de la déproverbialisation : l'aspect formel fait que le caractère dénominatif n'est pas gommé dans l'histoire et donc interdit au proverbe de passer pour une phrase générique exprimée par un particulier. C'est pour cette raison que des combinaisons telles que :

?Je crois que qui trop embrasse mal étreint
?Je pense que qui trop embrasse mal étreint
?Je trouve que à petite cloche grand son
?Je crois que loin des yeux loin du coeur

paraissent mal formées[24], alors que des structures telles que :

Je trouve que l'argent ne fait pas le bonheur
Je crois qu'il vaut mieux prévenir que guérir
J'estime que prudence est mère de sûreté

ne suscitent aucun sentiment de rejet, parce que la forme du proverbe inséré se prête à une déproverbialisation sans heurts. Insistons bien sur un point : ce n'est pas le côté connu du proverbe qui est à l'origine de la difficulté d'apparaître comme jugement individuel déproverbialisé. *Purdence est mère de sûreté* est un proverbe bien connu, qui, accepte pourtant d'être présenté comme le jugement d'un particulier, parce que sa forme ne suscite pas de difficultés d'insertion.

Une difficulté reste à surmonter. Si nore explication est correcte, pourquoi certains proverbes acceptent-ils d'être déproverbialisés par tel ou tel verbe d'opinion et non par tel ou tel autre ? Pourquoi le proverbe *Prudence est mère de sûreté* peut-il apparaître avec *j'estime que* et non avec *je trouve que :*

J'estime que prudence est mère de sûreté
?Je trouve que prudence est mère de sûreté.

---

[24] Michaux (1996) avait entrevu ce facteur en notant qu'outre le caractère de prédication originelle, «le caractère métaphorique du proverbe influence lui aussi le type de lecture possible.»

La raison en est simple : c'est parce que la phrase générique obtenue ne peut être un complément de *je trouve que*, mais de *j'estime que* ... C'est à cette place que peuvent entrer en ligne de compte les analyses et les résultats obtenus par Michaux (1995 et 1996). Il est en effet naturel que chaque verbe d'opinion sélectionne son type de complément propositionnel et qu'un proverbe dégradé au niveau de simple phrase générique doive satisfaire aux conditions posées par le verbe enchâsseur.

CONCLUSION «INDIVIDUELLE»

Peut-on en tirer parti pour distinguer différentes classes de proverbes ? Nous laisserons pour aujourd'hui cette question ouverte : elle mérite d'être traitée avec le plus grand soin, notamment dans sa comparaison avec les différentes catégories de phrases génériques[25]. Il nous suffit d'avoir atteint le but que nous nous somme fixé : celui d'avoir montré, grâce à l'hypothèse de la dénomination et de l'opération de déproverbialisation, pourquoi les proverbes ne pouvaient pas être des jugements individuels, malgré des données apparemment contraires. Corollairement, se trouve confortée — et c'est réjouissant — la tendance «optimiste» qui estime que les proverbes forment une classe suffisamment homogène pour être définie linguistiquement.

RÉFÉRENCES

ANSCOMBRE, J.C. (1989) : «Théorie de l'argumentation, topoï et structuration discursive», *Revue québécoise de linguistique*, 18, n° 1, pp. 13-56.

ANSCOMBRE, J.C. (1990) : «Les syllogismes en langue naturelle ? Déduction logique ou inférence discursive ?», *Cahiers de linguistique française*, 11, pp. 215-240.

ANSCOMBRE, J.C. (1994) : «Proverbes et formes proverbiales : valeur évidentielle et argumentative», *Langue française*, 102, pp. 95-107.

ANSCOMBRE, J.C. (1995) : «La nature des topoï», *in* Anscombre J.C. (dir), *Théorie des topoï*, Paris, Kimé, pp. 49-84.

ARNAUD, P.J.L. (1991) :«Réflexions sur le proverbe», *Cahiers de lexicologie*, 59-2, pp. 6-27.

ARNAUD, P.J.L. (1992) : «La connaissance des proverbes français par les locuteurs natifs et leur sélection didactique», *Cahiers de lexicologie*, 60-1, pp. 195-238.

ARNAUD, P.J.L. et R. MOON (1993) : «Fréquence et emploi des proverbes anglais et français», *in* Plantin C. (dir), *Lieux communs, topoï, stéréotypes, clichés*, Paris, Kimé, pp. 323-341.

BERRENDONNER, A. (1981) : *Éléments de pragmatique linguistique*, Paris, Minuit.

DUCROT, O. (1975) : «Je trouve que», *Semantikos*, 1,1, pp. 62-88.

FOREST, R. (1996) : «Noms propres, idiotismes et polyphonie», *Bulletin de la Société de Linguistique de Paris*, t. XCI, 1, pp. 55-76.

[25] Nous espérons pouvoir mener une telle investigation sur les bases de notre approche des phrases génériques (Kleiber et Lazzaro, 1987; Kleiber, 1988 et 1989b).

FRANKEN, N. (1995) : «Sur les détournements de proverbes», Projet ARC «Typologie textuelle et théorie de la signification», Rapport de recherches n° 3, Bruxelles, Université Libre de Bruxelles.

GOUVARD, J.M. (1996), «Les formes proverbiales», *Langue française*, 110, pp. 49-63.

GOUVARD, J.M. (à paraître), «Les proverbes dans *Le Paysan parvenu*».

GROSS, G. (1996) : *Les expressions figées en français*, Paris, Ophrys.

GRÉSILLON, A. et D. MAINGUENEAU (1984) : «Polyphonie, proverbes et détournement ou *Un proverbe peut en cacher un autre*», *Langages*, 73, pp. 112-125.

KLEIBER, G. (1984) : «Dénomination et relations dénominatives», *Langages*, 76, pp. 77-94.

KLEIBER,G.(1987) :*Du côté de la référence verbale. Les phrases habituelles*, Berne, Peter Lang.

KLEIBER, G. (1988) : «Phrases génériques et raisonnement par défaut», *Le français moderne*, LXVI, 1/2, pp. 1-16.

KLEIBER,G(1989a) :«Sur la définition du proverbe», *Recherches Germaniques*, 2, pp. 233-252.

KLEIBER, G.(1989b) :«Généricité et typicalité», *Le français moderne*, LXVII, 3/4,pp.127-154.

KLEIBER, G. (1994) : *Nominales*, Paris, A. Colin.

KLEIBER, G. et H. LAZZARO (1987) : «Qu'est-ce qu'un SN générique ou Les carottes qui poussent ici sont plus grosses que les autres», *in* Kleiber G. (dir), *Rencontre(s) avec la généricité*, Paris, Klincksieck, pp. 73-111.

MEJRI, S. (1996) : *Le figement lexical*, Thèse d'État, Tunis, Université de Tunis.

MICHAUX, C. (1995) : «Parémies, collocations verbales et actes de parole ou comment les verbes aident à la classification et à la maîtrise des proverbes», Projet ARC «Typologie textuelle et théorie de la signification», Rapport de recherches n° 3, Bruxelles, Université Libre de Bruxelles.

MICHAUX, C. (1996) : «Proverbe et jugement individuel : deux incompatibles ?», Projet ARC «Typologie textuelle et théorie de la signification», Rapport de recherches n° 4, Bruxelles, Université Libre de Bruxelles.

NORRICK, N. (1985) : *How Proverbs mean. Semantic Studies in English Proverbs*, Berlin, Mouton.

OLLIER, M.L. (1976) : «Proverbe et sentence : le discours d'autorité chez Chrétien de Troyes», *Revue des Sciences Humaines*, XLI, 163, pp. 329-357.

RODEGEM, F. (1984) : «La parole proverbiale», *in* Suard F. et C. Buridant (dir), *Richesse du proverbe. Typologie et fonctions* II, Lille, Presses Universitaires de Lille, pp. 121-129.

SCHULZE-BUSACKER, E. (1984) : «Proverbe ou sentence : essai de définition, *in* Di Sefano G. et R.G. Mc Gilliway (dir), *La locution*, Montréal, Éditions Ceres, pp. 134-167.

SOMOLINOS RODRIGUEZ, A (1993) : «Arcaismos sintacticos en los proverbios franceses», *Paremia*, 1, pp. 55-63.

SPERBER, D. et D. WILSON (1989) : *La pertinence*, Paris, Minuit.

TAYLOR, A. (1931), *The Proverb*, Cambridge, Mass.

À LA RECHERCHE DE LA MÉMOIRE PERDUE. OU POUR UN DICTIONNAIRE HISTORIQUE DE L'ARABE

Taïeb BACCOUCHE

Université de Tunis I, Tunis, Tunisie

Quand on considère l'histoire de la lexicographie arabe, on est frappé par un paradoxe patent :

Nous sommes d'un côté en présence d'une histoire lexicographique et lexicologique fabuleuse qui surprend par la maturité et l'originalité du premier dictionnaire que la tradition attribue à El-Khalil (2eS.H; 8eS.J.)[1] ainsi que par la quantité, la qualité et la variété des lexiques et des dictionnaires conservés. Mais d'un autre côté, une tradition aussi riche n'a pas abouti jusqu'à nos jours à la confection ni d'un dictionnaire proprement étymologique, ni d'un dictionnaire historique. Cette lacune de taille représente à l'heure actuelle le point le plus faible de la lexicographie arabe au point de laisser à l'observateur l'impression que les lexicographes arabes, tant anciens que modernes se soucient moins de conserver la mémoire des mots de leur langue que de les consigner et de les expliquer tout simplement, répondant ainsi à un besoin essentiellement didactique. On pourrait donc légitimement se poser la question de savoir pourquoi cette carence. Y a-t-il eu des tentatives et pourquoi n'auraient-elles pas abouti ? Existe-t-il dans la tradition lexicographique arabe des indices d'une vision historique du vocabulaire.? En fait, les indices ne sont pas totalement absents. Il suffit d'examiner les termes utilisés dans le classement du vocabulaire arabe pour se rendre compte de l'existence d'une «conscience» historique du vocabulaire. Il est en effet de tradition depuis les premiers traités de linguistique arabe de distinguer dans le vocabulaire :

- الفصيح [al-fasi:**h**] «le bon usage» conforme à la norme représentée notamment par le Coran, les dits du prophète Mohamed et la poésie classique (deux à quatre siècles après l'avènement de l'Islam selon qu'on est en zone citadine ou bédouine).

- الغريب [al-ɣari:b] «étrange», «hors usage», désignant en particulier des mots vieillis et sortis de l'usage ou même des expressions et des mots idiolectaux comme ceux attribués au prophète.

---

[1] *Al-ʕajn,* du nom de la première lettre d'après son classement. Ce dictionnaire étant classé selon l'articulation des phonèmes (laryngales ---> labiales). Toutes les combinaisons des phonèmes sont envisagées; ce qui permit à l'auteur de distinguer المستعمل [al- mustaʕmal] (unités faisant effectivement partie du système de la langue et المهمل [al-muhmal], (unités non utilisées par le système, mais représentant un potentiel latent et disponible).

- الدجيل [ad-daχi:l] «intrus», désignant les emprunts qui sont souvent perçus comme des xénismes mais qui peuvent accéder au statut de *fasi:**h*** s'ils sont intégrés dans des schèmes arabes leur permettant de fonctionner normalement dans le système dérivatif arabe; dans ce cas, ils sont qualifiés de معرب [muʔarrab] «arabisés»; ils restent néanmoins *fasi:**h*** de seconde zone.

- المولد [al-muwallad] «généré» et المحدث [al-muhdaθ] «néologisme» désignent les mots créés après la fin de l'ère de la fasa:ha pour répondre aux besoins nouveaux d'une civilisation arabo-islamique en expansion.

On peut ajouter العامى «populaire», qui désigne soit des néologismes populaires, soit des mots arabes «corrompus» phonétiquement, morphologiquement ou sémantiquement.

Cette catégorie peut aller jusqu'au vulgaire السوقي [as-su:qij]. Toutes ces catégories qui se classent en marge du «bon usage» sont quelquefois réunies sous le terme عجمة [ʔuʒma] ou أعجمى [ʔaʔʒamij] «étranger, non arabe» ou «non conforme à la norme arabe».

D'autres termes sont utilisés comme مذموم؛ رديء؛ ضعيف؛ متروك etc., qui n'ajoutent aux diverses variantes que de simples nuances. Il y a là bien évidemment une conscience historique du vocabulaire englobant le vieillissement, la néologie et le contact des langues. Il est cependant paradoxal de constater que ce ne sont pas les linguistes arabes anciens qui ont formulé clairement cette conscience mais les penseurs, représentés brillamment sur ce point précis par l'historien sociologue Ibn Khaldoun (8^{e}S.H.;14^{e}S.J.) qui considère dans ses fameux «Prolégomènes»[2] l'arabe citadin et l'arabe bédouin de son temps comme deux langues distinctes, différentes de l'arabe classique du temps du Prophète. Chacune de ces langues, fonctionne selon un système propre qu'il appelle ملكة [Malaka] et qu'on pourrait aujourd'hui traduire sans trop forcer par «compétence linguistique».

Pour répondre à la question : pourquoi les lexicographes arabes n'ont-ils pas poussé la réflexion à un niveau leur permettant de tirer profit de cette conscience historique et de jeter les bases d'une lexicographie historique ?, il est de tradition d'expliquer ce fait par le facteur religieux; le Coran étant un texte sacré, il est le modèle de la *fasa:**ha*** par excellence. Étant un texte atemporel, transhistorique, son support linguistique, l'arabe, est donc perçu comme en dehors du temps, échappant à l'historicité. Cette transcendance continue plus ou moins inconsciemment à hanter les lexicographes arabes jusqu'à nos jours, les empêchant de saisir la mémoire des mots dans sa dynamique.

Mais malgré cela, des tentatives plus ou moins timides ont été faites. Le rôle de pionnier dans ce domaine revient à un orientaliste allemand, Auguste Fischer (1865-1949), qui a le premier tenté d'élaborer un dictionnaire historique de l'arabe. L'idée a commencé à germer au milieu des années trente, à la suite de la création de l'Académie arabe du Caire, lorsque Fischer, membre de cette Académie, lui a présenté son projet; l'accord lui fut donné en 1938.

Ce travail avait l'ambition de couvrir cinq siècles (4^{e} → 9^{e}S.J.); 264 sources ont été à cette fin compulsées. Malheureusement, ce travail, qui a été partiellement relayé après la disparition de son promoteur, n'a pas abouti, faute de se voir transformé en une oeuvre de longue haleine, avec des structures et des moyens stables; cet échec traduit

---

2 *al-muqaddima* , éd. Égypte, S.d. (ch.38,39,40,41), p. 554-562.

l'absence de conviction ou de volonté surtout politique à travers les structures théoriquement habilitées à entreprendre un travail de ce genre. Nous pensons en particulier aux Académies, à l'Alecso et à l'Union des universités arabes.

Pour illustrer cette carence, nous avons entrepris d'étudier le traitement réservé par les dictionnaires arabes à six mots du vocabulaire : quatre concernent les points cardinaux et deux, les deux dernières religions monothéistes, le Christianisme et l'Islam. Nous verrons pourquoi cette troisième paire n'est pas sans liens avec les deux autres qui forment un groupe de mots courants, actuels, très polysémiques et lourdement connotés tant en arabe qu'en français. Nous partirons des définitions données par le لسان العرب [lisan el-ʔarab](L), dictionnaire quasi-encyclopédique dû à Ibn Mandûr (+1311 J.C.), que nous comparerons à celles données par المنجد [el-munʒid] (M1=éd 1906 et M2=éd 1986) et المحيط [al-muhi:t] (Mht), le dernier né des dictionnaires arabes modernes[3]. Nous comparerons ces définitions à celles données par le *GLLF* (éd. 1971) et *le Robert* (éd. 1985) en vue de dégager la part réservée à la mémoire de ces mots dans les descriptions lexicographiques respectives.

1. شرق [ʃarq] «orient»

Mis à part le sens étymologique qu'évoque dans **L** la racine trilitère [ʃ rq] à savoir «fendre, éclater...», le sens principal attesté de *ʃarq*, nom verbal substantivé est «lieu où le soleil se lève» avec une variante morphologique sur le schème du nom de lieu مشرق [maʃ riq] **M1** n'ajoute presque rien, tandis que **M2** en affine la formulation, l'étendant au point cardinal de l'est. **Mht** ajoute le sens plus moderne de «régions et pays de l'est». Pour ce dernier sens, le *GLLF* avance la date 1080 et les dates : 1949 pour *Proche-Orient*, début XX^{e} siècle pour *Moyen-Orient,* et 1968 pour *Extrême-Orient.*

	GLLF	Robert	Muhit	M2	M1	L
I-1- Partie du ciel, côté de l'horizon où semble se lever le soleil.	+	+	+	+	+	+
2-Vx ou litt. = direction de l'orient, pt cardinal de l'est.	+	+	+	+	+	
3-Vx ou litt. = direction, orientation	+	+				
4- Fig. Féerie de lumière.	+					
5- Class. et litt. = commencement brillant, heureux.	+					
6- Reflet des perles.	+	+				
7- Fig. Couleur évoquant (6)	+	+				
II-1- Zone ou ensemble des pays de l'est, P-0, M-0, E-0.	+	+	+			
2- Habitants et nations de l'orient.	+	+				
III- Dans la franc-maçonnerie, loges de province.		+				

Tableau 1 : Orient - شرق [ʃarq]

[3] Éd. Beyrouth, Paris, 1993, 3 vol.

Il est à remarquer que l'arabe ne distingue pas **orient** et **est**, qui n'ont pas forcément la même distribution en français alors que les deux variantes morphologiques arabes ont une distribution pratiquement identique.

En outre, les cases qui semblent vides en arabe, par comparaison aux sens français, sont en partie remplies par d'autres formes dérivées. Notons enfin que le sens politique né de la guerre froide n'est pas explicitée dans le dictionnaire français et est absent dans les dictionnaires arabes. Mais l'édition 92 du *Petit Robert* relève ce nouveau sens dans **est**.

Remarquons enfin que le dictionnaire arabe le plus récent, **Mht**, qui donne comme premier sens à شرق [ʃ arq], «soleil» (par métonymie), ne suit ni l'ordre logique ni l'ordre chronologique. Il ressort de cette comparaison que les dictionnaires français fournissent des descriptions beaucoup plus détaillées et datées.

