Les emprunts : du repérage aux analyses. Diversité des objectifs et des traitements

Jean-François Sablayrolles, Christine Jacquet-Pfau

HAL JFS 28 2008d

Sablayrolles Jean-François (Paris 13 et LDI UMR 7187) et Jacquet-Pfau Christine (Collège de France et LDI)

« Les emprunts : du repérage aux analyses. Diversité des objectifs et des traitements »

Neologica n° 2, Garnier, 2008, p. 19-38.

Résumé

Deux exposés, complémentaires et non opposés, visent à illustrer la complexité du concept d'emprunt : des points de vue et des objectifs différents conduisent à des pratiques différentes. Cette complexité du concept d'emprunt est responsable de nombre de malentendus à dissiper. JFS s'intéresse au processus (l'emprunt est un des – trois – types fondamentaux traditionnellement reconnus de la néologie) et examine les difficultés concrètes auxquelles on se heurte dans l'analyse des néologismes relevés et dans l'identification de leur matrice lexicale. Quelques propositions sont esquissées. La dimension processuelle est en revanche absente des préoccupations de CJP. Son objectif est le repérage, dans des productions écrites, d'unités lexicales dont certaines particularités formelles révèlent une origine étrangère, donc non intégrée au système de la langue emprunteuse. Ces informations sont utilisables dans différentes étapes et en vue de différentes applications de l'analyse automatique : recherche d'information et indexation, identification de types de textes, de spécialité en particulier, domaines du savoir dont ils traitent, reconnaissance des langues utilisées, notamment dans le contexte de la traduction…

Mot-clés : néologisme, emprunt, processus, particularités graphiques, TAL

Les emprunts : du repérage aux analyses. Diversité des objectifs et des traitements

Un consensus de façade sur le terme d'*emprunt* cache en fait des conceptions et des pratiques linguistiques diverses. L'objectif de cette contribution à deux voix consiste à explorer deux approches non pas divergentes mais complémentaires — et inconciliables. La première se situe dans le cadre de la néologie : l'emprunt est l'apparition d'une nouvelle unité lexicale, c'est un événement linguistique qui surgit à un moment donné, dans des circonstances données. La deuxième approche consiste dans le repérage, dans des productions écrites, d'unités présentant des traits ne relevant pas du système français et dans l'étude des informations que ce repérage donne sur la nature des types de textes où elles se trouvent. Une telle différence d'approches, parmi d'autres, devrait conduire les linguistes à toujours préciser leurs objectifs et les outils dont ils se dotent ou auxquels ils recourent quand ils travaillent sur l'emprunt.

1. Néologismes et emprunts : quelques problèmes

L'emprunt constitue, avec la néologie de forme et la néologie de sens, un des trois grands groupes définis par la majorité des typologies de néologismes. Mais, comme les emprunts affectent soit la forme soit le sens, certaines d'entre elles ne distinguent que ces deux derniers groupes. Cette solution, malgré sa logique, est bien expéditive, car en tant que matrice externe s'opposant aux matrices internes relevant du système de la langue, l'emprunt est bien un concept pertinent dans le traitement de la néologie. Mais, loin d'être évidente, son utilisation pose de nombreux problèmes tant théoriques que pratiques, que deux exemples suffiront à illustrer. Lors de la fabrication d'une base de

données incluant une grille de matrices lexicales comprenant l'emprunt[1], une collègue de Paris 7 (Hélène Béciri) a scindé cette grille en deux, en créant deux champs indépendants, avec chacun leur menu déroulant, l'un pour les matrices internes et l'autre pour la matrice externe qu'est l'emprunt. Ce changement n'est pas un simple réaménagement de surface mais il correspond à des conceptions fondamentalement différentes, que cette contribution se propose d'éclairer. Par ailleurs les collectes de quatre chercheurs de l'équipe néologie du LLI (Laboratoire de Linguistique Informatique, de Paris 13) sur un même corpus de presse, dans une expérience sur le sentiment néologique, s'avèrent très disparates. 86 items ont été relevés et identifiés comme emprunts au moins une fois. Mais la somme des emprunts des quatre listes ne s'élève qu'à 110. Ce qui signifie que le taux de recouvrement des quatre collectes est faible. Si les quatre collecteurs les avaient tous relevés, on obtiendrait un total de 384. On est donc loin du compte. De fait, seule une identification est commune aux quatre collecteurs, 3 le sont à trois, et 15 le sont à deux. Seuls 19 items sur 86 font donc l'objet d'une analyse partagée par au moins deux des quatre collecteurs. Ce qui prédomine, ce sont les analyses strictement individuelles qui s'élèvent à 67 (22+21+21+3). Ces difficultés ne peuvent être surmontées et des solutions proposées qu'à condition que soient préalablement levées quelques ambiguïtés sur le terme *emprunt*, que soient délimités explicitement les objectifs que l'on se fixe et que soient précisés les outils que l'on se donne.

Notre objectif étant l'étude des néologismes français contemporains, il est nécessaire que l'identification des matrices qui les ont produits soit cohérente et fiable. C'est le seul moyen de voir quels types de néologismes prévalent dans tel ou tel domaine du savoir, dans telle ou telle situation d'énonciation, etc. C'est aussi le seul moyen de mesurer le plus objectivement possible le poids et les manifestations de l'influence de langues étrangères dans la néologie actuelle.

Une première confusion à lever — mais ce n'est ni évident ni facile, tant elle est ancrée dans des habitudes intellectuelles héritées et jamais questionnées — repose sur l'amalgame de deux niveaux d'analyse, tous deux pertinents mais spécifiques. Il s'agit, d'une part, de l'analyse, en synchronie, de la structure morphologique d'une lexie et, d'autre part, de l'identification de la matrice lexicale responsable de l'apparition de cette lexie à un moment donné de l'histoire de la langue[2]. Cette confusion tient au fait que ces deux niveaux coïncident parfois, mais c'est loin d'être toujours le cas. Si *emprunt* signifie « fait d'emprunter », le phénomène dure ce que dure un néologisme : quelques années avant disparition ou intégration. En revanche si *emprunt* s'applique à tous les mots dont on sait, soit par l'interprétation de marques (parfois trompeuses), soit par érudition, qu'ils ne sont pas de création française, ils sont alors très nombreux puisqu'ils restent emprunts sans jamais perdre ce statut, comme *alcool*, *pintade*, *redingote* ou *week end*, par exemple. C'est une des principales causes de divergence d'incorporation entre les quatre collecteurs dans l'expérience en cours évoquée précédemment. Travaillant sur la néologie, un phénomène dynamique, nous décidons logiquement de réserver le terme *emprunt* à la première acception, c'est-à-dire à la matrice qui fait introduire dans des énoncés français des lexies existant dans d'autres langues et absentes dans un état immédiatement antérieur de la langue française.

---

[1] Cette grille inspirée de Jean Tournier (1985) mais présentée remaniée dans Sablayrolles (2000 et 2003) est fortement hiérarchisée (il y a jusqu'à cinq niveaux d'enchâssement). Elle oppose d'abord l'ensemble des matrices internes à la matrice externe qu'est l'emprunt. Les matrices internes se répartissent en matrices morpho-sémantiques, syntactico-sémantiques, morphologiques et pragmatico-sémantiques. Chacun de ces ensembles se subdivise encore en sous-ensembles, dont certains se subdivisent à leur tour, etc. Les matrices lexicales sont ainsi regroupées sur la base de similitudes et ne constituent pas un simple catalogue.

[2] La forme verbale *désagrémente* (actif, indicatif, présent, 3^{e} personne du singulier) est analysable en quatre morphèmes ou morphogrammes (*dés-agré-ment-e*). Son sens en contexte 'provoquer une sensation désagréable' conduit à lui attribuer comme matrice activée pour sa création la conversion à partir du nom *désagrément* et non la préfixation à partir du verbe *agrémenter*. Les mots *père* et *look* sont tous les deux morphologiquement simples et non analysables. Mais du point de vue des matrices, leur situation n'est pas identique : le second est un emprunt (daté de 1977 par le *Petit Robert*) alors que le premier, étant hérité et faisant partie du fonds ancien de la langue, n'est pas créé par une matrice.

Une seconde confusion, liée à la première, tient à l'identification comme emprunt de toute lexie qui présente un élément étranger. Il faut y regarder de plus près. On ne peut en effet emprunter à quelqu'un ou à quelque chose que ce qu'il possède. De ce point de vue, les emprunts sont moins nombreux qu'on pourrait le penser. Inversement, de réels emprunts peuvent passer inaperçus, faute de marques formelles qui permettent de les identifier automatiquement.

Nous examinerons successivement les cas où la matrice externe est inadéquate, puis ceux où elle est possible mais non assurée et enfin ceux où elle est (quasi) indiscutable, ce qui nous conduira à faire des propositions dans les outils d'analyse.

1.1. Cas d'inadéquation de la matrice externe

1.1.1. Les créations françaises sur des bases étrangères empruntées

Les cas les plus évidents de l'inadéquation de la matrice externe sont les créations françaises sur des bases étrangères empruntées, puisque les mots en question n'existent pas et ne peuvent pas exister dans la langue d'où les bases sont issues ou par laquelle elles ont transité. Plusieurs types de cas se présentent[3] :

i) conversion d'emprunts nominaux en verbes[4] : *tai chi chuer*, *talibaner*…

ii) adjectivation par suffixation sur base nominale : *cartoonesque*, *clippeux*, *showbizien*…

iii) suffixation, avec le suffixe verbal *-is-* : *bestselleriser*, °*bigbrotheriser*, verbe possible non attesté (< *big brother*), base de la nominalisation *bigbrotherisation*…

iv) préfixation : *déstresser*…

v) dérivation inverse : *caster* < *casting*…

1.1.2. Les « faux emprunts »

On appelle *faux emprunts* les créations françaises mettant en œuvre des formants d'origine étrangère et se conformant le plus souvent aux principes de la langue étrangère d'où sont issus ces formants. Aucun doute n'est permis quand ces lexies n'existent pas dans la langue étrangère d'où elles sont censées venir. Quand elles existent, des recréations indépendantes ne sont pas à exclure. Ces faux emprunts se répartissent essentiellement en deux ensembles : les composés savants[5] et les faux anglicismes. Remarquons qu'on peut puiser dans un stock de formants étrangers intégrés (*anthrop*, *mane*, *phile*) ou en emprunter de nouveaux (*psoph-iques*)[6].

1.1.2.1. Les composés savants

La vérification dans le dictionnaire grec-français *Bailly* atteste que *anatopisme*, *athéologie*, *chronophage*, *gérontophobie*, *logomaturge*, et *onirogène* n'existent pas en grec ancien : il s'agit bien de créations postérieures. Des commentaires ne laissent parfois aucun doute à ce sujet. *Anatopisme* est

---

[3] Les données proviennent de diverses sources orales (bulletins d'informations, conversations…) ou écrites (essentiellement la presse, mais aussi des publicités, etc.) récentes pour la plupart. En l'absence d'exemples récents, de plus anciens sont utilisés.

[4] Comme le changement réside seulement dans les marques flexionnelles, sans ajout ni suppression de suffixes dérivationnels, il s'agit de conversion et non de dérivation, contrairement aux analyses traditionnelles. Sont exclus de cet ensemble et considérés comme emprunts des verbes comme *blacklister, bruncher*, *canceler*, *forwarder*, *looker*, *printer*… qui existent comme verbes dans la langue d'origine et qui adoptent seulement la flexion verbale française.

[5] Même si ces formants n'ont pas d'emplois libres, leur utilisation pour créer de nouvelles unités est plus proche de la composition que de la dérivation. Si *logographe* est un composé du grec ancien, la combinaison des deux mêmes éléments dans l'ordre inverse, *graphologue*, est une création moderne (fin XIX^{e} siècle) qui ressortit à un type particulier de composition.