2. غرب [γarb] «occident»

Le sens principal qu'évoque la racine trilitère [γrb] attesté dans tous les dictionnaires arabes est «limite», lié dans **L** aux sens «départ, éloignement, absence, disparition...». Le sens «limite» engendre «tranchant» s'agissant d'armes blanches. Celui d'«absence» engendre «obscurité». La forme غرب - γarb, nom verbal substantivé, avec sa variante morphologique de nom de lieu مغرب [maγrib], signifie « lieu où le soleil se couche». **M2** précise cette définition en l'étendant au point cardinal de l'ouest. **Mht** ajoute le sens plus moderne de «pays situés à l'ouest» qu'il place en deuxième position au lieu de le placer en dernier lieu conformément à la chronologie d'évidence. Pour ce même sens, le *GLLF* date le sens «Partie ouest du continent européen», v.1119, puisqu'on parlait déjà *d'Empire d'occident* et *d'Église d'occident* par opposition à *Empire d'orient* et *Église d'orient*.

	GLLF	Robert	Muhit	M2	M1	L
1- Un des points cardinaux situé du côté de l'horizon où le soleil se couche.	+	+	+	+	+	
2- Partie ouest de l'Europe + ensemble des pays et des peuples qui l'occupent.	+	+				
3- Civil, culture des peuples de l'E.occidentale.	+					
4- Ensemble des États membres du pacte de l'Atlantique Nord.	+	+	+			
5- Dans une loge maçonnique, côté où se tiennent les surveillants	+					
6- Class. et fig. = déclin, ruine, catastrophe.	+					

Tableau 2 : Occident غرب [γarb]

Remarquons là encore que l'arabe ne distingue pas occident et ouest. On peut appliquer ici la même comparaison effectuée à propos d'*orient*. Cependant, le sens politique né de la guerre froide se retrouve dans les dictionnaires français : ensemble des États membres du pacte de l'Atlantique Nord, par opposition aux États de l'est de l'Europe et à ceux d'Asie, daté par le *GLLF,* 1959. Ce sens est ici explicite car *occident* peut se substituer à *Bloc de l'ouest* ou *pacte de l'Atlantique Nord* dans le discours politique moderne alors que *orient* ne se substitue pas à *Bloc de l'Est* ou *pacte de Varsovie.*

Ce sens, fréquent dans la littérature politique et médiatique arabe moderne, n'est pas attesté dans les dictionnaires arabes les plus récents.

Il ressort de ce qui précède que les dictionnaires arabes couvrent beaucoup moins de sens et ne respectent pas l'ordre chronologique.

3. شمال [ʃama:l] «nord»

Il est frappant de constater que **L** ne définit pas شمال mais l'évoque indirectement à partir d'un sens dérivé «vent du nord», que nous retrouvons en français daté par le *GLLF* début du XIII^{e} siècle.

L définit ce mot «vent soufflant du côté du pôle», قطب, employé seul, ne désigne que le pôle Nord. On y trouve quelques considérations astrologiques sur le point de départ de ce vent. On pourrait chercher la raison de ce paradoxe dans l'ambiguïté résultant de l'existence d'une variante en *i* [ʃimâl] désignant à la fois, le vent, le nord géographique et la gauche.

Le nord étant situé à la gauche de celui qui se place en face du soleil levant, ainsi [ʃa/imâl] s'oppose-t-il d'un côté à جنوب [ʒanu:b] «sud» et d'un autre côté à يمين [jami:n] «droite».

M1 et **M2** continuent à reproduire cette ambiguïté, alors que **Mht** consigne la spécialisation consacrée par l'usage : [ʃama:l] désignant le point cardinal et tout ce qui s'ensuit et [ʃima:l] la gauche géographique avec les mêmes connotations de mauvais augure que nous retrouvons dans *sinistre*.

Mht donne pratiquement tous les sens que nous trouvons en français mis à part le sens lié à la franc-maçonnerie. On y trouve également le sens récent de «pays développés et industrialisés» par opposition à جنوب [ʒanu:b] «pays pauvres» qu'on ne rencontre même pas dans l'édition 1992 du *Petit Robert.*

	GLLF	Robert	Muhit	M2	M1	L
1- Celui des 4 points card. qui correspond à la direction marquée par l'étoile Polaire	+	+	+			
+ le Nord = le vent du nord.	+	+	+	+	+	+
2- Lieu au Nord d'un point > (Nord)	+	+	+			
+ région du globe du côté nord		+				
3- Franc-maçonnerie = côté gauche en entrant en loge	+		+			
4- ≠Midi		+				
+ Pays industriels, développés.			+			

Tableau 3 : Nord شمال [ʃama:l]

4. جنوب [ʒanu:b] «sud»

L ne parle que du vent du sud, sec et chaud, avec quelques précisions astrologiques relatives à son point de départ. Il est paradoxal de remarquer que ce vent brûlant est considéré de bon augure parce qu'il souffle du يمين [Jami:n] «droite» (le sens de toutes les formes dérivées de la racine [jmn] évoque le bon augure, contrairement à celui de [ʃ ml] et sa variante étymologique [ʃ ?m]. **M1** et **M2** ajoutent à cela le sens géographique «point opposé au point nord».

Remarquons que **Mht** ne donne pas ici le sens «pays pauvres ou en développement» qu'il a cité par opposition dans [ʃ ama:l] «nord», probablement parce qu'il ne l'a pas introduit comme sens autonome mais comme exemple « حوار الشمال والجنوب » «dialogue Nord-Sud».

	GLLF	Robert	Muhit	M2	M1	L
1- Celui des 4 pts cardinaux corresp. à la direction de l'étoile polaire.	+	+	+	+	+	
2- Contrée située ds l'hémisphère Sud.	+	+				
3- Région située au sud dans un ensemble géographique.	+	+				
4- Ensemble des habitants d'une région sud.	+					
5- Adj.	+					
+ midi (de la France)		+				
+ vent du Sud		(+)	+	+	+	+

Tableau 4 : Sud جنوب [ʒanu:b]

Il ressort de l'opposition N-S que les tentatives d'actualisation existent comme le prouve **Mht** mais demeurent néanmoins non systématiques, puisque ce dictionnaire ne mentionne le nouveau sens «pays riches» et «pauvres» que dans l'article شمال [ʃama:l], et peu élaborées, puisqu'il ne lui accorde pas le statut de définition; se contentant du statut d'exemple.

5. إسلام [isla:m]

Nous retrouvons le sens étymologique de ce terme qui est un nom verbal substantivé «soumission, résignation à la volonté de Dieu» dans tous les dictionnaires arabes, ainsi que dans les dictionnaires français. Tous, excepté **L**, donnent bien évidemment le sens «religion des musulmans » et le sens dérivé «peuples qui pratiquent l'Islam et civilisation qui les caractérise». **Mht** évoque ce dernier sens à travers la définition de l'histoire musulmane; «histoire des peuples qui ont embrassé l'Islam». Quant à **L**, le sens «religion...» y est implicite à travers la définition non pas du concept mais du nom verbal, «le fait d'embrasser l'Islam».

	GLLF	Robert	Muhit	M2	M1	L
0- Soumission, Résignation à la volonté de Dieu.	+	+	+	+	+	+
1- Religion des Musulmans. 2- Peuples qui pratiquent l'islam	+	+	+	+	+	+
+ civil . qui les caractérise	+	+	(+)	+	+	
+ militant islamiste → partisan de l'islamisme.		+				

Tableau 5 : Islam إسلام [ʔisla:m]

Il faut noter que le sens moderne d'Islam militant ou politique n'existe dans aucun dictionnaire arabe. En français ce sens se trouve de plus en plus sous la variante *Islamisme.* Mais seul le *Robert* relève ce nouveau sens sous la forme dérivée *islamiste* dans une citation du *Nouvel observateur* datée de 1984. Les dictionnaires arabes modernes semblent omettre ce sens en occultant l'opposition entre مسلم «musulman» et إسلامي «islamiste» pourtant si fréquente dans la production arabe moderne.

On a donc l'impression que cette carence dans les dictionnaires arabes est due à un blocage idéologique qui fait que seul ce qui est idéologiquement admis peut figurer dans les dictionnaires.

6. مسيحية [masi:hijja] «christianisme»

Le terme مسيحية dans le sens de «christianisme» n'existe ni dans **L** ni dans **M1** et **M2** pourtant élaborés par des pères. On le retrouve dans **Mht** comme synonyme de نصرانية, ce dernier terme qui veut dire littéralement «Nazaréen» se trouve dans le Coran sous la forme de نصارى [nasa:ra:] «chrétiens nazaréens» car les chrétiens les plus connus à l'époque

étaient surtout les nestoriens, les jacobites et les sabéens. Ce nom, comme le précise le *GLLF* est celui «donné aux premiers chrétiens d'après le nom de la ville de Nazareth en Palestine», et c'est encore ce terme qu'on retrouve dans les parlers arabes actuels. Ainsi, مسيحية est-il un terme plus récent utilisé fréquemment par les lettrés pour désigner le christianisme. Il est donc paradoxal de constater que les dictionnaires arabes, même les plus récents ne disent pas mot de cette évolution pourtant frappante vers une forme de spécialisation :

نصرانية : usage ancien d'origine religieuse qui se perpétue dans la tradition populaire sous la forme du pluriel نصارى

مسيحية : usage plus récent qui se généralise dans les écrits modernes.

	GLLF	Robert	Muhit	M2	M1	L
1- Religion (doctrine de J.C.).	+	+	نصرانية+	+	+	
2- Foi chrétienne de qn.	+		مسيحية+			

Tableau 6 : Christianisme مسيحية [masi:hijja] / نصرانية [nasra:nijja]

CONCLUSION

De l'examen de ces six termes et de leurs variantes, en arabe et en français, il est possible de dégager les principales remarques suivantes :

1- En français, les termes exprimant les points cardinaux s'organisent en paires qui s'opposent nettement : N-S; Ori-Occ (ou E-O).

Par contre, cette opposition nette géographiquement ne l'est pas linguistiquement en arabe. En effet, le même terme [ʃima:l] désigne dans les dictionnaires arabes le nord et la gauche qui est définie par rapport à l'orient et non par rapport au nord qui sert d'habitude de référence et de repaire.

Ces dictionnaires continuent à entretenir l'ambiguïté malgré la tendance très nette dans l'usage vers la spécialisation des variantes : [ʃama:l] «Nord» et [ʃima:l] «gauche».

2- Le sens politique né de ces oppositions dans l'usage moderne est généralement absent dans les dictionnaires arabes : Est-Ouest (guerre froide) et Nord-Sud (post-guerre froide).

Cependant l'opposition Nord-Sud, développée après la guerre froide, n'existe pas encore dans les dictionnaires français mais sera sûrement retenue dans les éditions à venir.

3- L'acquisition par le mot *Islam* et en particulier par sa variante *Islamisme* d'une connotation politique, qui l'oppose dans les écrits actuels non pas à *Christianisme* mais à *Occident* n'est pas mentionnée dans les dictionnaires arabes; elle l'est partiellement dans les dictionnaires français. Des livres et des colloques portent ce titre apparemment paradoxal «Islam et Occident». C'est la dimension politique récente de cette relation qui explique ou justifie la mise en opposition de ces deux entités.

4- La complexité des relations qui caractérisent les mots étudiés et leur caractère dynamique (nouvelle relation Islam-Occident) prouvent que les dictionnaires arabes ont

besoin d'évoluer dans leur approche afin de pallier leurs insuffisances et rattraper un retard patent.

Partant de ces constatations qui ont porté sur un échantillon très limité, nous avançons quelques propositions pour une méthodologie de la dimension historique du dictionnaire arabe :

1- Jusque là, la langue arabe, pour des raisons religieuses et idéologiques, a sombré dans une a-historicité bloquant toute initiative rendant compte de son évolution naturelle. Ce blocage était d'autant plus important que toutes les tentatives de descriptions historiques ont tourné court parce qu'elles ont voulu partir du Coran et de la littérature préislamique; d'où l'importance du point suivant : au lieu de suivre le cours du temps en faisant des découpages plutôt arbitraires, ne vaudrait-il pas mieux suivre le chemin inverse en partant des synchronies actuelles où l'on dispose de matériaux fiables, faciles à observer et à analyser, évitant ainsi les conjectures souvent liées à la méthode opposée et remonter petit à petit le temps vers les origines. C'est à partir de l'observable et du tangible qu'on pourrait façonner des outils méthodologiques fiables, capables de dépasser le retard accumulé et d'éclairer la mémoire obscurcie par tant de retard. Ainsi redonnerait-on à la langue une historicité qui lui revient de droit.

2- Il serait souhaitable que l'on procède par des descriptions portant sur des champs lexicaux où il serait plus aisé de dégager les valeurs respectives des items lexicaux et de reconstituer les structures sémantico- logiques organisant leur polysémie.

Les domaines de la recherche terminologique fondée sur l'évaluation du patrimoine lexical, la collecte des usages en cours dans les différentes régions et sur les possibilités offertes par le système, présentent aux lexicographes des champs bien structurés, facilitant ainsi cette recherche à rebours.

3- Si l'historicité est réhabilitée et les blocages idéologiques et religieux dépassés, les tentatives hésitantes d'actualisation observées dans certains dictionnaires récents pourrait être renforcées et les acquis méthodologiques de la lexicographie moderne bien mis à profit pour raviver et entretenir la mémoire des mots de l'arabe.

Remerciements

Je remercie mon ami Salah Mejri qui, après lecture et discussion de ce travail, a contribué à enrichir ses conclusions.

DÉNOTATION ET PROBLÈMES DE POLYSÉMIE DANS L'ÉLABORATION D'UN DICTIONNAIRE ÉLECTRONIQUE FRANÇAIS-ARABE

Bassam BARAKÉ

Université Libanaise, Tripoli, LIBAN

Au cours de mon travail sur un dictionnaire français-arabe[1], j'ai affronté maintes fois des problèmes de polysémie dans l'une et l'autre langue. Les cas de mots monosémiques ne sont pas très fréquents et, sur le plan de l'analyse contrastive, ils ne posent pas de problèmes majeurs, sémantiques ou syntaxiques[2]. Mon propos ici se limite à la présentation des problèmes de dénotation et de polysémie dans une perspective de comparaison entre le français et l'arabe, cela à partir de l'analyse de quelques exemples, analyse dont le but est de présenter certains types de différences linguistiques et culturelles entre les deux langues et de délimiter les traits sémantiques, syntaxiques et socio-linguistiques qu'il faudrait prendre en compte dans l'élaboration d'un dictionnaire français-arabe.

* * *

Quelques remarques préliminaires

Le dictionnaire dont il s'agit est conçu dans l'optique du décodage de la langue source, le français, et de l'encodage en langue cible, l'arabe (dictionnaire de version français-arabe, pour usagers de langue arabe).

Nous adoptons l'arabe classique tel qu'il est actualisé de nos jours. Le canevas de départ est un dictionnaire du français contemporain et les équivalents arabes appartiennent généralement à l'arabe contemporain. Comme tout dictionnaire, mono- ou bilingue, est nécessairement dérivé d'une base de données, la tâche est relativement facile pour le français (les travaux dictionnairiques, monolingues ou encyclopédiques, sont, dans notre cas et pour l'usage actuel de la langue, plus développés dans la langue source que dans la langue cible).

---

[1] Il s'agit du *Dictionnaire Larousse Français - Arabe* à paraître en décembre 1997.

[2] Les problèmes rencontrés avec les vocables monosémiques concernent moins les équivalences entre le français et l'arabe que les spécificités culturelles de la langue arabe (homonymie, synonymie, polylexie, etc.).

Si l'équivalence peut être parfaite lorsqu'il s'agit de termes techniques à caractère monosémique, certains cas de monosémie française trouvent une panoplie de «synonymes» arabes. Cela est dû à deux faits : d'une part, la communication entre les chercheurs et les écrivains arabes est difficile, pour des raisons d'ordre géographique et/ou politiques; d'autre part, la notion est généralement d'origine étrangère (une science ou un concept nouveau) et les locuteurs la désignent en arabe avec des termes qui reflètent la compréhension et l'origine «scientifique» de chacun (ex : le terme «linguistique» a connu quelque cinq équivalents arabes).

Vu les progrès réalisés par les moyens informatiques, principalement en ce qui concerne les supports, les outils, la mémoire et les possibilités de stockage et de traitement, nous pouvons envisager d'établir une base de données (sur CD-ROM ou sur un disque dur) dont le volume, les instruments d'utilisation et les voies de consultation dépassent de loin ce que nous pouvons faire sur le papier. Ainsi, les exemples que nous allons présenter sont traités dans la perspective de l'élaboration d'un dictionnaire électronique français-arabe.

Le dictionnaire bilingue, qu'il soit électronique ou sur papier, établit une relation entre les vocables de deux langues différentes. Il n'a pas à s'occuper du référentiel, dans le sens que sa destinée n'est pas d'être un dictionnaire encyclopédique; il confronte deux lexiques et deux découpages du monde, mais s'il s'intéresse d'abord et avant tout au découpage en signes, il ne pourrait pas négliger le côté référentiel, surtout lorsqu'il y a divergence culturelle entre les deux langues.

* * *

Prenons les deux vocables suivants :

fleuve et *rivière*

Le *Petit Robert* donne (entre autres) les définitions suivantes :

fleuve : 1. Grande rivière
rivière : I. 1.Cours d'eau naturelle de moyenne importance
Maupassant : «La rivière [l'Oued Saïda], fleuve là-bas, ruisseau pour nous, s'agite dans les pierres sous les grands arbustes épanouis».

Nous partons du principe fondamental qu'un dictionnaire bilingue français-arabe, sur papier ou électronique, est censé donner un synonyme en arabe de l'adresse en langue française. Mais, si nous prenons en considération le fait qu'une langue a pour rôle de décrire le réel, de le découper, nous aboutissons à la conclusion que les problèmes d'équivalence se posent sur deux plans : le plan de la réalité (niveau socio-culturel) et le plan de la langue. D'une culture à l'autre, il y a des différences de «vision» de la réalité et, par conséquent, des différences de vocables la désignant.

Dans les dictionnaires français-arabe, il existe un seul équivalent arabe du mot «fleuve», c'est «*nahr*».

Ce même mot arabe est aussi l'équivalent du mot «rivière». Celui-ci partage avec «ruisseau» un autre équivalent arabe : «*jadwal*».

fleuve	*nahr*
rivière	*nahr* et *jadwal*
ruisseau	*jadwal*

Ce qui nous importe ici c'est que là où le locuteur français dispose de deux mots (fleuve et rivière) pour désigner deux états d'un cours d'eau (de moyenne ou de grande importance), le locuteur arabe ne dispose que d'un seul mot. Les Égyptiens prennent conscience de l'incapacité du mot à désigner la majesté et l'immensité du Nil et l'appellent «mer du Nil».