[6] C'est ainsi qu'au début du XVIII^{e} siècle a été intégré le formant *néo* pour créer *néologique* et les autres mots de cette famille. C'est à partir d'eux que ce nouveau formant s'est largement répandu comme l'attestent les dictionnaires qui en comportent de plus en plus.

particulièrement intéressant puisqu'il correspond en fait à trois créations homonymiques, sans filiation directe des sens. C'est d'abord un terme de psychiatrie dénommant les troubles psychiques des personnes déracinées. C'est ensuite, dans la définition donnée sur une affiche publicitaire d'une région touristique, l'agréable sensation de dépaysement ressentie par les vacanciers qui y séjournent. C'est enfin un procédé d'écriture de science-fiction, parallèle à l'anachronisme, consistant à faire apparaître des réalités hors de leur cadre géographique propre (des dromadaires au pôle nord). Remarquons encore que, si ces mots ne sont pas grecs, rien n'empêche que certains aient été créés dans d'autres langues, auxquelles le français les aurait empruntés. Il faut distinguer la langue d'origine et la langue à laquelle on emprunte : certaines lexies passent ainsi par plusieurs langues successivement. Une solution sera proposée en conclusion.

1.1.2.2. De faux anglicismes récents

Si les anciens *tennisman* ou *camping car* sont assurément des créations françaises, on peut s'interroger sur l'origine exacte de *footbusiness*, *kidstore*, *lolitrash*, *peopolisation*, *trashoïd*. Comme *business*, *people*, *store*, *trash*, *tabloïd*[7] sont fréquemment utilisés dans des énoncés français, rien n'interdit de penser que les mots en question ont été créés en français, qu'ils soient ou non attestés en anglais. Cette hypothèse reçoit le renfort de cas où le contexte révèle la création à coup sûr française comme *kitchen music* (une artiste française qualifie ainsi son type de création) et *big sister* (féminisation de *big brother* pour qualifier la communication dans une lettre d'humeur d'un lecteur stigmatisant la « big brotherisation » de la société). Des recherches plus approfondies sont nécessaires pour trancher entre emprunt et création française, d'une manière plausible mais jamais complètement assurée si le terme existe dans la langue potentiellement source.

1.1.2.3. Les hybrides

Les hybrides mettant en jeu un élément français et un élément étranger sont a priori de facture française. Quand les deux éléments sont d'origine étrangère, seule une enquête permet de dire, au cas par cas, si l'unité est plutôt de création française ou plutôt empruntée.

Sont donc passibles d'une matrice interne des lexies hybrides avec un élément français comme *e-connaissance*, *flashpasteurisé*, *khmers verts*, *one meuf show*, *papyboom*, *pré-teenage*, *tabacophobe*, *ultralooké*…

La réponse est moins évidente pour *e-design*, *télécoaching*, *tsunamigène*… dont aucun élément n'est d'origine française. Des enquêtes d'attestation, de datation et de circulation de ces lexies devront être faites pour pouvoir apporter une réponse plausible. Nul doute cependant qu'il doit y avoir beaucoup d'emprunts.

1.2. Cas de concurrence des deux analyses, interne et externe

1.2.1. Des cas non décidables

Quand les mêmes résultats sont engendrés par une matrice interne ou par la matrice externe et qu'il n'y a pas de raison de privilégier a priori l'une des deux solutions, on ne peut pas décider à la seule vue de la morphologie du néologisme. Les dérivés *économisation* (« le sport devenant un marché » et non le « fait de faire des économies ») et *végétalisation* (des terrasses) sont-ils des créations françaises ou des emprunts ? Comme il s'agit de mots récents, le recours à des dictionnaires ne suffit pas toujours. Des recherches sur internet sont plus probantes, mais elles sont dévoreuses de temps sans toujours fournir une réponse complètement assurée. *Ciste*, le « trésor » à découvrir dans une sorte d'énorme jeu de piste, est-il un emprunt ou un néologisme sémantique à partir de *ciste* « corbeille » ? Il s'est implanté en 2002 et a pour origine lointaine le mot grec *kiste*. Mais par où a-t-il transité ?

Le problème se pose d'une manière à peine différente avec certains internationalismes. Nombre d'innovations se répandent rapidement et mondialement, avec les termes qui les dénomment,

[7] Il est significatif qu'un seul, le quatrième, soit inconnu de mon (vieux) correcteur orthographique.

sans qu'il soit possible d'étudier leur trajectoire : la diffusion est tous azimuts et quasi immédiate. Ces dénominations peuvent d'autant moins être analysées nécessairement comme emprunts à une langue particulière qu'elles font appel à des formants savants, latins ou grecs. Néanmoins, l'avancée technologique des USA fait que ces mots sont souvent d'origine anglo-américaine. C'est peut-être le cas de la technique et du terme *dermabrasion* et aussi de *émoticône*. Faut-il dès lors les considérer comme des emprunts ? La question reste encore ouverte. Mais la distinction entre le référent et la dénomination qui apparaîtra clairement avec les calques, traductions et autres équivalences, invite à privilégier une solution interne, avec juste un aménagement.

1.2.2. Des cas décidables

1.2.2.1. Les séries construites sur des moules d'origine externe

Les séries *serial menteur*, *serial dragueur*…, *sanitairement correct*, *scientifiquement correct*…, *homotéléréalitus*… sont construites par analogie sur des modèles devenus des moules : *serial* radical monosyllabique +*er*/*eur*, adjectif +*ment correct*, *homo* radical +*us*. Parfois c'est le schéma qui est importé comme dans substantif ou adjectif + *attitude* : *bus attitude, négative attitude*…

La question est double : qu'est-ce qui est néologique, le moule ou les résultats ? Et une fois que ces structures d'origine étrangère sont devenues des moules productifs en français, ces derniers sont-ils encore des emprunts ou sont-ils intégrés au système ? Il vaut mieux considérer toutes les occurrences citées comme des créations françaises et non des emprunts : les formules latines de base (*homo* X*us*) étaient déjà de création moderne, et la plupart des autres unités n'existent pas dans la langue qui fournit le modèle. Il semble que c'est la matrice du détournement[8] (plutôt que la composition) qui est la plus pertinente. Chaque application à une nouvelle base produit un néologisme par détournement. Ce n'est effectivement pas le moule qui est néologique (seule une lexie peut l'être et le moule n'est pas une lexie), mais ce sont les unités qu'il génère.