Le recours aux «champs» syntaxique et sémantique rendra-t-il compte de cette différence (différence relative à une polylexie française) ? Ici, les traits syntactico-sémantiques sont indispensables pour l'analyse du lexique français (pour définir le signifié du vocable et déterminer son comportement syntaxique), mais ils ne sont pas suffisants pour décrire la polysémie de leur synonyme arabe. Dans un dictionnaire français-arabe (sur papier ou électronique), il faudrait rendre compte d'une telle différence entre le français et l'arabe et marquer l'équivalent de «fleuve» du trait «grand» et celui du mot «rivière» du trait «moyen». Cette différence n'étant pas due à l'emploi syntaxique dans l'une ou l'autre langue, le recours à la réalité socio-linguistique s'avère indispensable pour rendre compte des idiosyncrasies et des différences culturelles liés aux deux langues. «L'importance des écarts culturels nécessite, selon A. Rey, [...] une explicitation des différences, chaque fois qu'une simple équivalence lexicale ou idiomatique ne suffit pas» (Rey 1991 : 2865).

Prenons un exemple de polysémie du vocable français :

Campagne : I. Vaste étendue de pays découvert.
II. Étendue de terrain (zone de combat)
État de guerre, les combats
Travaux, action de communication limités à une période déterminée (campagne de fouilles archéologiques, campagne de pêche, campagne électorale, campagne de publicité)

La polysémie du mot français est présentée par une polylexie arabe. Ici, plusieurs lexèmes arabes correspondent à plusieurs sens d'un seul lexème français. C'est donc la polylexie de la langue cible qui articule la polysémie de la langue source. Chacune de ces deux acceptions est traduite par un mot différent (*rîf* et *hamlah*). La différence de lexème équivalent suffit à rendre compte de la polysémie de la source lorsqu'il s'agit d'un dictionnaire de version (usager arabe). La détermination du champ syntaxique et de celui des classes d'objets (pour l'équivalent arabe) est indispensable pour l'usager français (thème).

À la suite des travaux de Gaston Gross, Brigitte Lépinette définit la description lexicographique par l'analyse des comportements syntaxiques. Dans le cadre du traitement automatisé du langage, elle préfère distinguer, dans un premier temps, les différents comportements syntaxiques du terme.

> «Dans un second temps, en partant de l'hypothèse selon laquelle il y a souvent homogénéité entre les comportements syntaxiques et sémantiques (autrement dit, une construction et un environnement lexical donnés correspondent à un lexème et à un seul), il paraît possible d'organiser, en regard de ces unités en LS déterminées par leur comportement syntaxique, la présentation des équivalents traductifs de ces dernières en LC» (Lépinette 1997 : 60).

En effet, le recours à la notion de traits syntaxiques s'avère un moyen efficace pour lever les ambiguïtés liées aux phénomènes polysémiques. Prenons les vocables suivants : *chasseur* et *pêcheur*; Les lexèmes «chasseur» et «pêcheur» se traduisent tous les deux par un seul mot arabe «*sayyâd*», ainsi que les verbes chasser et pêcher (*sâda*). La polysémie du vocable arabe (par rapport à la langue source) ne rend pas compte de tous les sèmes dénotatifs de chacun des deux lexèmes français. Le dernier-né des dictionnaires français-arabe, celui de Abdelnour, réagit à ce fait comme s'il considérait que l'équivalent de chasseur est l'hyperonyme de celui de pêcheur et propose pour le premier «*sayyâd*» et pour le deuxième «*sayyâdu samak*», c'est-à-dire qu'il ajoute au premier le mot «poisson» pour rendre compte que le pêcheur est un «chasseur» de poissons. Cependant, en arabe, le verbe *sâda* (et le substantif de la même racine *sayyâd*) s'appliquent aussi bien au premier lexème qu'au second. Abdelnour a eu recours à la collocation syntaxique pour traduire la spécificité et la différence d'un des deux lexèmes français. Mais la solution qu'il propose, d'une part, ne permet pas de rendre compte du contenu lexical de «chasseur» (l'usager arabe peut comprendre le mot comme étant un synonyme de pêcheur) et, d'autre part, n'indique pas leur classe d'objets ni leur domaine. La faute n'incombe pas à l'auteur, c'est la propriété de tous les dictionnaires sur papier. En effet, les possibilités offertes par l'informatique permettent de doter chacun des lexèmes des traits indispensables à la reconnaissance de son contenu sémantique et de ses traits distinctifs.

Passons maintenant à un problème lié à la connotation de l'équivalent arabe. En effet, les écarts linguistiques, qui font l'objet du dictionnaire bilingue dans son ensemble et concernent principalement le plan sémantique et syntaxique, traduisent parfois des écarts culturels plus ou moins importants, qui concernent non seulement le plan référentiel (désignation), mais surtout le plan connotatif (culture).

Dans un poème d'Apollinaire nous lisons ceci :

> Le mai le joli mai en barque sur le Rhin
> Des dames regardaient du haut de la montagne
> Vous êtes si jolies mais la barque s'éloigne
> Qui donc a fait pleurer les saules riverains
> Or des vergers fleuris se figeaient en arrière
> ...

Si nous voulons traduire ce poème en arabe la première difficulté que nous rencontrons est le vocable «mai». Le dénoté est le mois, cela est élémentaire. Mais l'équivalent arabe n'est pas aussi simple. Nous avons deux mots : *'Ayyâr* et *Mayo*. Or, si tous les deux mots rendent compte du sens lexical, aucun ne traduit la portée significative de la phrase. Le poète utilise un réseau de mots qui expriment le bonheur de vivre et la lumière éclatante qui accompagne l'arrivée du printemps. Notre vocable renvoie à «dames», «jolies», «fleuris», etc. Le «mai» ne saurait donc se traduire que par un mot qui rende compte de ces

significations connotant la joie et la lumière. Il s'agit de «*nawwâr*» dont la racine (*nûr*) exprime à la fois la fête, la lumière et le rayonnement. Dans l'élaboration d'un dictionnaire français-arabe général, et pour une meilleure utilisation du vocable, il faudrait tenir compte des faits suivants :

- La polylexie arabe
- Le niveau de langue du vocable
- les sens connotatifs liés au terme de la langue source et à son équivalent (quitte à prendre en considération, dans ce dernier cas, le facteur diachronique et la mémoire culturelle des mots.

Comme le souligne Alain Duval, «L'équivalence parfaite exige un même niveau de dénotation (c'est-à-dire la référence à un même élément de la réalité extérieure), et un même niveau de connotation, c'est-à-dire le même réseau d'associations culturelles liées au terme dans les deux langues» (Duval 1991 : 2818-2819).

CONCLUSIONS

L'analyse contrastive qui se veut le premier pas vers l'élaboration d'un dictionnaire bilingue doit prendre en considération le sens dans lequel celui-ci est constitué (langue maternelle ⟶ langue étrangère ou l'inverse), le public (l'usager) auquel il est destiné et l'usage qu'il en fera. Dans cette perspective, il serait possible de prendre comme point de départ le profil sémantico-syntaxique fourni par le dictionnaire monolingue (des deux langues), en l'adaptant en fonction des résultats de l'étude différentielle des deux langues.

La différence entre un dictionnaire bilingue et un dictionnaire monolingue réside dans le fait que le bilingue est censé tenir compte des différences structurelles qui peuvent exister entre les deux langues, aussi bien sur le plan de la grammaire, de la dénotation et de la syntaxe que sur le plan du découpage du monde, de la culture et des structures sociales. Le travail d'élaboration doit passer par une étape d'analyse contrastive dont les résultats doivent être incorporés, d'une façon ou d'une autre, dans le dictionnaire en question.

L'analyse contrastive nous montre que l'équivalence n'est pas réciproque et que la symétrie entre les deux langues n'est possible que dans des cas rares.

L'élaboration d'un dictionnaire bilingue devrait passer par l'analyse contrastive des deux langues à partir de la comparaison de deux bases de données. La phrase constituera l'unité de base de l'analyse de chacune des deux langues, et l'introduction d'un champ sémantique qui puisse rendre compte des spécifités du signifié de l'équivalent (en comparaison avec la source) me paraît indispensable.

L'analyse contrastive de certains vocables de la paire de langues concernée révèle que la réalité dénotée par la langue source ne fait pas toujours partie de l'univers culturel des locuteurs de la langue cible, ou n'est pas reconnu en tant que tel par la majorité d'entre eux (comme nous l'avons vu avec l'exemple de «fleuve» et de «rivière»). Le réel dénoté n'existe que dans l'univers culturel et le lexique de la langue source et le vocable n'a de réalité que dans cette langue. C'est la conclusion à laquelle aboutit Alain Rey dans son analyse des contenus culturels dans les dictionnaires bilingues. Il soulève le problème du découpage référentiel d'univers culturellement différents.

«Les différences de découpage, écrit-il, proviennent a) de la nature des référents, b) des constructions conceptuelles effectuées par la culture à leur égard, c) de la prise en charge de ces conceptualisations par un usage d'une langue, lui même conditionné par les moyens — morphologiques, syntagmatiques, sémantiques... — qu'une langue met à la disposition d'une culture» (Rey 1991 : 2869).

Dans l'élaboration d'UN dictionnaire bilingue, l'analyse syntactico-sémantique fondée sur l'emploi est indispensable pour définir l'usage dans l'une et l'autre langue. Cependant, le recours aux champs conceptuel, culturel et socio-linguistique est lui aussi une nécessité.

RÉFÉRENCES

CADIOT, Pierre et Benoît HABERT (1997) : «Aux sources de la polysémie nominale», *Langue française*, Larousse, mars, n° 113.

DUVAL, Alain (1991) : «L'équivalence dans le dictionnaire bilingue», F. J. Hausmann et alii (dir), *Dictionnaires, Encyclopédie internationale de lexicographie*, Tome troisième, Berlin et New York, Walter de Gruyter, pp. 2817-2824.

LADMIRAL, Jean-René et Henri MESCHONNIC (dir) (1981) : «La traduction», *Langue française*, Larousse, septembre, n° 51.

LÉPINETTE, Brigitte (1997) : «Le rôle de la syntaxe dans la lexicographie bilingue», H. Béjoint et Ph. Thoiron (dir), *Les dictionnaires bilingues*, AUPELF-UREF/Duculot, pp. 53-69.

REY, Alain (1991) : «Divergences culturelles et dictionnaire bilingue», F. J. Hausmann et alii (dir), *Dictionnaires, Encyclopédie internationale de lexicographie*, Tome troisième, Berlin et New York, Walter de Gruyter, pp. 2865-2870.

UN MODÈLE HYBRIDE POUR L'EXTRACTION DES CONNAISSANCES : LE NUMÉRIQUE ET LE LINGUISTIQUE

Ismaïl BISKRI[1et3]; Jean-Guy MEUNIER[1];
Christophe JOUIS[2et3]

(1) *Laboratoire de l'ANalyse Cognitive de l'Information, Université du Québec à Montréal, Canada*; (2) *IDIST/CREDO, Université Charles de Gaulle-Lille 3, France*;
(3) *LALIC-CAMS, Université de la Sorbonne-Paris IV, France*

1. INTRODUCTION

De nos jours, un nombre croissant d'institutions accumulent très rapidement des quantités de documents qui ne sont souvent classés ou catégorisés que très sommairement. Très vite, les tâches de dépistage, d'exploration et de récupération de l'information présente dans ces textes, c'est-à-dire des «connaissances», deviennent extrêmement ardues, sinon impossibles. Pour y faire face, il devient nécessaire d'explorer de nouvelles approches d'aide à la lecture et à l'analyse de texte assistées par ordinateur (LATAO).

Du point de vue méthodologique, la question de l'extraction des connaissances dans les textes rencontre des difficultés épistémologiques sérieuses. En raison de sa nature sémiotique et langagière, le traitement informatique traditionnel d'un texte est de nature linguistique. Un texte est vu comme une suite de phrases qu'on doit soumettre à des analyseurs linguistiques. Cette approche semble tout à fait naturelle, elle correspond théoriquement au processus naturel de lecture d'un humain. Cependant, cette approche s'avère problématique dès lors qu'il s'agit d'une grande masse de données textuelles.

Dans ce cadre, le traitement d'un texte par ordinateur en appelle à des dépôts de connaissances préconstruites acquises via des enquêtes cognitives (analyse de protocole) auprès des experts ou puisées dans le répertoire encyclopédique du savoir partagé. Ceux-ci sont alors utilisés comme gabarit dans le dépistage et la reconnaissance. De plus, les systèmes experts qui opèrent dans ce domaine doivent être dotés des mécanismes habituels (moteur d'inférence, maintien de cohérences, tests de plausibilité, etc.) leur permettant d'effectuer des déductions et des tests d'hypothèses avec un haut niveau de confiance et de réussite. Les connaissances comportent des représentations d'objets, de propriétés, de relations d'événements et de situations propres à l'objet à traiter, en l'occurrence le contenu informationnel du texte. En possession de ce savoir, ce système informatique de type expert pourrait alors réussir à «comprendre» le texte et donc en extraire les connaissances. De nombreuses recherches ont d'ailleurs montré la nécessité d'avoir les connaissances de multiples niveaux (syntaxiques, psycholinguistiques, lexicales, sémantiques,

encyclopédiques, etc.) (Regoczei et al., 1988; Shaw & Gaines, 1988; Jacobs & Zernik, 1988; Moulin et Rousseau, 1990; Zarri, 1990).

Du point de vue de la lecture et de l'analyse de texte assistées par ordinateur (LATAO), le problème de l'extraction des connaissances d'un corpus textuel se présente de manière totalement différente. Il est en effet délicat de donner a priori à l'ordinateur, les connaissances que le texte avait pour fonction de transmettre sauf peut-être, pour celles qui sont de nature générale, encyclopédique ou technique. Dans le cadre de LATAO, la connaissance se trouve dans le texte lui-même et doit en être extraite. Et les techniques qui ont donné des résultats intéressants en IA sur de petits textes bien maîtrisés (scénario de restaurant, etc.) s'avèrent vite problématiques lorsqu'elles sont appliquées à des domaines dont on ignore en partie ou en totalité la teneur. Un texte contient normalement de nombreux énoncés originaux qui n'ont pas encore été lus et dont le contenu tant lexical, sémantique, qu'encyclopédique est inconnu au préalable par le lecteur, et qu'il découvrira dans le parcours du texte lui-même.

Le deuxième problème est de nature plus technique. Même si on possédait des analyseurs linguistiques raffinés et robustes pouvant décrire un texte selon ses diverses catégories linguistiques (morphologiques, syntaxiques, sémantiques, discursives) il faudrait prévoir que ce traitement prenne un certain temps. Dans la meilleure des situations, la technologie actuelle ne permet guère d'analyser des phrases en deçà de quelque 10 à 20 secondes par phrase. On peut imaginer le temps requis pour traiter des milliers de pages. La situation de LATAO ne permet pas ce type de traitement. Il faut modifier l'approche. Des stratégies, peut-être plus grossières dans leurs approches premières, permettent ultimement des extractions fines de connaissances. C'est dans cette perspective que nous explorons les approches par classification numérique et plus particulièrement les classifieurs de type connexionniste. Il nous semble que, dans le traitement de grande masse d'informations, il faut y aller comme en archéologie. Un bon archéologue ne commence pas directement sa fouille par le plus fin et le plus précis de ses outils. Au contraire, il commence sa recherche par un parcours général de son territoire. Il utilise pour ce faire des outils généraux (sonar, résonance magnétique, géomatique, etc.). Ce n'est qu'une fois qu'il a cerné le lieu potentiel des vestiges archéologiques qu'il en appelle à des outils plus fins. La pelle, la cuillère, la brosse, etc. Et ce n'est qu'à la fin qu'il prendra son microscope électronique. En d'autres termes deux grandes étapes sont nécessaires, une première étape utilisant un outil que nous dirons «bulldozer» pour classifier d'une manière grossière les données textuelles et ainsi permettre à un utilisateur de sélectionner dans une deuxième étape les parties du texte sur lesquels il veut extraire des connaissances d'une manière plus fine et ce au moyen de méthodes linguistiques.

2. STRATÉGIES NUMÉRIQUES

La littérature technique relative au traitement de l'information textuelle a montré qu'il était possible d'explorer des outils d'extraction des connaissances dans des textes (*data mining*). Or, l'extraction de connaissances peut être vue sous plusieurs angles. Dans notre perspective, elle n'est pas une «compréhension» du texte, ni une paraphrase, ni un rappel d'information, mais un processus de traitement classificatoire qui identifie des segments de textes qui contiennent un «même» type d'information. Autrement dit, l'extraction des connaissances est définie comme résultant d'une opération de classification fondée sur l'un ou l'autre critère d'équivalence.

Pour les chercheurs dans le domaine de LATAO, cette problématique n'est pas nouvelle. Dans la recherche antérieure, plusieurs techniques et méthodes ont déjà été proposées pour tenter d'organiser le contenu d'un texte en des configurations interprétables. Ces méthodes, souvent moins fines certes que les approches linguistiques et conceptuelles, n'en permettent pas moins un premier parcours général et robuste du texte. Elles sont en mesure, par exemple, d'identifier dans un corpus des classes ou des groupes de lexèmes qui entretiennent entre eux des associations dites de cooccurrence et donc de détecter leurs réseaux sémantiques. Et les recherches actuelles commencent d'ailleurs à les privilégier de plus en plus (Church et al., 1989; Lebart et Salem, 1988; Salton, 1988, etc.). Parmi les modèles les plus couramment utilisés, on trouve habituellement l'analyse des cooccurrences, l'analyse corrélationnelle, l'analyse en composante principale, l'analyse en groupe, l'analyse factorielle, l'analyse discriminante, etc. Malgré le succès qu'elles ont obtenu, on a dû constater que ces méthodes particulières posent deux problèmes importants. Premièrement, les modèles classiques ne peuvent traiter que des corpus stables. Toute modification du corpus exige une reprise de l'analyse numérique. Ceci devient un problème majeur dans des situations où le corpus est en constante modification (par exemple les reposoirs de l'autoroute électronique). Deuxièmement, les types de résultats qu'ils produisent ne sont pas sans problèmes théoriques. Ils posent des problèmes d'interprétation linguistique importants (Church et Hanks, 1990). Les associations des mots dans les classes ne sont pas toujours facilement interprétables. Pourtant, malgré leurs limites, ces approches ont été reconnues des plus utiles pour l'extraction des connaissances et plus particulièrement les connaissances terminologiques. D'une part, ces stratégies classificatoires permettent une immense économie de temps dans le parcours exploratoire d'un corpus, et à ce titre, elles sont incontournables lorsqu'on est confronté à de vastes corpus textuels. D'autre part, elles servent d'indices pour détecter rapidement certains liens sémantiques et textuels. Cependant, lorsqu'associées à des stratégies linguistiques plus fines et intégrées dans des systèmes hybrides (i.e., avec analyseurs linguistiques d'appoint), elles livrent une assistance précieuse pour des analyses globales. Elles permettent un premier déblaiement général du texte. Peuvent alors suivre des analyses plus fines.