1.2.2.2. Les calques, traductions et autres équivalences

Gratte-ciel est traditionnellement considéré comme un emprunt par calque morphologique de *sky scraper*. Cependant, contrairement à la tradition, l'analyse par la matrice interne de la composition paraît préférable pour des raisons linguistiques. Il faut en effet établir une différence entre l'emprunt véritable qu'aurait été l'introduction de *sky scraper* dans l'usage français et la création française *gratte-ciel*. D'autres solutions étaient par ailleurs possibles. On a créé un sigle, *IGH* pour *Immeuble Grande Hauteur* (avec un statut administratif particulier, mais absent des dictionnaires courants) ou un néologisme sémantique comme *tour*. On aurait pu aussi créer un équivalent par traduction comme **gratteur*. Une solution de remplacement à un anglicisme s'est imposée : il serait dommage, et surtout erroné, de ne pas en tenir compte en l'étiquetant purement et simplement emprunt.

Ainsi *action collective en justice*, *flexicurité*, *jeu en réseau* (*LAN party*), *plainte de groupe*, *télétravail* (*kiworking*) et *télétravailleur* (*kiworker*) sont des créations françaises dénommant des réalités d'origine étrangère. Cette influence étrangère ne doit pas être niée, mais elle s'exerce sur le référent, pas sur la forme du mot[9]. Aussi proposons-nous la création d'un champ spécial dans la base de données pour noter ces influences. Cette solution résout un réel problème de notre grille d'analyse initiale et évite de scinder l'ensemble des matrices en deux comme l'avait fait la collègue de Paris 7.

---

[8] Cette matrice, qui ne figure pas dans la grille de Tournier (1985), forme à elle seule l'ensemble des néologismes sémantico-pragmatiques. Elle a été ajoutée pour rendre compte des innovations affectant des séquences lexicales figées. Ces unités ne sont le plus souvent interprétables qu'en référence aux séquences figées qu'elles détournent, comme des palimpsestes. Ainsi la séquence *splendeurs et misères de la science économique* ravale-t-elle celle-ci, par détournement du titre du roman de Balzac, au rang, réprouvé, des courtisanes dont l'éclat factice est peu durable.

[9] La structure de la création française ne correspond pas nécessairement à celle du mot d'origine. On relève par exemple l'inversion de l'ordre des éléments dans les composés, le recours à des synapsies, etc.

1.3. Les emprunts incontestables

1.3.1. Des anomalies révélatrices d'emprunts externes

Les candidats néologismes qui présentent des anomalies par rapport au système français sont a priori des emprunts. Comme la deuxième partie de cet article aborde ce repérage, contentons-nous de signaler ici les cas de discordance entre la graphie et la prononciation : *girly, swing state, nipple gate* et les homonymies dues à l'emprunt sémantique : *supporter* 'être partisan / soutien de' sans rapport avec 'endurer'.

1.3.2. Des informations contextuelles

Parfois, le mot ou la chose sont explicitement indiqués par le locuteur comme venant d'autres pays parlant d'autres langues. Ainsi *skidou*, *géocacheur* sont des francisations, essentiellement graphiques, de *skidoo* et *geocacher*, dénommant respectivement un mode de locomotion sur les grandes étendues neigeuses du pays des Inuits, et un humain qui s'adonne au *géocaching*, à la recherche de « trésors » sans valeur marchande cachés dans la nature, jeu importé des USA. La matrice de l'emprunt s'impose dans ces cas. Elle s'impose à plus forte raison quand les commentaires de l'émetteur ou du transmetteur, plus rarement du récepteur, s'appliquent à des mots dont la forme révèle une origine étrangère comme *Dating market*, *free ride* et *free rider*, *one-to-one*, *ostalgie*, *no show*, *split screen*…

L'usage des marques typographiques que sont les guillemets et les italiques constitue un indice intéressant mais à manier avec précaution car elles ont bien d'autres usages. C'est une cause importante des divergences d'incorporation dans l'enquête évoquée précédemment : à côté de néologismes par emprunt (« *best-of-ing* », « *big five* », « *call center* », « *care* », « *chart* », etc.), il y a de nombreux items en italiques et / ou entre guillemets qui ne sont pas néologiques ou qui ne sont pas des emprunts.

1.3.3. Un cas paradoxal : les emprunts internes

L'introduction de lexies appartenant à des variantes du français, régional ou non hexagonal, est traditionnellement considérée comme emprunt. C'est le cas de *cagole*, *magasinage*, *pourriel*, *vivoir*. Cela paraît séduisant mais, comme le principe qui nous guide est celui de la conformité au système de la langue et non celui de l'origine géographique, et qu'il n'y a aucune raison d'adopter une attitude nombriliste voire impérialiste privilégiant le français hexagonal, nous préférons considérer que ces cas relèvent également des matrices internes et indiquer « régionalisme » dans la case réservée aux influences étrangères ou externes.

On traitera d'une manière analogue les « emprunts » à des états de langue passés que nous avons proposé, sans grand succès encore, de nommer « paléologismes »[10], comme *épinglette* pour évincer *pin's*.

1.4. Une solution pratique en guise de conclusion

Il vaut mieux considérer une lexie comme de facture française chaque fois que c'est possible et ne la considérer comme emprunt que lorsque aucune analyse interne (morphologique ou sémantique) n'est possible. Cette décision a pour conséquence la réduction de la part de l'emprunt par rapport à des solutions internes, non pas par chauvinisme linguistique, mais parce que cela semble correspondre à des réalités linguistiques de plusieurs types.

Les calques, les traductions, les équivalents sont des tentatives, spontanées ou planifiées, de trouver des solutions conformes au système de la langue. Comment pourrait-on faire la différence entre ces procédés de remplacement et l'emprunt proprement dit si tout était indistinctement mêlé sous l'étiquette emprunt ? On amalgamerait des unités lexicales de facture hétérogène.

---

[10] Pour les différencier des archaïsmes qui existent sans solution de continuité et qui sont chargés de connotations. Voir Sablayrolles 2000, p. 191 sq.

Le sentiment linguistique du locuteur français lambda, non linguiste ou historien de la langue, est également à prendre en compte. La *souris* informatique est clairement une métaphore à partir de l'animal. En fait la dénomination métaphorique est un emprunt, mais, à part les spécialistes, personne ne le sait, et surtout on peut s'en passer. L'explication fonctionne bien en français. Est-ce à dire que l'emprunt sémantique n'existe pas ? Non. Il y a des cas où le nouveau sens ne peut pas être dérivé du ou des acceptions françaises antérieures : comme l'ancien *réaliser* 'comprendre' ou le plus récent *supporter* 'soutenir, encourager'.

Il semble néanmoins légitime de noter que de nouvelles acceptions ou des nouveautés formelles, notamment issues des matrices internes morphosémantiques (dérivation et composition essentiellement), se font sur le modèle ou sous l'influence d'une langue étrangère. Aussi décidons-nous, dans la base d'analyse des données[11], de maintenir la matrice externe qu'est l'emprunt au sein d'un champ unique des matrices, mais de créer un autre champ indiquant une influence externe dans l'apparition d'un néologisme dû à une matrice interne. Les composés savants empruntés à des langues vivantes sont passibles du même traitement : dans la mesure où rien dans leur formation n'indique cet emprunt et qu'ils auraient pu tout aussi bien être créés en français, il y a intérêt à les considérer comme de facture française, créés par la matrice 'composition savante', mais de noter que cette innovation lexicale s'effectue sous l'influence de telle ou telle langue vivante.

2. La reconnaissance des emprunts en TAL

2.1 Utilité de la reconnaissance des emprunts en TAL

Si la nécessité de définir l'emprunt comme un processus se fait sentir pour l'étude de la néologie, il semble tout aussi important de s'interroger, parallèlement, sur le bien-fondé de la reconnaissance des unités lexicales en tant que formes ayant été empruntées. Et, s'il ne s'avère pas moins utile d'étiqueter comme telles ces unités lexicales, se pose alors la question de savoir quels moyens mettre en œuvre pour les reconnaître.

Pour essayer de répondre à ces deux questions, la seconde étant contrainte, nous le verrons, par la première, nous avons choisi d'aborder la réflexion sur cette notion de formes empruntées à travers le rôle qu'elle peut remplir dans les applications relevant du traitement automatique des langues (TAL), et plus particulièrement du traitement automatique de corpus écrits.

Cette analyse nécessite quelques précisions méthodologiques préalables. Le TAL étant un champ d'application encore en friche, il est nécessaire d'avancer avec lenteur et prudence et surtout de ne jamais perdre de vue des contraintes scientifiques fortes, nous semble-t-il, telles que l'exigence de :

- préciser l'axe retenu (ici synchronique),
- définir le champ d'application,
- définir les finalités de l'analyse envisagée,
- coller au plus près d'une description linguistique qui doit rester aussi neutre que possible par rapport à l'application, pour en assurer éventuellement le transfert dans une application différente ou plus générale.

2.1.1. Niveau d'analyse

Ce travail sur les emprunts linguistiques repérables en français a été initialement mené dans le cadre du CERTAL[12], groupe de recherche dont l'une des caractéristiques était de mener des travaux parallèles sur des corpus écrits de langues différentes (allemand, arabe, français, tchèque, italien…). Il se situe dans une tradition qui tente de réduire le recours à des dictionnaires et d'utiliser des connaissances s'appuyant sur les régularités morphophonologiques, morphosyntaxiques et

[11] Elle comporte actuellement un peu moins d'une vingtaine de champs, certains obligatoires et à remplir dès la création de la fiche (entrée, description grammaticale, contexte, source de l'énoncé, etc.), d'autres, sont facultatifs ou peuvent être remplis ultérieurement (indication du domaine, définition, analyse morphologique, etc.).

[12] CERTAL, Centre d'études et de recherche en traitement automatique des langues (Paris, INALCO, 1988 -2006.

morphosémantiques de la langue traitée. La justification de ce point de vue est à la fois linguistique et méthodologique. On sait d'une part qu'un dictionnaire ne peut jamais être exhaustif et que d'autre part les phénomènes de créativité lexicale reposent essentiellement sur des fonctionnements morphologiques fondamentaux. Un système dont la compétence lexicale s'appuie sur la connaissance des mécanismes de la langue plutôt que sur un dictionnaire constitué comme un inventaire de formes est capable de faire des calculs sur les unités lexicales absentes des lexiques qu'il utilise.

Nous nous plaçons donc ici au niveau morphologique, nous situant dans le cadre plus large de la reconnaissance des mots inconnus (les « non-attendus » de G. Sabah, 1989 : 151-184), c'est-à-dire non reconnus par le système, soit qu'ils ne figurent pas dans les lexiques (mots hors-lexique), soit qu'ils répondent à des règles de formation non prises en compte par le système, soit enfin qu'ils soient erronés. Le développement du niveau morphologique permet, à un niveau relativement simple et « léger », d'attribuer des traits minimaux à des unités lexicales non reconnues et, ainsi de ne pas bloquer l'analyse des niveaux « supérieurs », syntaxiques et sémantiques notamment. Ce blocage contraindrait en effet à avoir recours à des processus beaucoup plus lourds. Le niveau morphologique est ici pris dans un sens large : il englobe tous les systèmes participant à la construction des formes lexicales apparaissant dans un texte, c'est-à-dire non seulement le système de flexion et de dérivation, mais également tout système de formation des unités lexicales (la composition, les abréviations, etc.). L'étude fine des unités lexicales montre que leur forme et leur structure permettent souvent d'en déterminer les valeurs morphosyntaxiques et morphosémantiques. Il s'agit donc de mettre en évidence tout indice qui permet de guider leur interprétation. L'observation des données a permis d'élaborer des stratégies d'analyse reposant sur l'application de techniques de reconnaissance de formes utilisant des ensembles de règles et un dictionnaire minimal constitué de différentes listes restreintes de formants. Les règles décrivent la structure et les propriétés des unités lexicales et utilisent des listes de morphèmes et/ou de patrons graphémiques.