Les recherches récentes permettent de penser qu'on peut améliorer ces techniques de classification de l'information. En effet, de nouveaux modèles classifieurs dits émergentistes commencent à être explorés pour ce type de tâche. Ils ont pour fondement théorique que le traitement «intelligent» de l'information est avant tout associatif et surtout adaptatif. Parmi ces modèles dits «de computation émergente» on distingue les modèles «génétiques», markoviens (Bouchaffra et Meunier, 1993) et surtout connexionnistes. Parmi ces derniers, on trouve une grande variété de modèles, entre autres, les modèles matriciels linéaires et non linéaires, les modèles thermodynamiques et les modèles basés tantôt sur la compétition, tantôt sur la rétropropagation, mais surtout sur des règles complexes d'activation et d'apprentissage (Kohonen, 1982). Les principaux avantages de ces modèles tiennent au fait que leur structure parallèle leur permet de satisfaire un ensemble de contraintes qui peuvent être faibles et même, dans certains cas, contradictoires et de généraliser leur comportement à des situations nouvelles (le filtrage), de détecter des régularités et ce, même en présence de bruit. Outre les propriétés de généralisation et de robustesse, la possibilité pour ces modèles de répondre par un état stable à un ensemble d'inputs variables repose sur une capacité interne de classification de l'information.

Cependant, tous ces modèles classifieurs émergentistes opèrent sur des données bien contrôlées et qui toutes doivent être présentes au début et tout au long du traitement. De plus, ils exigent souvent divers paramètres d'ajustement qui relèvent souvent d'une description statistique du domaine. Il s'ensuit que les résultats de classification obtenus sont valides pour autant qu'ils portent sur les données bien contrôlées où peu de modifications sont possibles. Si, après la période d'apprentissage, pour quelque raison que ce soit, les systèmes sont confrontés à des données qui n'étaient pas prévues dans les données de départ, ils auront tendance à les classer dans les prototypes déjà construits, donc à produire une sous-classification.

Or, le domaine dans lequel nous opérons, à savoir le texte, présente précisément ce type de problème. Chaque nouvelle page peut contenir des informations que le système peut ne jamais avoir rencontrées, et donc qu'il ne peut se permettre de classer dans ses prototypes antérieurement construits. Il faut donc, outre la dynamicité de l'apprentissage, un système qui soit aussi plastique.

3. LA GRAMMAIRE CATÉGORIELLE COMBINATOIRE APPLICATIVE DANS LE CADRE DE LA GRAMMAIRE APPLICATIVE ET COGNITIVE

La Grammaire Applicative et Cognitive (Desclés, 1990) postule trois niveaux de description des langues :

a- le niveau phénotypique (ou le phénotype) où sont représentées les caractéristiques particulières des langues naturelles (par exemple, l'ordre des mots, les cas morphologiques, etc.). Les expressions linguistiques de ce niveau sont des unités linguistiques concaténées, la concaténation est notée par : 'u_1-u_2-...-u_n';

b- le niveau génotypique (ou le génotype) où sont exprimés les invariants grammaticaux et les structures sous-jacentes aux énoncés du niveau phénotypique. Le niveau génotypique est structuré comme un langage formel appelé «Langage génotype»; il est décrit par une grammaire appelée «Grammaire applicative»;

c- le niveau cognitif où sont représentées les significations des prédicats lexicaux par des schèmes sémantico-cognitifs.

Les trois niveaux font appel à des formalismes applicatifs typés où l'opération d'application d'un opérateur à un opérande est considérée comme primitive. Les niveaux deux et trois s'expriment dans le formalisme de la logique combinatoire typée de Curry et Feys (1958). Cette logique fait appel à des opérateurs abstraits — appelés «combinateurs» — qui permettent de composer intrinsèquement des opérateurs plus élémentaires entre eux (Desclés, 1990). Les combinateurs sont associés à des règles d'introduction et d'élimination. Ceux que nous utiliserons dans cet article[1] sont **B**, $\mathbf{C}_*$, avec les règles d'élimination (ß-réduction) suivantes (U_1, U_2, U_3 sont des expressions applicatives typées) :

[1] Le combinateur $\mathbf{C}_*$ est souvent noté T.

$((\mathbf{B}\ U_1\ U_2)\ U_3) > (U_1\ (U_2\ U_3))$
$((\mathbf{C}_*\ U_1)\ U_2) > (U_2\ U_1)$

Le modèle de la Grammaire Catégorielle Combinatoire Applicative (GCCA) relie explicitement les expressions phénotypiques à leurs représentations sous-jacentes dans le génotype[2]. Le système consiste en :

(i) une analyse syntaxique des expressions concaténées du phénotype par une Grammaire Catégorielle Combinatoire;
(ii) une construction à partir du résultat de l'analyse syntaxique d'une interprétation sémantique fonctionnelle des expressions phénotypiques.

Les Grammaires Catégorielles assignent des catégories syntaxiques à chaque unité linguistique. Les catégories syntaxiques sont des types orientés engendrés à partir de types de base et de deux opérateurs constructifs '/' et '\'.

(i) N (syntagme nominal) et S (phrase) sont des types de base.
(ii) Si X et Y sont des types orientés alors X/Y et X\Y sont des types orientés[3].

Une unité linguistique u de type orienté X sera désigné par '[X : u]'.

Les deux règles d'application (avant et arrière) sont notées :

$$\frac{[X/Y : u_1]\ -\ [Y : u_2]}{[X : (u_1\ u_2)]}\ (>) \qquad ; \qquad \frac{[Y : u_1]\ -\ [X\backslash Y : u_2]}{[X : (u_2\ u_1)]}\ (<)$$

Les prémisses dans chaque règle sont des concaténations d'unités linguistiques à types orientés considérées comme étant des opérateurs ou des opérandes, la conséquence de chaque règle est une expression applicative avec un type orienté.

La Grammaire Catégorielle Combinatoire (Steedman, 1989) généralise les Grammaires Catégorielles classiques en introduisant des opérations de changement de type et des opérations de composition des types fonctionnels. Dans la GCCA les règles de la Grammaire Catégorielle Combinatoire de Steedman introduisent les combinateurs **B**, $\mathbf{C}_*$ dans la séquence syntagmatique. Cette introduction permet de passer d'une structure concaténée à une structure applicative. Les règles de la GCCA sont :

Règles de changement de type :

---

[2] Dans le phénotype, les expressions linguistiques sont concaténées selon les règles syntagmatiques propre à la langue. Dans le génotype, les expressions sont agencées selon l'ordre applicatif.

[3] Nous choisissons ici la notation de Steedman (1989) : X/Y et X\Y sont des types orientés fonctionnels. Une unité linguistique 'u' avec le type X/Y (respectivement X\Y) est considérée comme un opérateur (ou une fonction) dont l'opérande de type Y est positionné à droite (respectivement à gauche) de l'opérateur.

$$\frac{[X : u]}{[Y/(Y\backslash X) : (\mathbf{C}_{*}\ u)]}\ \mathbf{>T} \quad ; \quad \frac{[X : u]}{[Y\backslash(Y/X) : (\mathbf{C}_{*}\ u)]}\ \mathbf{<T}$$

$$\frac{[X : u]}{[Y/(Y/X) : (\mathbf{C}_{*}\ u)]}\ \mathbf{>Tx} \quad ; \quad \frac{[X : u]}{[Y\backslash(Y\backslash X) : (\mathbf{C}_{*}\ u)]}\ \mathbf{<Tx}$$

Règles de composition fonctionnelle :

$$\frac{[X/Y : u_1]\text{-}[Y/Z : u_2]}{[X/Z : (\mathbf{B}\ u_1\ u_2)]}\ \mathbf{>B} \quad ; \quad \frac{[Y\backslash Z : u_1]\text{-}[X\backslash Y : u_2]}{[X\backslash Z : (\mathbf{B}\ u_2\ u_1\ )]}\ \mathbf{<B}$$

$$\frac{[X/Y : u_1]\text{-}[Y\backslash Z : u_2]}{[X\backslash Z : (\mathbf{B}\ u_1\ u_2)]}\ \mathbf{>Bx} \quad ; \quad \frac{[Y/Z : u_1]\text{-}[X\backslash Y : u_2]}{[X/Z : (\mathbf{B}\ u_2\ u_1)]}\ \mathbf{<Bx}$$

Les prémisses des règles sont des expressions concaténées typées; les résultats sont des expressions applicatives (typées) avec éventuellement introduction d'un combinateur. Le changement de type d'une unité u introduit le combinateur $\mathbf{C}_{*}$; la composition de deux unités concaténées introduit le combinateur **B**.

Pour l'exemple suivant *La liberté renforce la démocratie* nous avons l'analyse suivante :

1.	[N/N : *la*]-[N : *liberté*]-[(S\N)/N : *renforce*]-[N/N : *la*]-[N : *démocratie*]	
2.	[N : (*la liberté*)]-[(S\N)/N : *renforce*]-[N/N : *la*]-[N : *démocratie*]	**(>)**
3.	[S/(S\N) : (**C*** (*la liberté*))]-[(S\N)/N : *renforce*]-[N/N : *la*]-[N : *démocratie*]	**(>T)**
4.	[S/N : (**B** (**C*** (*la liberté*)) *renforce*)]-[N/N : *la*]-[N : *démocratie*]	**(>B)**
5.	[S/N : (**B** (**B** (**C*** (*la liberté*)) *renforce*) *la*)]-[N : *démocratie*]	**(>B)**
6.	[S : ((**B** (**B** (**C*** (*la liberté*)) *renforce*) *la*) *démocratie*)]	**(>)**
7	[S : ((**B** (**B** (**C*** (*la liberté*)) *renforce*) *la*) *démocratie*)]	
8.	[S : ((**B** (**C*** (*la liberté*)) *renforce*) (*la démocratie*))]	**B**
9.	[S : ((**C*** (*la liberté*)) (*renforce* (*la démocratie*)))]	**B**
10.	[S : ((*renforce* (*la démocratie*)) (*la liberté*)))]	**C ***
11.	[S : *renforce* (*la démocratie*) (*la liberté*)]	

Ainsi pour cet exemple, à l'étape 1 des types catégoriels sont assignés aux unités linguistiques. À l'étape 2, la règle (>) est appliqué aux unités linguistiques *la* et *liberté*. À l'étape 3 une règle de changement de type (**>T**) est déclenchée pour construire un opérateur (**C*** (*la liberté*)) à partir de l'opérande (*la liberté*). Cet opérateur est composé avec l'opérateur *renforce* à l'étape 4 par une opération de composition (**>B**) de façon à former un opérateur complexe (**B** (**C*** (*la liberté*)) *renforce*). Deux autres opérations respectivement (**>B**) et (**>**) suivent.

À l'étape 7 commence la réduction des combinateurs. Cette série de réduction se fait dans le génotype.

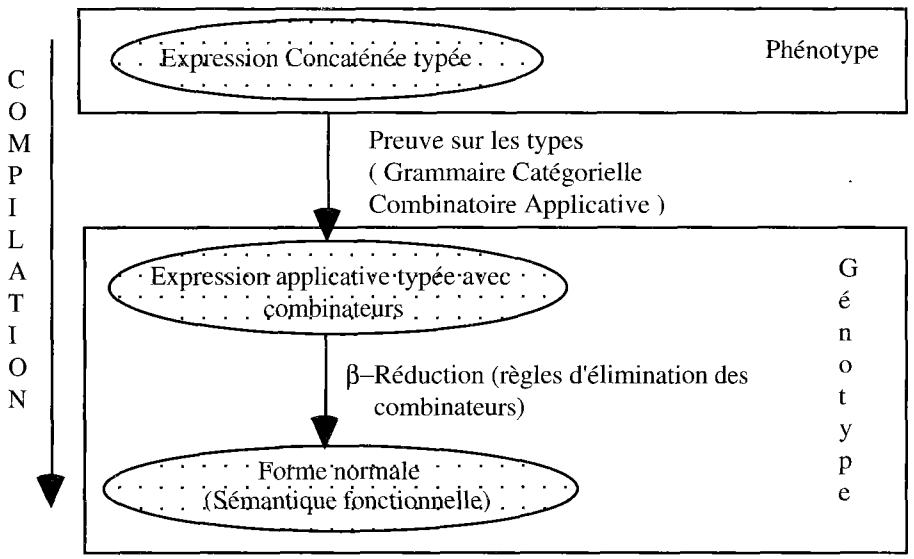
À l'étape 11 nous produisons la structure prédicative dont nous aurons besoin pour aider le terminologue.

Un traitement complet basé sur la Grammaire Catégorielle Combinatoire Applicative s'effectue en deux grandes étapes :

(i) la première étape s'illustre par la vérification de la bonne connexion syntaxique et la construction de structures prédicatives avec des combinateurs introduits à certaines positions de la chaîne syntagmatique;

(ii) la deuxième étape consiste à utiliser les règles de ß-réduction des combinateurs de façon à former une structure prédicative sous-jacente à l'expression phénotypique. L'expression obtenue est applicative et appartient au langage génotype. La GCCA engendre des processus qui associent une structure applicative à une expression concaténée du phénotype. Il nous reste à éliminer les combinateurs de l'expression obtenue de façon à construire la «forme normale» (au sens technique de la ß-réduction) qui exprime l'interprétation sémantique fonctionnelle. Ce calcul s'effectue entièrement dans le génotype. Les formes applicatives obtenus en bout de parcours seront retenues pour être stockées dans des bases de données à des fins d'aide terminologique à l'utilisateur.

Le traitement fondé sur la GCCA prend la forme d'une compilation dont les étapes sont résumées dans la figure 1 :



4. LE MODÈLE HYBRIDE

Dans sa forme concrète le modèle hybride que nous proposons consiste en deux grandes étapes :

- Un filtrage numérique grossier du corpus qui permet de classifier et de structurer le corpus en des classes de termes qui serviront d'indices de régularités d'associations lexicales que le terminologue peut utiliser comme tremplin pour approfondir les étapes ultérieures d'interprétation, de construction de réseaux sémantiques, et finalement d'élaboration de ses fiches terminologiques. Une plate-forme réalisée au LANCI, en l'occurrence, la plate-forme ALADIN (Seffah et Meunier, 1995) permet d'exécuter une chaîne de traitement qui réalise un tel filtrage. La chaîne présente les étapes suivantes : elle commence par une gestion du document, suit alors une description morphologique (lemmatisation) et une transformation matricielle du corpus. Vient ensuite une extraction classificatoire par réseaux de neurones FUZZYART. Ainsi dans la première étape de sa gestion, le texte est reçu et traité par des modules d'analyse de la plate-forme ALADIN-TEXTE. Cette plate-forme est un atelier qui utilise des modules spécialisés dans l'analyse d'un texte. Dans un premier temps, un filtrage sur le lexique du texte est fait. Par divers critères de discrimination, on élimine du texte certains mots accessoires (mots fonctionnels ou statistiquement insignifiants, etc.) ou ceux qui ne sont pas porteurs de sens d'un point de vue strictement sémantique, et dont la présence pourrait nuire au processus de catégorisation, soit parce qu'ils alourdiraient indûment la représentation matricielle, soit parce que leur présence nuirait au processus interprétatif qui suit la tâche de catégorisation. Vient ensuite une description morphologique minimale de type lemmatisation.

Puis une transformation est opérée pour obtenir une représentation matricielle du texte. Cette transformation est encore effectuée par des modules d'ALADIN explicitement dédiés à cette fin. On produit ainsi un fichier indiquant pour tout lemme choisi sa fréquence dans chaque segment du texte. Suit ensuite un post-traitement pour construire une matrice dans un format acceptable par le réseau de neurone FUZZYART[4].

Le réseau neuronal génère une matrice de résultats qui représentent la classification trouvée. Chaque ligne (ou vecteur) de cette matrice est constituée d'éléments binaires ordonnés. La ligne indique pour chaque terme du lexique original s'il fait ou non partie du prototype de la classe. Ainsi est créé un «prototype» pour chacune des classes identifiées. On dira alors que la classe numéro X est «caractérisée» par la présence d'un certain nombre de termes. Autrement dit, chaque classe identifie quels sont les termes qui se retrouvent dans les segments de textes qui présentent, selon le réseau de neurones une certaine similarité. Ainsi, les classes créées sont caractérisées, arbitrairement, par les termes qui sont présents également dans tous les segments du texte qui ont été «classifiés» dans une même classe.

Les résultats du réseau de neurones se présentent donc (avant interprétation) sous la forme d'une séquence de classes que l'on dira «caractérisées» par des termes donnés et incluant un certain nombre de segments.

---

[4] Le réseau de neurones FUZZYART utilisé pour l'expérimentation d'ALADIN a été développé sur une plate-forme de programmation matricielle disponible sur le grand marché appelé MATLAB.

- Un traitement linguistique plus fin des segments sélectionnés selon les thèmes choisis par le terminologue. Ce dernier sélectionne donc des segments dont il veut une analyse plus fine et en extraire une représentation des connaissances plus structurée. Le terminologue peut décider pour un segment donné de focaliser son attention sur un terme donné et en construire son réseau sémantique. La Grammaire Catégorielle Combinatoire Applicative peut organiser les phrases dans lesquelles apparaît le terme, choisi par le terminologue, sous forme de structures prédicatives `Prédicat argument1 argument2... argumentn`. Ainsi pour une sélection de phrases on peut engendrer une liste d'expressions prédicatives. Nous pouvons avoir dans cette liste des structures prédicatives ayant des arguments en commun.

Par exemple, le cas suivant :

Prédicat1 argument1 argument2
Prédicat2 argument3 argument1

et là un terminologue comprendrait la relation sémantique entre les arguments 2 et 3 par rapport à l'argument 1.

Nous pouvons conserver une liste d'expression de la forme 'Prédicat argument1 argument2... argumentn' dans une base de données.

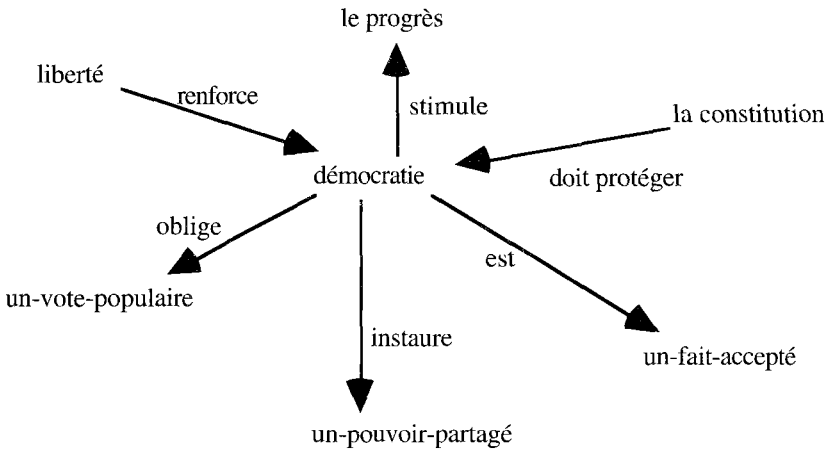
Le terminologue la consultant peut déduire le sens sémantique de chaque argument (donc terme) en ce sens qu'il peut en construire le réseau sémantique. Prenons les phrases suivantes :

La liberté renforce la démocratie.
La démocratie stimule le progrès.
La démocratie oblige un vote populaire.
La démocratie est un fait accepté.
La constitution doit protéger la démocratie.
La démocratie instaure un pouvoir partagé.

Les structures prédicatives de ces phrases obtenues par une analyse linguistique catégorielle sont respectivement :

renforce (la démocratie) (la liberté)
stimule (le progrès) (la démocratie)
oblige (un-vote-populaire) (la démocratie)
est (un-fait-accepté) (la démocratie)
(doit protéger) (la démocratie) (la constitution)
instaure (un-pouvoir-partagé) (la démocratie)

un terminologue ayant dans sa base de données ces structures prédicatives pourra représenter le réseau sémantique de (la démocratie).



Remarque

De tels exemples cependant ne mettent pas en évidence la pertinence de l'approche par classifieurs numériques car il n'y a pas d'ambiguïté dans les exemples. Le mot pôle *démocratie* a un sens unique.

Prenons le mot pôle *ferme*

Nous pouvons avoir trois groupes possibles :

Groupe 1
Une démocratie ferme est l'espoir d'un peuple.
Un vote ferme est passé à l'assemblée.
C'est une position ferme du gouvernement.

Groupe 2
Les paysans ont occupé les fermes et les villages.
Toutes les fermes ont abandonné leur récolte.
Les fermes sont laissées à l'abandon.

Groupe 3
Le vote ferme la discussion.
La décision du gouvernement ferme les options.
Le président de l'assemblée ferme le vote.

C'est le traitement classifieur qui a permis de séparer les champs lexicaux en trois groupes. Mais c'est au moyen de la grammaire catégorielle combinatoire applicative qu'une analyse en profondeur sera faite.

5. CONCLUSION

Nous venons de présenter un modèle d'extraction des connaissances pour terminologues. L'idée d'associer des modèles linguistiques à des modèles numériques est très prometteuse. Elle est également très pertinente, en ce sens qu'elle associe la finesse d'analyse des méthodes linguistiques à la capacité des méthodes numériques d'absorber de gros corpus. L'ordre stratégique d'appliquer une méthode numérique avant de faire intervenir une méthode linguistique résulte du compromis nécessaire pour faire `cohabiter` ces deux approches. En effet, la méthode numérique est plus à même de `débroussailler` un gros texte et de permettre à un terminologue de soumettre des segments choisis à l'analyseur linguistique plus fin.

Ce type d'approche permet de solutionner une des critiques importantes qu'on fait classiquement aux approches de cooccurrence et collocation : leurs difficultés d'interprétation. Notre approche permet d'entrevoir des outils de raffinement de l'analyse des résultats livrés par les approches numériques trop grossières et générales. Il y a compensation entre les deux. Les Grammaires Catégorielles sont trop fines et donc trop lentes sur un corpus ample. Mais bien placées elles ne travaillent que sur des sous-corpus qui ont effectué un premier travail de désambiguïsation.

En fait cette approche oblige à effectuer la désambiguïsation dans un traitement différent de celui de la grammaire. La désambiguïsation joue sur la différentialité des contextes de l'ensemble d'un corpus (relation paradigmatique) alors que l'analyse catégorielle opère sur la dépendance des contextes immédiats (relation syntagmatique). Ainsi, le système n'est pas obligé de faire les deux analyses en même temps ou dans une même passe ce qui, pensons nous, le rend plus efficace.

Enfin la configuration des résultats telle que permise par les expressions prédicatives permet au lecteur analyste d'avoir une organisation plus limpide sur le plan ergonomique et cognitif. On envisage la possibilité de livrer ces résultats dans des réseaux structurés mais aussi dans des reposoirs structurés genre dictionnaires, thesaurus, bases de connaissances, etc.

Enfin, signalons que nous travaillons également sur un autre modèle hybride numérique/linguistique, associant réseaux de neurones et le modèle d'exploration contextuelle (système SEEK) (Jouis, 1993; Jouis, Biskri, Desclés, Meunier et al., 1997). Ce système permet de détecter des liens sémantiques de types cinématiques ou dynamiques (mouvements d'objets, changements d'états d'objets, relations de causalité, recherche des contextes définitoires d'un terme, etc.). Cela permettra d'élargir les interprétations par le terminologue des niveaux descriptifs statiques du domaine vers des descriptions évolutives.

RÉFÉRENCES

BISKRI, I. (1995) : *La Grammaire Catégorielle Combinatoire Applicative dans le cadre de la Grammaire Applicative et Cognitive*, Thèse de doctorat, EHESS, Paris.

BISKRI, I. et J.P. DESCLÉS, (1995) : «Applicative and Combinatory Catégorial Grammar (from syntax to functional semantics)», *Colloque RANLP*, Bulgarie 1995.

BOUCHAFFRA, D. & G. MEUNIER (1993) : «Theory and Algorithms for Analysing the Consistent Region in Probabilistic Logic», *An International Journal of Computers & Mathematics with Applications*, vol. 25, n° 3, February, edit. Ervin Y. Rodin, Published by Pergamon Press.

BURR, D. J. (1987) : «Experiments with a Connectionnist Text Reader», *IEEE First International Conference on Neural Networks*, San Diego, pp. 717-724.

CARPENTER, G. & G. GROSSBERG (1991) : «An Adaptive Resonnance Algorithm for Rapid Category Learning and Recognition», *Neural Networks*, 4, pp. 493-504.

CHEESEMAN, P., SELF, M., KELLY, J., STUTZ, J., TAYLOR, W. & D. FREEMAN (1988) : «Bayesian Classification», *Proceedings of AAAI 88*, Minneapolis, pp. 607-611.

CHURCH, K., GALE, W., HANKS, P. & D. HINDLE (1989) : «Word Associations and Typical Predicate-argument Relations», *International Workshop on Parsing Technologies*, Carnegie Mellon University, Aug. 28-31.

CHURCH, K. W. & P. HANKS (1990) : «Word Asociation Norms, Mutual Information, and Lexicography», *Computational Linguistics*, 16, pp. 22-29.

CURRY, B. H. & R. FEYS (1958) : *Combinatory logic*, Vol. I, North-Holland.

DELANY P. & P. LANDOW (Eds) (1993) : *The Digital Word : Text Based Computing in the Humanities*, Cambridge, Mass, MIT Press.

DELISLE, S. (1994) : *Text Processing Without A Priori Domain Knowledge : Semi Automatic Linguistic Analysis for Incremantal Knowledge Acquisition*, PhD Thesis, Ottawa University.

DESCLÉS, J. P. (1990) : *Langages applicatifs, langues naturelles et cognition*, Paris, Hermès.

DESCLÉS, J. P. & I. BISKRI (1996) : «Logique combinatoire et linguistique : Grammaire Catégorielle Combinatoire Applicative», *Revue Mathématiques, Informatiques et Sciences Humaine,* Paris.

FREY, S., REYLE, U. & C. ROHRER (1983) : «Automatic Construction of a Knowledge Base by Analysing Texts in Natural Language», *Proc. of IJCAI 83*, pp. 727-729.

GARNHAM, A. (1981) : «Mental Models and Representation of Texts», *Memory and Cognition*, 9, pp. 560-565.

GREFENSTETTE, G. (1992) : «Sextant : Exploring Unexplored Contexts for Semantic Extraction from Syntactic Analysis», *Proc. of the 30th Annual Meeting of the ACL*, pp. 324-326.

GREFENSTETTE, G. (1992) : «Use of Syntactic Context to Produce Term Association Lists for Text Retrieval», *Proc. of SIGIR 92 ACM*, Copenhagen, June 21-24.

GROSSBERG, S. & S. CARPENTER (1987) : «Self Organization of Stable Category Recognition Codes for Analog Input Patterns», *Applied Optics*, 26, pp. 4919-4930.

JACOBS, P. & U. ZERNIK (1988) : «Acquiring Lexical Knowledge from Text A Case Study», *Proceedings of AAAI 88*, St Paul, Min.

JANSEN, S., OLESEN, J., PREBENSEN, H. & T. THARNE (1992) : *Computanional approaches to text Understanding,* Copenhaguen, Museum Tuscalanum Press.

JOUIS, C. (1993) : *Contributions à la conceptualisation et à la modélisation des connaissances à partir d'une analyse linguistique de textes. Réalisation d'un prototype : le système SEEK*, Thèse de doctorat, EHESS, Paris.

JOUIS, C., BISKRI, I., DESCLÉS, J. P., LE PRIOL, F., MEUNIER, J.-G., MUSTAFA, W. & G. NAULT (1997) : «Vers l'intégration d'une approche sémantique linguistique et d'une approche numérique pour un outil d'aide à la construction de bases terminologiques», *Actes du colloque JST97*, Avignon, France.

KOHONEN, T. (1982) : «Clustering, taxonomy and topological Maps of Patterns», *IEEE Sixth International Conference on Pattern Recognition*, pp. 114-122.

LEBART, L. & A. SALEM (1988) : *Analyse statistique des données textuelles*, Paris, Dunod.

LIN, X., SOERGEL, D. & G. MARCHIONINI (1991) : «A Self Organizing Semantic Map for Information Retrieval», *SIGIR 91*, Chicago, Illinois.

MEUNIER, J.-G. (1996) : «Théorie cognitive : son impact sur le traitement de l'information textuelle», V. Riale et D. Fisette, *Penser l'esprit des sciences de la cognition a une philosophie cognitive*, Presses de Université de Grenoble, pp. 289-305.

MOULIN, B. & D. ROUSSEAU (1990) : «Un outil pour l'acquisition des connaissances à partir de textes prescriptifs», *ICO*, Québec 3 (2), pp. 108-120.

REGOCZEI, S. et al. (1988) : «Creating the Domain of Discourse : Ontology and Inventory». Gaines &. Boose (Eds), *Knowledge Acquisition Tools for Experts and Novices*, Academic Press.

REGOCZEI, S. & G. HIRST (1989) : On extracting knowledge from Text. Modeling the Architecture of Language Users. (TR CSRI 225), Computer Systems Research Institute, University of Toronto.

SALTON, G. (1988) : «On the Use of Spreading Activation», *Communications of the ACM*, vol 31 (2).

SALTON, G., ALLAN, J. & C. BUCKLEY (1994) : «Automatic Stucturing and Retrieval of Large Text File», *Communications of the ACM*, 37 (2), pp. 97-107.

SEFFAH, A. & J.-G. MEUNIER (1995) : «ALADIN : un atelier orienté objet pour l'analyse et la lecture de textes assistée par ordinateur», *International Conference on Statistics and Texts,* Rome.

SHAUMYAN, S. K. (1987) : *A Semiotic Theory of Natural Language*, Bloomington, Indiana Univ. Press.

SHAW, M. L. G. & B. R. GAINE (1988) : «Knowledge Initiation and Transfer Tools for Expert and Novices», Boose &. Gaines (Eds), *Knowledge Acquisition Tools for Expert Systems*, Academic Press.

STEEDMAN, M. (1989) : *Work in progress : Combinators and grammars in natural language understanding*, Summer Institute of Linguistic, Tucson University.

TAPIERO, I. (1993) : *Traitement cognitif du texte narratif et expositif et connexionnisme : expérimentations et simulations*, Université de Paris VIII.

THRANE, T. (1992) : «Dynamic Text Comprehension», J. O. S. Jansen, H. Prebensen, T. Thrane (Eds), Copenhaguen, Museum Tuscalanum Press.

VERONNIS, J., IDE, N. M. & S. HARIE (1990) : «Utilisation de grands réseaux de neurones comme modèles de représentations sémantiques», *Neuronimes*.

VIRBEL, J. (1993) : «Reading and Managing Texts on the Bibliothèque de France Stations», P. Delany & P. Landow (Eds), *The Digital Word : Text Based Computing in the Humanities. Cambridge*, Mass., MIT Press.

YOUNG, T. & T. CALVERT (1987) : *Classification, Estimation, and Pattern Recognition*, Amsterdam, Elsivier.

ZARRI, G. P. (1990). «Représentation des connaissances pour effectuer des traitements inférentiels complexes sur des documents en langage naturel», Office de la langue française (Ed.), *Les industries de la langue. Perspectives 1990*, Gouvernement du Québec.

DÉVELOPPEMENT DE LEXIQUES À GRANDE ÉCHELLE

Pierrette BOUILLON, Sabine LEHMANN, Sandra MANZI et
Dominique PETITPIERRE

ISSCO, Université de Genève, Genève, Suisse

1. INTRODUCTION

Cet article donne un aperçu général de trois descriptions linguistiques développées en parallèle pour l'anglais, le français, l'allemand et l'italien, à l'aide de l'outil morphologique *mmorph*. Il insiste surtout sur sa convivialité et ses diverses possibilités.[1]

mmorph a été conçu dans le cadre du projet européen *Multext* («Multilingual Text Tools and Corpora», Projet LRE 62-050, 1994-1996). Ce projet s'est donné pour objectif de développer un ensemble d'outils pour le traitement de textes multilingues, destinés à des études de corpus ou à des applications dans le domaine du TALN. À son terme, il offre ainsi aux chercheurs des *corpus* et des *outils* pour les *annoter*. Ces derniers comprennent un *segmenteur* (qui segmente le texte en ses différentes unités : paragraphes, phrases et mots), un *analyseur morphologique* (*mmorph*) et diverses *descriptions linguistiques* (qui permettent de faire l'analyse morphologique du mot), un *étiquetteur probabilistique* (ou taggeur) (qui prend en entrée des mots avec des étiquettes ambiguës et lui attribue la plus probable en fonction de son contexte) et, enfin, un *aligneur* (qui établit des correspondances entre les diverses unités, paragraphes, phrases et mots, de deux textes de deux langues différentes). L'outil d'analyse morphologique que nous décrivons ici ne constitue donc qu'un des maillons dans la chaîne de traitement des corpus.

Dans la suite, nous décrivons d'abord l'outil *mmorph*, puis illustrons ses possibilités à l'aide d'exemples concrets tirés des descriptions linguistiques des différentes langues. Nous terminons en donnant une vue d'ensemble des lexiques.

2. L'OUTIL *mmorph*

mmorph est un outil qui permet de décrire la morphologie des mots d'une langue. Cette description est interprétée pour produire une base de données de mots fléchis. Celle-ci peut ensuite être manipulée de différentes façons, par exemple pour associer un mot fléchi à des informations morphologiques (analyse) ou pour dériver des formes fléchies à partir de la forme de base d'un mot et d'indications morphologiques (Petitpierre et Russell, 1995).

---

1 Nous remercions Susan Armstrong et Graham Russell qui ont participé à ce travail et Suissetra qui a contribué au financement du projet.

mmorph est basé sur le paradigme de la morphologie à deux niveaux (Koskenniemi, 1984; Karttunen et Wittenburg, 1983 et Ruessink, 1989), avec l'addition de traits et de règles qui décrivent la syntaxe interne des mots (Pulman et Hepple, 1993; Ritchie et al., 1992).

Pour *mmorph*, la **description morphologique** d'une langue comporte quatre parties , a) un ensemble de déclarations, b) un ensemble d'entrées lexicales, c) un ensemble de règles de production lexicale, et d) un ensemble de règles morphographémiques (orthographiques).

a. Les **déclarations** définissent et nomment certains concepts utilisés dans la description morphologique (types, traits, valeurs, alphabets, etc.), ce qui permet au programme de vérifier la cohérence de la description morphologique, et à l'utilisateur de travailler avec une notation adaptée au cadre linguistique dans lequel il opère.

b. Les **entrées lexicales** (**lexique**) associent une description morphosyntaxique à des mots complets ou partiels (racines ou affixes). La description morphosyntaxique est composée d'un type et d'un ensemble de traits (des paires *attribut=valeur).* Les mots sont spécifiés par une chaîne de symboles lexicaux.

Le mot *étudiant*, par exemple, reçoit la description donnée en (1).

```
(1) Nom [genre=masculin nombre=singulier forme=surface type=2]
"étudiant" = "étudiant"
```

Il a le type `Nom`. Les traits `genre=masculin` et `nombre=singulier` indiquent qu'il est au masculin singulier; le trait `type=2` qu'il accepte une flexion du féminin (pour former *étudiante*) et le trait `forme=surface` qu'il constitue un mot complet sans adjonction d'affixe (contrairement à la racine «appel» du verbe *appeler*).