2.1.2. Contextes où la reconnaissance des emprunts est utile en TAL

Plusieurs applications relevant du TAL peuvent être concernées par le traitement de l'emprunt. L'un des principaux champs d'application concerne la recherche d'informations dans de grandes bases de données structurées ou non structurées telles que celles d'internet, qu'il s'agisse de l'indexation et de l'extraction de mots-clés et de concepts, ou de la fouille de données[13], moins précise que la recherche d'informations, puisqu'elle n'a pas de correspondance à établir avec une requête et se donne comme but de proposer, dans des ensembles non-structurés, un repérage rapide et général de certains types d'informations[14]. L'archivage automatique de textes et l'étiquetage morphosyntaxique de corpus peuvent également utiliser avec profit ce repérage. Ou encore, un système de correction orthographique qui, rencontrant un néologisme, pourra l'identifier comme emprunt, n'aura plus besoin de chercher inutilement à le corriger. Le domaine de la traduction est également un des grands champs d'application, qu'il s'agisse de traduction automatique, de traduction assistée par ordinateur ou de mémoire de traduction. Doit-on ou non traduire un mot emprunté ? Si oui, avec quelles adaptations ? Fondement des mémoires de traduction, l'alignement de corpus, destiné à repérer des syntagmes textuels identiques dans plusieurs langues, trouvera dans la reconnaissance des emprunts de quoi améliorer ses performances. Faisons encore mention de l'enseignement assisté par ordinateur (E.A.O.) qui pourra identifier les unités lexicales périphériques au système de la langue étudiée. Toutes ces applications ont intérêt à repérer des unités lexicales marquées, porteuses d'informations sur elles-mêmes, les traitements à leur appliquer et sur l'environnement textuel où elles apparaissent (nature de la langue, type de texte, etc.).

La reconnaissance des emprunts en TAL répond à la nécessité, pour certains traitements bien précis, d'attribuer à des éléments périphériques au système de la langue d'accueil un certain nombre de traits parfois indispensables pour ne pas bloquer le processus, partiellement ou totalement, et/ou de leur reconnaître des propriétés différentes de celles des unités lexicales autochtones. L'objectif d'une

[13] Extraction de connaissances utiles pour une tâche donnée dans des textes particuliers.

[14] La fouille de données permet par exemple de repérer, par une analyse des mails, les adresses susceptibles d'intéresser les annonceurs de certains types de publicité.

reconnaissance des emprunts peut donc viser un résultat déjà relativement complet ou, au contraire, l'attribution de caractéristiques minimales.

Aux éléments reconnus comme emprunts sont en effet associées des informations sur le traitement à appliquer à l'unité lexicale concernée. Parmi les différences de traitement générées par cet « étiquetage », signalons, entre autres :

- l'application de règles d'analyse et/ou de génération spécifiques,
- l'adaptation des règles de construction du sens à partir des morphèmes aux divers systèmes des langues auxquelles ils sont empruntés,
- la reconnaissance d'une langue différente de celle de l'ensemble du corpus analysé,
- la possibilité de ne pas traduire (choix volontaire ou intraduisibilité),
- la reconnaissance des langues traitées.

L'emprunt est encore une source d'information intéressante, permettant également de dresser une typologie des textes (médicaux, économiques, informatiques…) ou de reconnaître les domaines culturels dans une analyse littéraire par exemple. Certains domaines sont très riches en emprunts, notamment les langues de spécialité ou encore d'autres types de discours, tels que ceux de la mode, de la musique, etc.

2.2. Comment définir l'emprunt en TAL ? Une des voies possibles

La reconnaissance graphémique – rappelons que nous traitons ici de corpus écrits – de ce type d'unités lexicales doit d'autant plus retenir l'intérêt que ces dernières sont très souvent des néologismes et ne se trouvent alors pas (encore) dans la base de données lexicale du système.

L'hétérogénéité des emprunts, la difficulté de s'accorder sur une définition et d'en établir une typologie nous obligent, en TAL, à repérer, dans un premier temps[15], ceux d'entre eux qui sont identifiables comme éléments pas encore ou seulement partiellement intégrés au système. Autrement dit, il s'agit de constituer des patrons de reconnaissance formelle permettant d'identifier des emprunts et de leur attribuer certaines caractéristiques (langue à laquelle ils sont empruntés, catégorie lexicale, grammaticale…). Pour ce faire nous procédons à une analyse des propriétés morphographémiques des unités lexicales.

Conformément au choix exposé précédemment, nous n'avons retenu que les formes empruntées « visibles », c'est-à-dire les unités lexicales reconnaissables formellement. Cette catégorie définit un large spectre dont l'homogénéité est caractérisée par les comportements que vont induire chez le récepteur (système automatique, mais aussi apprenant, traducteur…) les unités lexicales ainsi identifiées. Ce champ englobe aussi bien des unités lexicales empruntées à une autre langue non ou partiellement assimilées dans la langue réceptive que des unités lexicales « savantes », empruntées ou créées à partir d'éléments eux-mêmes empruntés au latin et surtout au grec ancien.

2.2.1. Distinction unités lexicales autochtones / unités lexicales non autochtones

Il est théoriquement possible de partitionner le lexique du français pour distinguer les unités lexicales autochtones des unités lexicales non autochtones. Les contraintes d'une analyse formelle compliquent cette partition.

Les unités lexicales non autochtones constituent elles-mêmes un sous-lexique linguistiquement non homogène, dont les caractéristiques morphographémiques sont déterminées par des degrés différents d'intégration à la langue emprunteuse. Elles ne sont donc plus toujours reconnaissables en tant que telles.

Le degré d'intégration au code de la langue emprunteuse dépend notamment de la période d'emprunt et du domaine de langue. Sur l'axe synchronique il est possible de distinguer trois stades d'évolution dont seulement les deux derniers présentent des caractéristiques morphographémiques différentes du système de la langue qui les emprunte et sont donc pertinentes pour nos préoccupations :

---

[15] Cette première étape exclut, entre autres, les emprunts sémantiques et les propositions de traduction de néologismes étrangers.

(1) intégrée

jardin (all. *Garten*), *vasistas* (all. *was is das ?*), *bouledogue* (angl. *bull dog*), *paquebot* (angl. *packet-boat*), *embarcadère* (esp. *embarcadero*), *agriculture* (latin *agricultura*) *archéologie* (grec *arkhaiologia*) ;

(2) adaptée, moyennant des adaptations phonétiques ou/et graphiques :

bug/bogue, *disquette* (angl. *diskette*), *spoul, spouler* (angl. *spool, spooleur*), *acétamide* (all. *Azetamid*)

ou morphologiques :

debugging/débogage, *packaging/package* ;

(3) conservée sous sa forme initiale :

flirt <ce mot ayant lui-même été adapté quand il a été emprunté par l'anglais au français *fleurette*>, *software, stress, blockhaus, bunker, putsch*...