Le lexique décrit donc les mots complets ou partiels. Les règles de production définissent comment ils se combinent.

c. Les **règles de production lexicale** sont des règles de réécriture qui définissent ce qu'est un mot bien formé et comment les traits des morphèmes se combinent dans le mot. La règle (2), par exemple, crée un nom *pluriel* en concaténant un nom *singulier* (de type `Nom`) avec un suffixe nominal pluriel (de type `Nom_suffixe`), par exemple «mère + s» → *mères*. Elle précise que le nom pluriel aura le même *genre* que son homologue au *singulier* (le `$G` indique une variable) et qu'il aura le trait `pluriel`.

```
(2)        Nom[nombre=pluriel genre=$G]
           ¨ Nom[nombre=singulier genre=$G]
             Nom_suffixe [nombre=pluriel]
```

Les règles de production combinent donc les chaînes entre elles. Cette combinaison provoque cependant différents changements orthographiques (suppression de lettres : «nez+s» → *nez;* ajout : «gros+e» → *grosse;* changement : «beau+s» → *beaux).* Ceux-ci sont pris en compte par les règles morphographémiques.

d. L'ensemble de **règles morphographémiques** établit une correspondance entre les chaînes de symboles lexicaux (qui proviennent de la concaténation de mots complets et partiels, par exemple «beau+s» en (3)) et celles de surface (le mot après modification orthographique, comme *beaux*).

forme de surface	b	e	a	u		x
forme lexicale	b	e	a	u	+	s

(3) Tableau

Les règles morphographémiques se présentent comme en (4).

```
(4) Opérateur Contexte gauche - Focus - Contexte droit Contraintes
```

Ici, `Opérateur` indique s'il s'agit d'une règle optionnelle (=>), de coercion (<=) ou obligatoire (<=>) ; `Focus`, la correspondance entre le niveau lexical et le niveau de surface; `Contexte gauche` et `Contexte droit`, les contextes dans lesquels la correspondance est possible (règle optionnelle), autorisée (règle de coercion) ou obligatoire. Enfin, `Contraintes` est une liste d'ensembles typés de traits qui restreint l'application de la règle à certaines constructions. `Focus`, `Contexte droit` et `Contexte gauche` spécifient des paires de symboles lexicaux et de surface, comme en (5).

```
(5) symbole de surface / symbole lexical
```

Prenons comme exemple le cas de la règle qui supprime le morphème «s» du pluriel après un «z» final («nez+s» → *nez*) :

```
(6): <=> z/z +    - <>/s - ~ Plur_suffixe[]
```

Celle-ci spécifie que le «s» ne correspond à aucun symbole de surface (`<>/s`) si le contexte gauche est un «z» au niveau lexical et de surface (`z/z`). Cette règle est contrainte par les informations du champs `Contraintes` : elle ne s'appliquera qu'en présence d'affixes de type `Plur_suffixe` (c'est-à-dire les suffixes du pluriel). Le symbole (+) indique le point d'attache des affixes et le tilde (~), la fin du mot.

mmorph facilite la description des règles morphographémiques par l'usage de macros. Par exemple, la règle de (6) s'applique aussi après le «s» et le «x» («prix+s» → *prix*; «repas+s» → *repas*). Plutôt que de définir plusieurs règles, le linguiste peut remplacer le *Contexte gauche* par la macro `s_ou_x_ou_z`, comme en (7a). Celle-ci correspond à la disjonction des trois paires `s/s`, `x/x` et `z/z`, comme l'illustre (7b). Ainsi, la règle s'appliquera après un «s», un «x», ou un «z».

```
(7)    a. <=> s_ou_x_ou_z + <>/s - ~ Plur_suffixe[]
       b. s_ou_x_ou_z : s/s x/x z/z
```

mmorph est donc un système basé sur règles, qui s'oppose aux listes de mots complets : le fait d'utiliser des règles permet de tirer des généralisations linguistiques. Celles-ci facilitent ainsi considérablement la modification et l'extension de lexiques (Bouillon et Tovena, 1991).

La suite montre cette flexibilité et le pouvoir de généralisation. Elle décrit d'abord les règles puis les lexiques.

3. LES RÈGLES

3.1 Français

Notre morphologie du français comporte quarante-sept règles de production et sept règles morphographémiques. Le français présente trois particularités, par rapport à une langue comme l'anglais : tout d'abord, elle nécessite souvent le changement de plusieurs lettres à la fois («chameau+e» → *chamelle*); ensuite, l'application des règles ne peut être prédite de la forme lexicale («appel+e» → *appelle*, mais «pel+e» → *pèle*). Enfin, certaines règles sont facultatives. Nous les examinons successivement.

a. **Mise en correspondance d'une séquence de symboles** : dans beaucoup d'analyseurs morphologiques de ce type (CLE, par exemple, voir Rayner *et al.*, 1996 et Carter, 1995), les règles morphographémiques ne peuvent faire correspondre qu'une paire de symboles à la fois. Pour traiter des cas comme «chameau+e» → *chamelle*, le linguiste doit donc écrire deux règles : l'une qui établit la correspondance entre le «a» et le «l» devant la paire «l/u» et l'autre qui fait correspondre «u» à «l» devant le suffixe du féminin «e» muet. Il s'ensuit une multiplication des règles. Dans *mmorph,* celle-ci peut être évitée (contrairement à ce que soutient Ruessink, 1989), puisqu'une même règle peut porter sur des séquences de symboles. La règle en (8), par exemple, spécifie que les symboles lexicaux «a» et «u» de *chameau* correspondent tous les deux au symbole de surface «l» après un «e» et devant un «e» muet.

```
(8) <=> e/e - l/a l/u - + e/e_muet
```

La description des verbes du troisième groupe de conjugaison est ainsi grandement facilitée («peign+ra» → *peindra*; «dev+e» → *doive*», etc.).

b. **Contrainte sur l'application des règles** : dans *mmorph*, différentes possibilités peuvent être envisagées pour contraindre l'application de règles : tout d'abord, on peut choisir de ne décrire que les règles générales et d'énumérer les exceptions (comme dans Bouillon et Tovena; 1990, Ostling Anderson, 1987). En français, cependant, une chaîne lexicale peut avoir plusieurs équivalents au niveau de surface, sans que l'on puisse décider quel est le cas général et quelles sont les exceptions. Par exemple, les verbes, les noms et les adjectifs qui se terminent par «el» et «et» peuvent doubler le «l» ou le «t» devant un «e» muet («appel+e» → *appelle*) ou changer le «*e*» préfinal en «*è*» («pel+e» → *pèle*). Dans ce cas, quelle est l'exception ? Deux autres solutions peuvent aussi être envisagées en *mmorph* : on peut introduire un

caractère lexical spécial dans les entrées lexicales pour distinguer les différentes chaînes entre elles ou ajouter des contraintes dans les règles morphographémiques. Illustrons ces deux possibilités avec le traitement des mots se terminant en «el». Pour distinguer les mots qui doublent la consonne de ceux qui changent le «e» préfinal en «è», nous remplaçons d'abord le «e» de la racine par le symbole lexical «e_fort» dans le premier cas et nous gardons le «e» dans le second, comme en (9) (dans l'entrée, un symbole lexical composé de plusieurs caractères commence par «&» et se termine par «;» pour éviter la confusion avec les symboles lexicaux qui ont une seule lettre).

```
(9)     a. "app&e_fort;l" = "appeler"
        b. "pel" = "peler"
```

Dans la règle morphographémique qui double le «l» (10), nous précisons ensuite que le changement ne se produira qu'en présence de ce symbole dans le contexte gauche. Elle ne pourra donc pas s'appliquer à *peler*.

```
(10): <=> e/e_fort - <l l>/l - + Fut_e_muet
```

Le problème est en fait plus complexe pour deux raisons différentes :

(i) d'une part, le «e» (qui correspond au premier caractère du suffixe du futur et du conditionnel) peut parfois être considéré comme un «e» muet qui affecte les lettres qui le précèdent, sans qu'il s'agisse d'une règle générale. Par exemple, «appel+erai» → *appellerai* (avec doublement du «l» au futur et au conditionnel), mais «céd+erai» → *céderai*, et non pas **cèderai*, comme on s'y attendrait au vu de «céd+e» → *cède*. Pour rendre compte de cette différence, nous utilisons différents symboles lexicaux qui distinguent parmi les affixes de type `Verbe_suffixe` ceux qui ne sont pas muets («ez») (11a), ceux qui le sont («e») (11b) et ceux qui peuvent être considérés comme tels («erai», «erais») (11c).

```
(11)    a. "ez" Verbe_suffixe[...]
        b. "&e_muet;" Verbe_suffixe[...]
        c. "&fut_e;rai" Verbe_suffixe[...]
```

Dans les règles morphographémiques, nous spécifions ensuite si la règle s'applique ou pas devant le `"fut_e"` du futur et du conditionnel. En (10) par exemple, la règle qui double le «l» s'applique devant le «e» muet, y compris le «e» du futur et du conditionnel, puisque `Fut_e_muet` est une macro qui définit les deux paires de (12).

```
(12) Fut_e_muet :       e/e_muet
                        e/fut_e
```

En revanche, la règle qui change le «é» en «è» ne s'applique que devant le «e» muet, comme exemplifié en (13).

```
(13) <=> - é/è - Consonne + e/e_muet
```

(ii) D'autre part, les noms et les adjectifs peuvent aussi doubler le «l» et le «t» après d'autres voyelles, comme le «i» («gentil+e» → *gentille*) et le «u» («nul+e» → *nulle*). Pour éviter que cette règle ne s'applique aux verbes («pil+e» → **pille*), nous limitons son application au morphème du féminin et nous ajoutons une contrainte dans la règle qui effectue le doublement : (14) ne s'appliquera que devant le morphème «e_muet» qui a le type `Fem_suffixe`.

```
(14)      <=> u_ou_i - <ll>/l - + e/e_muet Fem_suffixe []
```

c. **Les règles optionnelles** : certaines règles en français sont optionnelles, comme celle qui change le «y» en «i» si cette lettre suit un «a» et précède un «e» muet («pay+e» → *paie* ou *paye*) . Pour ce faire, nous utilisons l'opérateur => qui indique que le changement morphographémique est optionnel (15). Notons que ce changement peut aussi avoir lieu devant le «e» du futur et du conditionnel (`Fut_e_muet`).

```
(15) => a/a - i/y - + Fut_e_muet
```

3.2 Italien

Notre traitement de l'italien se fait par dix règles de production et seize règles morphographémiques. Ici, nous nous intéresserons à la règle d'insertion du «h» et au traitement des clitiques.

a. **Insertion du «h»** : la plupart des mots dont la racine se termine par «g» et «c» insèrent la semi-consonne «h» devant tout suffixe commençant par «e» ou «i» («elvetic+e» → *elvetiche, helvétiques;* «acciug+e» → acciughe, *anchois;* «trasloc+i» → «traslochi», *tu déménages*). Pour opérer ce changement, nous définissons la règle en (16).

```
(16) <=> c_ou_g - h/<> ¬ + e_ou_i
```

Les mots qui n'obéissent pas à cette règle reçoivent un caractère lexical spécial à la place du «c» et du «g», comme en (17) pour l'équivalent italien de *faux*.

```
(17)      "falc&no_h;" = "falce"
```

b. **Clitiques** : une des particularités de l'italien est la présence de clitiques qui s'attachent à certains modes et temps (impératif, gérondif, participe présent et infinitif) au verbe qui le sous-catégorise, de la manière résumée dans le tableau (18).

Impératif singulier avec les différentes réalisations de clitiques				
	Sans clitique	Clitique accusatif (clA)	Clitique datif (clD)	Clitiques accusatif et datif (clA et clD)
intransitif	parti ! *pars !*	-	-	-
transitif direct	vinci ! *vaincs !*	vincilo ! *vaincs-le !*	-	-
transitif indirect	obbedisci ! *obéis !*	-	obbediscigli ! *obéis-lui !*	-
bi-transitif	dedica ! *dédie !*	dedicalo ! *dédie-le !*	dedicami ! *dédie-moi !*	dedicamelo ! *dédie-le-moi !*

(18) Tableau

Pour traiter les clitiques et générer toutes ses formes, nous avons besoin des informations suivantes :

(i) Dans l'entrée lexicale, nous précisons si le verbe prend un clitique accusatif (clA) ou datif (clD). «*Dedicare*» (bi-transitif), par exemple, peut avoir un clitique accusatif et datif (la valeur de clA et clD est vide), alors que «*vincere*» (transitif) n'a qu'un clitique accusatif (la valeur de clD est aucun). Enfin, «*partire*» qui est intransitif n'accepte pas de clitique.

```
(19) a. Racine_verbe [clD=vide clA=vide ...] "dedic" = "dedicare"
     b. Racine_verbe [clD=aucun clA=vide ...] "vinc" = "vincere"
     c. Racine_verbe [clD=aucun clA=aucun ...] "part" = "partire"
```

(ii) Les règles de production combinent le verbe avec les clitiques qu'il admet. Ainsi, deux règles combinent un verbe avec le clitique datif ou accusatif si ceux-ci sont possibles. Une troisième prend pour entrée un verbe (avec le clitique datif) et y ajoute le clitique accusatif, en vérifiant certaines contraintes auxquelles sont soumis les clitiques italiens. Par exemple, la séquence «lele» est impossible («**decicalele*», dédie-le-lui).

(iii) Lors de la composition du verbe et du clitique se produisent différents changements orthographiques qui sont pris en compte par les règles orthographiques : celles-ci changent notamment le «i» final du clitique datif en «e», s'il est suivi d'un des clitiques accusatifs, soit «dedica+mi+lo» → *dedicamelo*.

3.3 Allemand

Notre morphologie de l'allemand comporte seize règles de production et cinquante-cinq règles morphographémiques. Nous discutons ici l'«umlautung» (insertion du tréma) et des cas de dérivation et montrons comment ils ont été traités dans *mmorph*.

a. **Insertion du tréma** : l'insertion du tréma est peut-être le phénomène morphographémique le plus connu de l'allemand. Il concerne le changement systématique de la voyelle de la *racine «a», «au», «o», «u»* (parfois *«e»*) en *«ä», «äu», «ö», «ü»* (et *«ie»*). Celui-ci se produit dans la morphologie flexionnelle (*«das Haus», la maison → «die Häuser», les maisons*) et dérivationnelle (*«der Tag», le jour → «täglich», quotidien*) (Trost, 1990).

Pour le traitement, nous avons recours aux symboles lexicaux «a_tréma», «o_tréma» et «u_tréma» qui marquent les voyelles sujettes à l'insertion du tréma. Comme ce phénomène ne se produit que devant des suffixes spécifiques, nous précisons à l'aide de traits le contexte dans lequel ce changement va avoir lieu. Ainsi pour les noms, la règle (20) indique que ce phénomène n'apparaît que lorsqu'on ajoute à la racine du nom un suffixe du pluriel :

```
(20) <=> - ä/a_tréma - Nom_suffixe[nombre=pluriel]
```

Le traitement n'est cependant pas encore complet : comme «a_tréma» ne correspond pas à une lettre de surface, nous devons spécifier comment il se réalise quand il se combine avec les autres suffixes. Au lieu de tous les énumérer dans une règle, nous pouvons exploiter les règles morphographémiques optionnelles, comme en (21). Celles-ci s'appliquent dans tous les cas sauf ceux qui sont traités par une règle obligatoire. Par exemple, lorsque nous ajoutons à un nom un des suffixes du génitif singulier, le symbole lexical «a_tréma» sera changé en «a» (*«das Haus», la maison → «des Hauses», de la maison*).

```
(21)       => - a/a_tréma -
```

b. **Morphologie dérivationnelle** : *mmorph* permet également de traiter la morphologie dérivationnelle. Notre morphologie ne traite pour l'instant que les cas suivants : les adjectifs et les infinitifs qui peuvent être nominalisés (*«rot», rouge → «das Rot», le rouge; «essen», manger → «das Essen», → le manger*) et les participes qu'on peut utiliser en tant qu'adjectifs (*«gelesen», lu → «das gelesene Buch», le livre qui a été lu*). L'exemple (22) montre une version simplifiée de la règle qui change un participe (présent, `prp` ou passé, `psp`) en un adjectif non fléchi. Dans (22), le signe «|» indique une disjonction : la règle s'applique si la valeur de `vform` est `prp` ou `psp`.

```
(22) Adjectif[type=non_fléchi degré=pos] Verbe[vform=prp|psp]
```

4. LEXIQUES

Les descriptions linguistiques existent pour le français, l'italien, l'allemand et l'anglais et sont disponibles sur demande[2]. Ils ont été créés de manière semi-automatique, à partir de dictionnaires existants et ont été validés à l'aide de diverses listes de mots

2 L'outil *mmorph* est disponible à l'URL http://www.issco.unige.ch/tools/.

complètes et de correcteurs d'orthographe. Le tableau (23) donne le nombre de mots de base et de formes dérivées.

	Français	Italien	Allemand	Anglais
Nombre de mots de base	22236	29621	45693	25955
Nombre de formes dérivées	235598	365929	3392697	53214

(23) Tableau

Précisons finalement que la description morphosyntaxique se base sur les spécifications développées dans le projet *Multext* (Calzolari et Monachini, 1995).

5. CONCLUSION

Dans cet article, nous avons montré la flexibilité et les possibilités de *mmorph*, à travers les descriptions linguistiques de plusieurs langues. Les lexiques continuent à être développés pour couvrir des domaines plus spécialisés et nous prévoyons plusieurs extensions de l'outil, en particulier pour le traitement des noms composés et les changements morphographémiques externes au mot, très fréquents en français («le++état» → *l'état;* «va++il» → *va-t-il*; etc.).

RÉFÉRENCES

ARMSTRONG, S. (1996) : «MULTEXT : Multilingual Text Tools and Corpora», H. Feldweg et E. W. Hinrichs, *Lexikon und Text*, Sonderdruck aus Lexicographica, Series Maior, Band 73, Max Niemeyer Verlag, pp. 107-119.

BOUILLON, P. et L. TOVENA (1990) : *L'analyse morphologique du français et de l'italien avec le lexique ALVEY*, ISSCO Working Paper, n° 57.

BOUILLON, P. et L. TOVENA (1991) : «Word formation and computational dictionaries», *Actes de TKE '90*, pp. 447-454.

CALZOLARI, N. et M. MONACHINI (1995) *: Common Specifications and Notation for Lexicon Encoding and Preliminary Proposal for the Tagsets*, Deliverable 1.6.1B, MULTEXT.

CARTER, D. (1995) : «Rapid development of morphological descriptions for full language processing systems», *Actes de ACL*, également sous SRI, Technical Report CRC-047.

KAPLAN, R. M. et M. KAY (1994) : «Regular Models of Phonological Rule Systems», *Computational Linguistics*, 20, 3, pp. 331-378.

KARTTUNEN, L. et A. K. WITTENBURG (1983) : «A two-level morphological analysis of English», J. Texas (dir), *Linguistics Forum*, 22, pp. 217-228.

KOSKENNIEMI, K. (1984) : «A General Computational Model for Word-form Recognition and Production», *Actes de COLING*, pp. 178-181.

OSTLING ANDERSON, A. (1987) : «L'identification automatique des lexèmes en français contemporain», *Acta Univ. Ups. Studia Romanica Upsuliensa* 39, Uppsala.

PETITPIERRE, D. et G. RUSSELL (1995) : *MMORPH - The Multext Morphology Program*, Version, 2.3, Deliverable 2.3.1, MULTEXT, ftp://issco-ftp.unige.ch/pub/multext/mmorph.doc.ps.tar.gz.

PULMAN, S. G. et M. R. HEPPLE (1993) : «A Feature-Based Formalism for Two-Level Phonology: A Description and Implementation», *Computer Speech and Language*, 7, pp. 333-358.

RAYNER, M., CARTER, D. et P. BOUILLON (1996) : «Adapting the Core Language Engine to French and Spanish», *Actes de la conférence internationale NLP+IA/TAL+AI*, Moncton, NB, Canada, pp. 224-232.