2.2.2. Un cas particulier et discutable : les emprunts vs *les créations à partir de formants gréco-latins*

L'analyse synchronique d'un corpus de terminologie traduit en français, l'*Atlas d'écologie* de D. Heinrich et M. Hergt (Jacquet-Pfau et Moreaux, 1998) nous a permis de mettre en valeur l'importance des éléments lexicaux empruntés au grec ancien, notamment les « formants » tels que les a définis Henri Cottez, qui ne cessent d'enrichir la langue, notamment scientifique et technique, s'ajoutant aux unités lexicales empruntées directement au grec (*bibliothèque*, *lampadophore, xylophage*). Une reconnaissance des emprunts en TAL a tout intérêt, pour les raisons énoncées ci-dessus, à prendre en compte ces unités lexicales créées à partir d'éléments eux-mêmes empruntés, ces mots « savants » constituant un « *code* particulier à l'intérieur du code général de la langue » (Cottez, 1989 : X).

Contrairement au substrat latin, nombre de ces éléments sont en effet marqués dans le système français et, pour cette raison précisément, sont porteurs d'informations morphologiques, syntaxiques et sémantiques, notamment pour le TAL, mais également pour l'apprenant. Cette spécificité conduit à penser qu'il ne serait pas pertinent de faire disparaître les repères graphémiques mis en place dans un système orthographique, qui n'a pas encore – contrairement à celui d'autres langues romanes, comme l'espagnol par exemple – proposé de simplifier les graphies héritées du grec ancien.

L'exploitation systématique des informations véhiculées par les éléments graphémiques et morpho-sémantiques conservés dans le système de la langue nécessite de dresser une typologie fine des éléments empruntés au grec ancien et de formaliser les principales règles de construction sémantique des unités lexicales complexes qu'ils contribuent à former (ordre centripète). Les formants empruntés au grec constituent, dans la langue emprunteuse, une liste en partie ouverte. Une première liste doit cependant pouvoir être enregistrée dans un dictionnaire. Ce dictionnaire pourra être enrichi de nouveaux formants puisés dans le stock lexical du grec ancien pour former de nouvelles unités lexicales en les combinant avec des formants eux aussi empruntés au grec ou au latin, ou encore à des formants autochtones (ainsi il sera possible d'analyser *tabacologue* ou encore *tricothèque*, mot relevé dans *Le Monde* du 16/12/2005).

2.2.3. Européanismes et internationalismes

Un nombre croissant d'unités lexicales, notamment – mais pas exclusivement – dans les domaines de spécialité, sont communes à plusieurs langues européennes et peuvent même dépasser ces frontières.

L'étude de ces « internationalismes » a été mise en place par Peter Braun et ses collègues, puis reprise par Christian Schmitt (Schmitt, 1991). Le premier corpus de C. Schmitt était constitué par les nouveaux mots introduits dans l'édition de 1988 du *Petit Robert*, qui se révèlent majoritairement appartenir à des langues de spécialité et issus de formes gréco-latines. Ces néologismes se retrouvent, sous une forme adaptée, non seulement dans les autres langues latines examinées, mais aussi en allemand et en anglais. Le développement de ce lexique commun est par ailleurs favorisé par une

politique linguistique visant à unifier les terminologies scientifiques. Les unités lexicales constituant ce lexique sont essentiellement des mots empruntés à l'anglais et au grec ou, plus souvent, des unités créées à partir d'éléments gréco-latins. Elles sont le plus souvent marquées par des adaptations orthographiques qui, dans la plupart des cas, sont formalisables à l'aide de règles de transcription simples et régulières, comme nous l'avons par exemple montré pour le français et l'allemand (*écologie / Ökologie ; radioactivité / Radioaktivität ; insecticide / Insectizid*…) (Jacquet-Pfau et Moreaux, 1988).

2.3. Reconnaissance des emprunts en TAL

2.3.1. L'emprunt comme élément périphérique au système

Dans le cadre défini précédemment, il s'agit de reconnaître des éléments graphiques non conformes (totalement ou partiellement) au code de la langue qui emprunte. Ces unités ne suivent en effet ni le fonctionnement du système graphique ni celui du système morphologique du français. Les particularités de la zone périphérique au système central de la langue d'accueil qu'elles déterminent sont diverses :

• le code graphique est différent (géminations vocaliques, redoublements consonantiques finals, présence de certains graphèmes simples ou complexes, isolés ou cumulés…) :

bazooka, stress, yorkshire, zapping, kronprinz.

C'est précisément cet élément distinctif qui nous permet de reconnaître le sous-ensemble des emprunts que nous pouvons identifier et auquel nous attribuons des propriétés spécifiques.

• l'instabilité orthographique du français (Nina Catach (1980) estimait à environ 10 000 le nombre de graphies « flottantes ») concerne prioritairement ce type d'emprunts. Elle se manifeste d'ailleurs de manière variable d'un dictionnaire à l'autre :

aquavit VAR. *akvavit* (< suédois) (*PR*[16] 2000)
cacahouète ou *cacahuète (*< aztèque) (*PR* 2000)
haschisch ou *haschich* VAR. *hachich* (< arabe) (*PR* 2000)
imprésario ou *impresario* (*PR* 2000), avec phonétique elle aussi variable
offshore, off shore (*PLI*[17] 2004) ou *offshore* (*PR* 2000)
orang-outan VAR. *orang-outang* (< malais) (*PR* 2000)
rösti VAR. *rœsti* (*PR* 2000) ;

• les propriétés morphologiques peuvent présenter des écarts par rapport au système du français :

- La productivité dérivationnelle est, sinon impossible, tout au moins restreinte et ne se développe que dans certains contextes (intégration totale de l'emprunt, usage intensif…) ; elle peut également fonctionner selon des règles spécifiques.
- Les règles flexionnelles appliquées sont parfois celles de la langue à laquelle est emprunté le mot. Il faut toutefois noter que les propositions de réforme de l'orthographe de décembre 1990 préconisent de soumettre les unités lexicales empruntées au système flexionnel du français (*un lied, des lieds* ou *des lieder* ; *un barman, des barmans* ou *des barmen*).
- Les mots empruntés sans adaptation morphologique peuvent conserver la même forme au féminin (*un clown, une clown ; un gourou, une gourou ; un judoka, une judoka ; un baby-sitter, une baby-sitter*) mais peuvent également avoir une forme de féminin empruntée à leur langue d'origine (*un pizzaïolo, une pizzaïolo* ou *une pizzaïola ; un torero, une torera ; un barman, une barmaid …*)

[16] *Petit Robert*.
[17] *Petit Larousse illustré*.