RITCHIE, G. D., RUSSELL, G. J., BLACK, A. W. and S. G. PULMAN (1992) *: Computational Morphology: Practical Mechanisms for the English Lexicon Cambridge*, MA, The MIT Press.

RUESSINK, H. (1989) : *Two-Level Formalisms*, Working Papers in Natural Language Processing, n° 5, Rijksuniversiteit Utrecht.

SHIEBER, S. M. (1986) : *An Introduction to Unification-Based Approaches to Grammar*, Stanford, CSLI.

THEOFILIDIS, A. et S. RIEDER (1995) : *Final Specifications on German Inflectional Morphology*, Deliverables D2.2 et D3.1, MLAB93-17.

TROST, H. (1990) : «The Application of Two-level Morphology to Non-Concatenative German Morphology», *Actes de COLING*, pp. 371-376.

COMMENT REPRÉSENTER L'EXPÉRIENCE INDIVIDUELLE QUI DONNE LEUR SENS AUX MOTS, APPROCHE INFORMATIQUE

Françoise FOREST

Groupe «Langage & cognition», LIMSI, Orsay, France

1. LE PROBLÈME DE LA REPRÉSENTATION DU SENS EN LANGUE

Du point de vue de l'informatique, notre travail se situe au carrefour de plusieurs domaines, le traitement automatique des langues, les systèmes de recherche d'information, la reconnaissance des formes, l'apprentissage automatique. Du point de vue des sciences cognitives, il a l'ambition d'étudier les possibilités de modélisation des relations qui se créent entre les mots, les perceptions et les abstractions au cours de l'acquisition et de la représentation des connaissances, de proposer une approche qui associe le perceptif, notamment dans sa dimension topologique et continue, et le conceptuel traditionnellement représenté comme discret. Les choix informatiques étant dictés par l'approche cognitive adoptée, nous commencerons par présenter et justifier cette dernière.

1.1 Le point de vue cognitif

L'hypothèse fondamentale sur laquelle s'appuie ce travail de recherche est l'**existence d'un sujet** dans le processus de compréhension. Il s'agit d'abord pour nous de représenter le sujet pour lequel l'énoncé émis fait sens. L'énoncé étant exprimé au travers d'une langue, nous nous intéressons aux liens que ce sujet a tissés, au cours de son histoire personnelle, entre les mots et les situations perçues, et tout particulièrement aux situations d'apprentissage langagier, celles au cours desquelles le sujet a pour la première fois mis en relation le monde perçu et les mots entendus. Pour développer cet aspect, nous nous inspirons principalement de travaux effectués en psycholinguistique par L.S. Vygotski (1985).

La deuxième hypothèse qui s'est avérée nécessaire est en relation avec la thèse d'un codage multiple des connaissances (Denis, 1989). Toute compréhension se ramenant à une transformation de représentation, il nous est apparu rapidement que la représentation informatique des situations perçues ne pouvait pas se réduire à une description propositionnelle. Nous faisons donc l'hypothèse d'une capacité que nous qualifions de **perception cognitive**. Il s'agit de rendre compte des situations perçues en termes d'entités qui interagissent dans un espace métrique. Cette hypothèse permet de proposer des représentations qui préservent les relations de proximité et de succession présentes dans la

situation. Nous nous inspirons principalement de travaux effectués par R. Thom (1977 et 1991).

La troisième caractéristique de notre travail, d'un point de vue cognitif, est la volonté de s'inspirer largement des particularités de la **langue des signes** dans la représentation de l'expérience acquise par le sujet. Deux aspects de cette langue nous intéressent. Tout d'abord son expression spatiale et temporelle permet de dessiner les situations, et donc d'en fournir simplement une représentation telle que celle souhaitée plus haut. Ensuite la capacité de la langue des signes à «dessiner» aussi simplement des situations abstraites que des situations concrètes suggère que la représentation spatio-temporelle proposée est adaptée à la mise au point de processus de compréhension par analogie qui s'apparentent à ceux qu'on peut imaginer à l'œuvre quand il s'agit d'exprimer des notions abstraites à l'aide de gestes.

1.1.1 Les études sur l'acquisition du langage chez l'enfant, les travaux de Vygotski

Les études sur l'acquisition du langage chez l'enfant mettent en évidence l'importance de la communication langagière avec les adultes au cours de la petite enfance, ainsi que l'importance de l'expérience perceptive et affective vécue par l'enfant.

Pour Vygotski, cette acquisition est la face visible d'une acquisition plus importante, celle des concepts.

Vygotski est un psycholinguiste russe du début du vingtième siècle qui s'est notamment intéressé au problème de l'apprentissage langagier chez l'enfant. Il aborde l'étude de la pensée verbale du point de vue développemental. Dans son ouvrage *Pensée et Langage*, il distingue les concepts «spontanés», qui s'acquièrent au cours de l'expérience quotidienne de l'enfant, expérience soutenue par l'interaction qu'il entretient avec les adultes, et les concepts «scientifiques», qui sont d'abord rencontrés par l'enfant sous forme de définitions verbales. Pour lui, un concept n'existe que dans un réseau de significations. C'est une structure dynamique qui se construit au cours de la résolution d'un problème, et évolue constamment au cours de l'évolution du développement de l'individu.

En ce qui concerne l'acquisition des concepts spontanés, Vygotski met en évidence l'existence intermédiaire d'agglomérats qu'il nomme **complexes**. Ces complexes réunissent un ensemble d'expériences individuelles présentant une certaine similitude aux yeux de l'enfant, et qu'une ou plusieurs caractéristiques perceptives rapprochent. Parmi ces caractéristiques perceptives communes, un rôle très particulier est joué par les mots énoncés par les adultes pour nommer les entités perçues. Vygotski justifie ainsi l'importance primordiale de l'interaction langagière avec les adultes dans l'étape intermédiaire de construction des complexes. Le mot joue dans cette construction un rôle central, puisqu'il est d'abord *objet de perception* avant de devenir également *élément de signification*. C'est sa présence dans le psychisme de l'enfant qui permet à celui-ci d'abstraire une notion de l'ensemble de ses expériences perceptives.

On trouvera dans Forest et Siksou (1994) un exposé de l'intérêt que nous lui trouvons encore aujourd'hui dans une perspective de modélisation du sens.

1.1.2 Un ancrage perceptif de la structuration des connaissances

Les résultats de travaux plus récents sur l'acquisition du langage chez l'enfant montrent qu'un enfant est, même dans son stade de développement préverbal, capable de distinguer les entités en jeu dans une situation qu'il perçoit (Rondal, 1983). Nous ajoutons qu'il est capable d'effectuer un repérage de ces entités dans son champ de vision (plus haut, plus bas, à gauche, à droite, contre, dans...). Cette capacité de reconnaissance perceptive est soutenue par le dialogue que l'enfant poursuit à chaque instant avec son entourage. En effet, on constate que les adultes en interaction avec un très jeune enfant utilisent de nombreux moyens pour l'aider à catégoriser le monde et à agir sur lui (gestes et intonations soutenant le message verbal, actions d'accompagnement, appels à la dimension affective...). Nous parlerons d'une *dimension cognitive de la perception* qui nous paraît ressortir de la même approche que celle utilisée dans des travaux comme ceux de Langacker (1986) sur la grammaire cognitive, de Thom et de Petitot (Petitot, 1991) sur la sémiophysique et la morphodynamique.

Thom appuie son travail sur la notion de discontinuité. Dans le domaine de la perception, il s'agit à la fois de discontinuités spatiales et temporelles. C'est la capacité de reconnaissance des discontinuités qui permet de percevoir les différentes entités en jeu dans une situation perçue. Thom parle à leur propos de formes saillantes dans un espace substrat. Concernant la représentation de cette perception dans le langage, il propose seize morphologies archétypes correspondant à seize verbes d'action ou d'état, et interprétables comme les transformations de relations topologiques entre les entités interagissant dans une situation. On rencontre aussi chez Langacker cette proposition d'une représentation topologique dynamique des objets du discours.

Ces approches ont en commun de proposer une représentation spatiale et dynamique des situations décrites ou observées. Ce type de représentation, associé à des descriptions langagières, nous rapproche de l'hypothèse de double codage (Pavio, Denis). Nous pensons que ce peut être un fondement productif pour une modélisation opérationnelle du phénomène de compréhension. Nous verrons plus loin comment ces approches sont prises en compte dans le modèle que nous proposons.

1.1.3 Les apports de la langue des signes

L'identification des morphologies archétypes proposées par Thom est particulièrement claire dans les langues gestuelles. La langue des signes, tout en disposant, comme les langues orales, d'une panoplie de signes conventionnels (équivalents des mots) pour distinguer les entités du monde perçu, conserve des éléments de nature non discrète servant à construire une représentation spatio-temporelle de la situation décrite dans l'énoncé. Cette représentation rend explicites les relations topologiques entre entités mises en jeu dans cette situation. La capacité de cette langue à exprimer aussi bien des notions abstraites que concrètes suggère une mémorisation spatio-temporelle de tous ces types de situations, les quatre dimensions devenant métaphoriques dans le cas de situations abstraites. L'étude de la forme des énoncés de la langue des signes nous conforte dans notre choix d'une représentation spatiale et temporelle des situations. Nous y voyons la possibilité d'une mise en action simple de processus de compréhension par analogie.

1.2 Le point de vue informatique

1.2.1 Refuser une formalisation a priori des relations de sens

Une des grandes préoccupations du traitement automatique des langues est de mettre en relation des énoncés exprimés en langage naturel et des représentations de ces énoncés traduisant leur signification dans un contexte d'application donné.

Les outils généralement utilisés sont les formalismes logiques manipulant des symboles, ces symboles étant des représentations des mots ou de leur signification (les concepts).

Lorsque les énoncés sont pris dans des textes plus longs que des phrases, ils ont pour rôle de décrire des situations statiques ou dynamiques qu'on se propose de représenter également, par exemple sous forme de schémas, à la Schank. Les structures de données manipulées et les outils logiques pour le faire sont généralement construits de façon ad hoc pour le domaine, généralement restreint, sur lequel ils interviennent.

On a beaucoup reproché à cette approche sa rigidité, son manque de validité sur des domaines plus larges, et les difficultés qu'on rencontre quand on doit faire évoluer les représentations, par exemple les problèmes posés par l'apprentissage de nouvelles connaissances.

1.2.2 Faire l'hypothèse d'une émergence statistique du sens

Les systèmes de recherche documentaire se fondent assez généralement sur cette hypothèse. Ayant pour objectif de retrouver une information stockée sous forme textuelle, dans une base généralement très importante de textes, les outils qu'ils mettent en œuvre sont des outils statistiques. Ils sont robustes et efficaces, en ce sens qu'ils permettent de retrouver en proportion satisfaisante des informations pertinentes malgré le peu de contraintes imposées à la requête de l'utilisateur. Par ailleurs, ils sont souples et permettent de s'adapter assez simplement à un nouveau domaine d'utilisation. En revanche, ne prenant en considération que des fréquences de cooccurrence de mots ou de syntagmes, ils négligent la structure sémantique des connaissances implicitement manipulées. Mais s'ils fonctionnent, c'est sans doute parce que, justement, la cooccurrence des mots dans les textes traduit certaines relations sémantiques. C'est cette idée d'une émergence statistique du sens que nous retenons de cette approche.

1.2.3 Une mise en œuvre de la notion d'expérience individuelle comme fondement de la notion de sens

Les deux approches précédentes ont pour défaut selon nous de se cantonner dans l'étude de données linguistiques. Les travaux effectués en psychologie montrent que la notion de sens, celle de compréhension, ne peuvent se saisir sans faire appel à l'existence du sujet pour lequel cette compréhension a lieu, ce sens se construit. C'est dire que la connaissance ne peut s'appréhender sans une mise en relation de la langue et de l'expérience individuelle, notamment au cours de l'apprentissage du langage. La prise en compte de l'expérience individuelle recouvre, du point de vue de l'informatique, deux types de techniques.

les techniques connexionnistes

L'aspect cumulatif de l'expérience acquise par un individu, les phénomènes de renforcement, d'assimilation, orientent le traitement vers des outils connexionnistes. L'étude des modèles traditionnels (réseaux à couches, cartes de Kohonen et même réseaux récurrents) et des liens qu'ils entretiennent avec les méthodes statiques d'analyse de données nous ont amenée à choisir de développer un réseau à propagation d'activité qui soit exactement adapté à la modélisation des phénomènes décrits dans la littérature psycholinguistique à laquelle nous nous référons (poids de l'affectif dans la mémorisation, importance des liens de cooccurrence entre les mots et leurs situations d'apprentissage, phénomène d'oubli...). Cette structure est détaillée dans le paragraphe de présentation du modèle MoHA. Elle a donné lieu récemment à un travail de thèse effectué par Jean-Pierre Gruselle et à soutenir en octobre prochain.

les techniques de reconnaissance de formes

L'expérience acquise par un individu est d'abord perceptive, généralement centrée sur le visuel. Si l'on dépasse le niveau 2D de perception visuelle, l'acquisition de cette expérience met en jeu des notions de voisinage, d'orientation, d'intensité, qui sont absentes du traitement automatique symbolique classique et qui pourtant entrent en compte dans son ancrage dans le réel perçu de l'expérience individuelle. Plus que la reconnaissance de formes proprement dite, c'est l'aspect topologique des représentations manipulées dans ce domaine de recherche qui me paraît intéressant. Il s'agit de transposer dans le domaine de la représentation de connaissances abstraites les notions de voisinage et de connexité qui permettent de dire que deux formes sont plus ou moins semblables, et de situer les formes les unes par rapport aux autres dans un espace.

Nous pensons qu'une représentation de ce type, utilisable pour les connaissances concrètes visuelles, doit l'être aussi pour les connaissances abstraites. Les travaux menés par les linguistes sur la langue des signes attestent de la possibilité d'une dimension spatiale dans l'expression de toutes les notions accessibles à l'esprit humain (Cuxac, 1993).

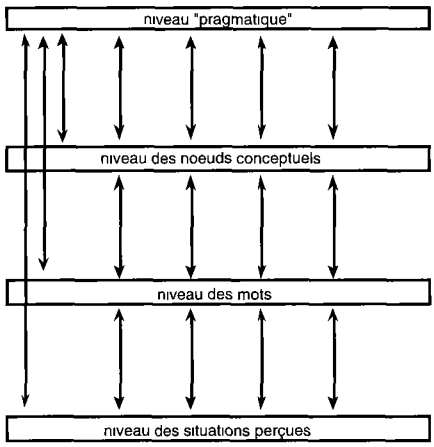
une modélisation opérationnelle sur machine SIMD

Si nous posons qu'une situation perçue est délimitable dans le temps et dans l'espace (Bachelard), que le lien entre le mot perçu et l'objet désigné est clair en situation d'apprentissage (Rondal), que le nombre des échanges langagiers entre enfants et parents est de l'ordre de quelques millions pendant la période d'apprentissage du langage chez le jeune enfant (Lambert et Rondal), une modélisation de la notion de sens nous paraît accessible à l'informatique. Il devient en effet justifié de représenter explicitement les mots, les situations, les occurrences d'entités observées, la taille des corpus autorisant une exploitation réaliste sur les machines massivement parallèles actuelles.

2. UNE PROPOSITION DE MODÉLISATION INFORMATIQUE

2.1 MoHA, Modèle Hybride d'Apprentissage

Le groupe de travail MoHA rassemble au sein du groupe «Langage et cognition» du LIMSI plusieurs chercheurs et doctorants, autour de la construction d'un modèle hybride d'apprentissage. Nous nous situons dans le domaine de l'acquisition de connaissances sémantiques et pragmatiques à partir de données linguistiques et perceptives. Nous faisons l'hypothèse que l'apprentissage ne se réduit pas à des processus ascendants, allant des perceptions vers des connaissances abstraites de différents types, mais qu'il nécessite de mettre en relation l'information nouvellement acquise avec les connaissances déjà structurées, celles-ci guidant le processus d'apprentissage au moyen d'interactions descendantes entre les différents niveaux. Nous faisons également l'hypothèse que l'apprentissage n'est pas un processus séparé des autres processus impliqués dans un système de compréhension, mais qu'il intervient de façon continue, tandis que l'expérience de l'individu s'enrichit. C'est dans l'objectif de modéliser ces phénomènes que nous proposons MoHA, un modèle hybride qui combine une approche numérique et une approche symbolique (Forest et Grau, 1992).



Les différents niveaux de représentation des connaissances et leurs interactions dans le système MoHA

La *première approche* consiste en un traitement massivement parallèle d'événements du monde réel tels que les perçoit un individu. Jusqu'à présent, seule la perception verbale était considérée, ces événements et leur perception verbale étant représentés dans un réseau de type connexionniste, appelé graphe de premier niveau. Il s'agit d'un graphe bipartite qui associe les situations mémorisées aux mots qui en ont accompagné la perception, les liens étant pondérés par des valeurs numériques représentant le poids affectif accordé à cet instant de la mémorisation. Notre objectif est d'en faire émerger des attracteurs qui constituent les concepts «stables». Le rôle du mot dans la formation de ces concepts stables a été étudié, avec l'objectif d'une implantation informatique, dans la thèse de J.P. Gruselle. Les différentes étapes de la construction y

sont spécifiés informatiquement : perception des mots décrivant des situations, mémorisation des situations en mémoire à long terme, émergence de catégories.

La *deuxième approche* est symbolique. Les concepts interviennent dans des informations de plus haut niveau que nous représentons sous forme de schémas. L'apprentissage incrémental de ces schémas s'effectue à l'aide de traitements symboliques tels que l'analogie, la généralisation et la spécification.

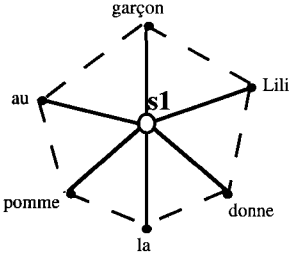
Le modèle MoHA est présenté plus en détail dans un article rédigé en collaboration avec François Bordeaux et Brigitte Grau (Bordeaux, Forest et Grau, 1992).

Nous détaillons ci-dessous les points du modèle MoHA que notre travail actuel aborde plus spécifiquement.

2.2 Le graphe de premier niveau : enrichir la représentation des situations mémorisées

Le graphe de premier niveau rassemble les informations concernant les mots et les situations perçues au cours de l'apprentissage de ces mots.

Chaque situation perçue contribue à ce graphe bipartite qui est constitué de nœuds-mots et de nœuds-situations, chaque nœud-mot étant relié à tous les nœuds-situations représentant des situations à la description desquelles le mot contribue.

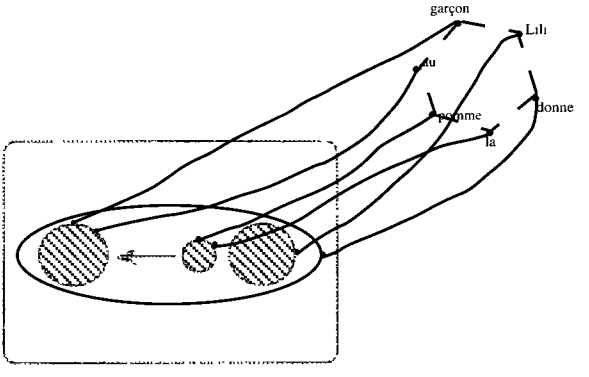


Contribution de la situation s1 décrite par l'énoncé «Lili donne la pomme au garçon» au graphe de premier niveau (1)

Le choix du nombre de mots et de l'importance relative de chaque mot intervenant dans la description ont été particulièrement étudiés par Jean-Pierre Gruselle qui propose plusieurs critères de sélection des mots perçus et différentes lois de diffusion d'activité dans le graphe, simulant le phénomène d'association et permettant notamment de proposer une explication au fait que, dans un contexte donné, un mot n'est généralement pas ambigu.

Pour enrichir la représentation de la situation perçue s1 dans le graphe de premier niveau, il faut substituer au nœud s1 une structure qui fasse apparaître des éléments constitutifs de cette situation, notamment des informations visuelles.

Des travaux ont déjà été réalisés dans notre groupe pour ce qui concerne les situations perçues visuellement. Mais en voulant partir de représentations visuelles en 2D, ces travaux se sont heurtés à des difficultés qui nous éloignaient de notre propos. Un moyen de contourner cet obstacle est de considérer cette modélisation sous l'angle de la morphodynamique, en acceptant d'emblée l'existence d'un niveau cognitif de la perception.



Contribution de la situation s1 au graphe de premier niveau (2) La situation n'est plus représentée uniquement par un nœud dans le graphe de premier niveau, mais par l'ensemble des actants de l'action de «donner» schématisée ici par la flèche

Nous proposons donc de représenter une situation par l'ensemble des actants de l'action décrite, leurs positions relatives dans la scène, et l'évolution de ces positions au cours du temps, notamment l'apparition ou la disparition éventuelle d'une des entités impliquées dans l'action. Nous détaillerons plus loin la forme proposée pour représenter ces informations.

2.3 Le passage au symbolique

Ce passage se manifeste par l'émergence des concepts et l'existence de liens rattachant les situations perçues aux informations conceptuelles et aux schémas du niveau pragmatique.

Si l'on suit l'approche de Vygotski sur la formation des concepts spontanés, le passage le plus difficile à modéliser est certainement le phénomène de «saut qualitatif», celui qui permet à l'enfant de passer de la pensée par complexes à la pensée par concepts.

On a vu comment nous proposons de modéliser les complexes et leur relation aux mots appris. Le passage à une véritable conceptualisation suppose la détection de sous-graphes dont la connectivité interne soit suffisamment forte et les liens aux autres nœuds suffisamment faible pour pouvoir parler d'une certaine «stabilité» du sous-graphe. J.P. Gruselle propose un algorithme de construction des nœuds-complexes associés à ces sous-graphes stables. On peut, dans un premier temps, les assimiler à des nœuds conceptuels, au sens de prototypes d'une catégorie.

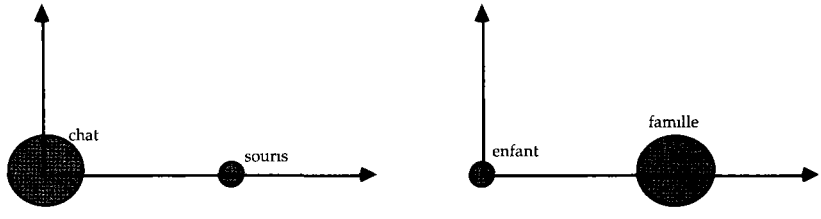
La représentation des entités mémorisées ne se limite pas à des boules dans un espace spatio-temporel. Par l'intermédiaire de l'apprentissage langagier, un certain nombre de caractéristiques verbales sont associées à chaque entité, à commencer par son nom. Si la catégorie abstraite correspondant à la pomme est déjà formée, on associe à la boule un pointeur vers une instance nouvellement créée du prototype de cette catégorie. Sinon, par l'intermédiaire du mot, cette nouvelle entité est mise en relation avec toutes les pommes déjà rencontrées, c'est-à-dire tous les constituants du complexe de pomme.

3. LA REPRÉSENTATION SPATIALE ET DYNAMIQUE DES SITUATIONS

Nous distinguons les *situations statiques*, considérées comme des images fixes, dont l'exploration mentale se fait entité par entité, comme on parcourrait des yeux un paysage, et les *situations dynamiques*, considérées comme une succession de situations statiques à l'intérieur desquelles on aurait la capacité de parcourir transversalement l'évolution d'une entité donnée suivant l'axe temporel. Si on suppose qu'une situation statique est représentée dans un espace à N dimensions, une situation dynamique le sera dans un espace à N+1 dimensions, la dernière représentant l'axe temporel.

3.1 Les situations statiques

Dans une situation statique, les entités présentes sont placées dans un espace métrique à une, deux ou trois dimensions. Ce sont des boules déterminées par leur centre qui indique leur placement dans l'espace, et leur rayon qui peut indiquer leur taille, mais aussi leur saillance dans la situation en fonction d'un certain critère.



Représentation du degré relatif de saillance de deux entités intervenant dans une situation

Un gros *chat* et une petite *souris* n'auront pas le même rayon. Parce qu'ils ne sont pas visuellement de la même taille. Mais aussi parce qu'il est possible de mettre la souris dans le chat, et la possibilité d'établir cette relation topologique suppose que la boule chat puisse inclure la boule souris. Pour des raisons semblables, les entités *enfant* et *famille* seront représentées par des boules de tailles différentes permettant de construire la relation topologique d'inclusion de l'enfant dans la famille.

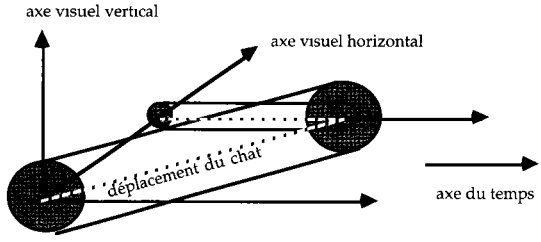
Au démarrage de la construction de MoHA, les premières scènes mémorisées doivent être des scènes concrètes. Mais lorsque le système dispose d'un nombre suffisant de situations concrètes mémorisées, mettant en jeu différentes relations topologiques, on

peut alors, par analogie, construire des représentations de scènes abstraites en utilisant les mêmes techniques. La grande quantité d'expressions et de verbes spatiaux présents dans la langue atteste de la réalité de cette analogie.

3.2 Les situations dynamiques

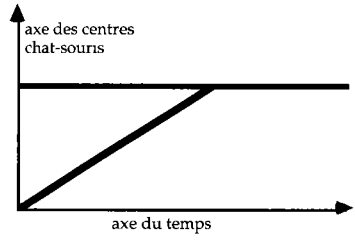
Si les situations statiques peuvent être considérées comme des *instantanés* subjectifs, les situations dynamiques sont des *films* subjectifs. Leur représentation nécessite une dimension supplémentaire qui est celle du temps.

Dans l'espace à N+1 dimensions (N dimensions spatiales et la dimension temporelle), la situation qui évolue est représentée par une forme. Si le chat se précipite sur la souris, la situation dynamique correspondant à la situation statique dessinée plus haut (qui en est le point de départ), aura la représentation donnée ci-après.



Représentation de la situation relative à l'énoncé : «le chat attrape la souris»

En projetant cette forme sur un plan parallèle à l'axe des temps et contenant l'axe des centres chat et souris, on trouve le schéma suivant.

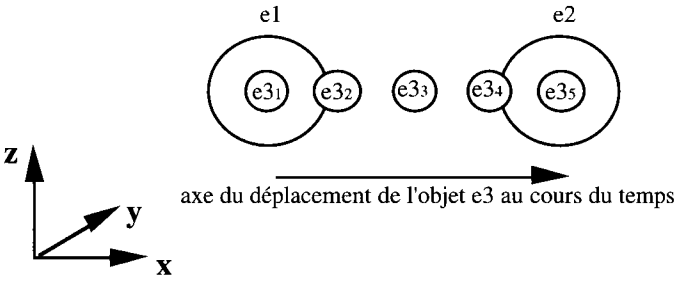


Projection sur le plan (temps, axe des centres chat-souris)

On reconnaît dans ce schéma l'une des morphologies archétypes (Thom, 1977).

3.3 Exemple de représentation dans R5 d'une situation dynamique

On choisit les trois dimensions spatiales, une dimension de prégnance exprimant l'importance relative des différentes entités intervenant dans la situation, et la dimension temporelle.



#donner :
e1 (agent) : (x=0, y=0, z=0, r=50, t=0), (x=0, y=0, z=0, r=50, t=10)
e2 (patient) : (x=200, y=0, z=0, r=50, t=0), (x=200, y=0, z=0, r=50, t=10)
e3 (objet) : (x=0, y=0, z=0, r=10, t=0), (x=200, y=0, z=0, r=10, t=10)

Représentation de la situation *#donner... quelque chose à quelqu'un.* Superposition des plans (x,y) pris à 5 instants successifs de l'évolution de la situation. e3 se déplace de e1 vers e2

La dimension de prégnance est subjective. C'est celle qui est caractérisée visuellement par la taille de chaque entité dans l'expression en langue des signes (ici par le rayon des boules). Elle permet d'effectuer des comparaisons analogiques qui se ramènent à des comparaisons de formes, à l'aide de rotations et d'homotéties, dans R5.

3.4 Les situations concrètes reconstruites à partir d'un énoncé

Toute scène intelligible est constituée d'entités qui interagissent. La perception de chaque entité se traduit par la création d'une entité virtuelle dans l'espace mental où se reconstruit cette scène.

Si la scène est visuelle et directement perçue, sa représentation est déduite de la scène réelle. En revanche, si la scène est décrite par un énoncé émis par un locuteur, le travail de l'interlocuteur consiste à reconstituer mentalement la scène à partir des indices linguistiques fournis par l'énoncé et de l'expérience qu'il a préalablement acquise en liaison avec sa propre utilisation de chacun des mots de l'énoncé. Dans ce cas, l'interlocuteur crée une entité virtuelle pour chaque entité dont il est capable de repérer l'existence dans le discours du locuteur, et place cette entité virtuelle dans son espace mental. La place de l'entité virtuelle dans son «champ de vision» mental est soit déterminée par le discours, soit inférée des connaissances préalables. L'existence d'un «champ de vision» suppose que cet espace est construit à partir des perceptions visuelles habituelles. Comme nous l'avons dit plus haut (§1.1.2), nous faisons l'hypothèse que les notions de haut et de bas, de droite et de gauche... y sont déjà présentes au moment de la perception.

Cas où toutes les entités en jeu sont nommées : «La souris est sur l'armoire» fait linguistiquement référence à une scène dans laquelle deux entités entrent en jeu (la souris et l'armoire) dont la relation topologique est déterminée par l'existence de la préposition «sur».

Cas où certaines entités en jeu ne sont pas nommées : «La souris est là-haut» fait linguistiquement référence à la seule souris. «là-haut» détermine sa place en hauteur dans le champ de vision. Cette connaissance peut s'exprimer par le prédicat Haut(souris). Comment en inférer l'existence d'un objet sur lequel la souris devrait se tenir pour que la scène soit acceptable ?

Si, à chaque fois qu'on a eu l'occasion d'observer une souris en hauteur, elle était posée sur un objet, l'examen des situations mémorisées contenant une souris devrait permettre de mettre en évidence la présence obligée de cet objet, même s'il n'en est pas fait mention dans l'énoncé. Le repérage de cette entité non nommée se fait par superposition de toutes les situations concernées, la superposition étant centrée sur l'entité virtuelle associée au mot «souris».

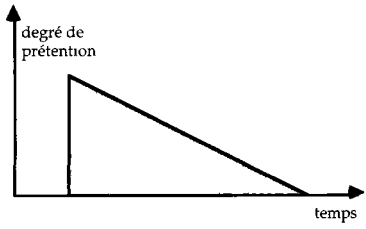
L'observation des relations topologiques qui la concernent dans chaque situation mémorisée permet de mettre en évidence l'existence d'un objet situé dessous, et en contact avec elle. Cela signifie que, dans chacune de ces situations, on trouvera, à la fois sous forme symbolique et sous forme géométrique les informations

Haut(souris), Sur(souris,x), Contact (souris,x).

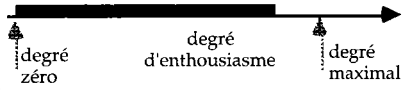
La précision de la description de l'entité devant s'identifier à x dans la situation à reconstruire dépend de l'approche qu'on adopte vis-à-vis de la catégorisation. Soit l'entité reste «floue», on ne lui associe que les caractéristiques communes à toutes les observations précédentes, et leur domaine de valeurs possibles. Soit on crée une instance du prototype de la catégorie. Soit on copie à l'identique l'une des instances rencontrées dans une des situations précédemment observées, par exemple celle de la situation la plus prégnante dans le contexte de l'énoncé, c'est-à-dire celle qui est le plus fortement remémorée à la suite de la perception de l'énoncé.

3.5 Les situations abstraites

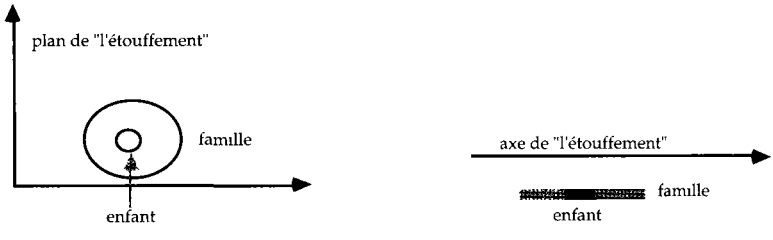
Même abstraite, une situation concerne toujours des entités qui interagissent. Mais l'interprétation des dimensions spatiales dans lesquelles cette situation est représentée est métaphorique. Chacune des dimensions vaut pour un trait, commun aux entités présentes, et qui ordonne ces entités suivant un ordre total. L'axe des temps demeure en général, même si les valeurs prises par les différentes entités y sont subjectives. Tous les autres axes peuvent être le lieu de relations topologiques entre les entités.



«une prétention qui diminue»

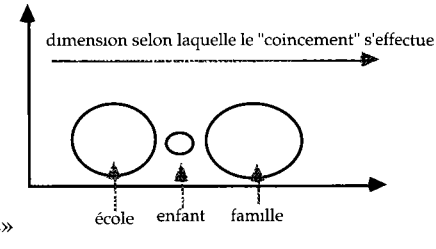


«un enthousiasme fantastique»



«l'enfant étouffé par sa famille»

La relation topologique perçue est la relation d'inclusion.



«l'enfant coincé entre l'école et sa famille»

Les volumes représentent la saillance de chaque entité dans la représentation que construit l'interlocuteur.

4. CONCLUSION

Cette représentation spatiale et temporelle des situations enrichit le graphe de premier niveau. Elle permet d'accéder aux éléments constitutifs de chaque situation dans un format non propositionnel qu'on peut assimiler à un codage de nature différente du codage linguistique utilisé dans le niveau symbolique.

La dimension continue de cette représentation et son caractère topologique permettent d'effectuer des opérations de comparaison, des mesures de proximité interdites dans l'autre type de codage. Elles permettent aussi de mettre en évidence des similarités entre situations qui sont de l'ordre de l'«analogie d'évolution». En assimilant les situations abstraites à des situations concrètes à dimensions effectivement spatiales, on se donne la possibilité d'anticiper l'évolution d'une situation abstraite par comparaison avec des situations concrètes. Dans le cadre d'une approche développementale des connaissances acquises, on voit tout l'intérêt de cette assimilation de l'abstrait au concret, les psycholinguistes ayant généralement observé que l'acquisition des mots désignant des entités concrètes, puis les relations qui les lient, précèdent chez l'enfant l'acquisition des connaissances abstraites.

RÉFÉRENCES

BORDEAUX, F., FOREST, F. et B. GRAU (1992) : «MoHA, an hybrid learning model: a model based on the perception of the environment by an individual», *IPMU'92-Advanced Methods in Artificial Intelligence*, B. Bouchon-Meunier, L. Valverde, R. R. Yager (Eds), *Lecture Notes in Computer Science*, n° 682, Springer-Verlag.

CUXAC, C. (1993) : *L'icônicité dans la langue des signes*, VENACO.

DENIS, M. (1989) : *Images et cognition*, Paris, PUF.

FERRET, O. et B. GRAU (1996) : «Construire une mémoire épisodique à partir de textes : pourquoi et comment ?», *RFIA'96*-Rennes, 16-18 janvier 1996.

FOREST, F. (1990) : «Le sens d'un énoncé est fondamentalement lié à l'expérience de l'individu qui le perçoit», *4ème colloque de l'ARC* - Paris, 28-30 mars 1990.

FOREST, F. (1991) : «Se donner les moyens d'une approche constructiviste de la représentation du sens, le traitement massivement parallèle des données, *Notes et documents du LIMSI*, n° 91-21, 12-1991.

FOREST, F. et M. SIKSOU (1994) : «Développement de concepts et programmation du sens, Pensée et Langage chez Vygotski», *Intellectica*, vol 1, n° 18, pp. 213-236.

LAMBERT et RONDAL (1979) : *Le mongolisme*, Bruxelles, Pierre Mardaga éditeur.

LANGACKER, R. (1986) : «An introduction to cognitive grammar», *Cognitive Science*, n° 10, pp. 1-40.

PETITOT, Jean (1991) : «Syntaxe topologique et grammaire cognitive», *L'objet, sens et réalité, Langages*, n° 103, septembre.

RIVIERE, A. (1990) : *La psychologie de Vygotsky,* Bruxelles, Mardaga.

RONDAL, Jean-A. (1983) : *L'interaction adulte-enfant et la construction du langage,* Bruxelles, Mardaga.

THOM, René (1977) : *Stabilité structurelle et morphogénèse,* 2e édition, Paris, Interéditions.

THOM, René (1991) : *Esquisse d'une sémiophysique,* Paris, Interéditions, 2e tirage corrigé.

VYGOTSKI, L.S. (1985) : *Pensée et langage,* Paris, Éditions sociales (traduction de Myschlenie y rech' écrit en 1933 et publié dans *Izbrannye psikhologicheshie issledovanya* en 1956 à Moscou).