Aussi l'emprunt constitue-t-il en TAL une sphère lexicale périphérique qui, par sa non coïncidence avec le système de la langue qui emprunte et la variété des langues auxquelles on emprunte – même si l'anglais devient majoritaire –, crée des difficultés au moment de processus tels qu'une analyse morphosyntaxique dont le but est d'étiqueter un corpus écrit. L'emprunt formel, s'il ne figure pas dans le lexique, ne permet pas, par exemple, excepté dans quelques cas très restreints, de faire des hypothèses sur la catégorie lexicale, grammaticale, voire sémantique de l'unité, comme permettent de le faire les morphèmes dérivationnels du français. Il est utile de préciser également que les emprunts ne sont pas seuls à constituer des zones périphériques qui, dans un système général, doivent être traitées en parallèle : les noms propres, par exemple, présentent également des caractéristiques particulières. La reconnaissance des emprunts telle que nous la présentons ici est postérieure à celle des noms propres.

2.3.2. Pour une reconnaissance automatique des emprunts : une graphotaxe particulière

Dans le contexte d'une analyse automatique de l'écrit menée au niveau morphologique, une des approches les plus simples est d'identifier les unités lexicales non conformes au code de la langue qui emprunte ou a emprunté parce que, précisément pour cette raison, ces unités lexicales doivent être traitées différemment au moment des processus d'analyse ou de génération.

Cette reconnaissance automatique des unités lexicales périphériques repose sur la reconnaissance de chaînes graphiques, simples chaînes de caractères (symboles, codes informatiques) ou graphèmes (éléments répondant à une propriété linguistique), non intégrées au code du français. À cette reconnaissance de formes a été adjoint un principe fondamental dans tout le système orthographique du français, à savoir le principe de position. L'analyse de notre système fait en effet apparaître que certains graphèmes n'apparaissent jamais à certaines positions dans le mot. Ce sont essentiellement ces deux critères (distribution et position) combinés qui nous ont permis de construire une graphotaxe particulière pour les « mots d'origine étrangère » non (encore) intégrés graphiquement au français.[18]

La graphotaxe est la description ou l'analyse des règles d'organisation des graphèmes à l'intérieur des mots d'une langue donnée. Dans ce cadre, les emprunts seront alors analysés selon des règles de reconnaissance des unités lexicales non intégrées au code morpho-graphique du français. Des patrons de reconnaissance formelle ainsi déterminés permettront de reconnaître :

- des graphèmes marqués,
- des suites de graphèmes marquées,
- des positions marquées de certains graphèmes.

2.3.2.1. Graphèmes marqués

Certains graphèmes simples sont marqués en français :

- la lettre *k*, marquée en latin, car elle n'était utilisée que pour la transcription des mots d'origine grecque. En français cette lettre, d'abord ignorée, a été réintroduite plus tard pour la transcription de mots étrangers ;
- les trois dernières lettres de l'alphabet français, *x, y, z*, importées tardivement par le latin du grec.

Si ces graphes ne sont pas déterminants en soi (l'unité lexicale à laquelle ils appartiennent est souvent parfaitement intégrée), leur repérage peut être utilisé en combinaison avec un autre critère pour améliorer le patron de reconnaissance :

- le graphème *K* (*bunker*) mais surtout C + *K* (*blockhaus*)
- le graphème *W* (*wagon*) mais surtout C + *W* (*software*)

Notons qu'il est souvent nécessaire de mettre en œuvre deux, voire trois filtres pour obtenir une fiabilité maximale.

[18] L'intégration d'un grand nombre de ces unités lexicales se fait moyennant quelques modifications graphiques et/ou phonétiques (*bug* → *bogue*, *supporter* → *supporteur* …).

Nous pouvons ajouter, plus marginalement (notamment dans les domaines de la publicité, de la mode, etc. ou encore dans celui des noms propres) le repérage de combinaisons d'un graphème avec un diacritique n'appartenant pas à l'alphabet de la langue d'accueil :

ä, ö, ã, õ, ñ, š etc. (*cañon* ou *canyon*, *angström* < NP en suédois, *Coca-Cola BlāK* (2006)…).

2.3.2.2. Combinaisons consonantiques et vocaliques marquées

Un certain nombre de combinaisons consonantiques et vocaliques sont périphériques par rapport au code du français. Nous en donnons ci-après quelques exemples :

* les séquences de plus de trois graphes consonnes (*chthonien, schlamm*)
* les voyelles géminées (*kraal, tweed, chiite, booster, coolie,* relooker ; ou dans des emprunts anciens tels que *alcool, zoo…*)

N.B. 1 : La suite vocalique *-uu* marque un emprunt direct au latin (*continuum*).

N.B. 2 : Les seules géminations vocaliques permises par le code français sont celles qui marquent une frontière morphologique, flexionnelle ou dérivationnelle, (*créé, criions, réélire, coorganisateur, simiiforme...*).

* redoublement du graphe *Z* (mots empruntés essentiellement à l'italien : *mozzarella, pouzzolane, paparazzi*…)
* le graphème *H* précédé et suivi par la même voyelle (*mahaleb*)
  excepté aux frontières morphologiques (*antihistaminique*)
* les graphèmes grecs *PH, RH, TH*
* le graphème *J* suivi de *i* (*jigger*)
* le graphème *Q* quand il n'est pas suivi de *-u* (*qat* vs. en finale : *coq, cinq*)
* le digraphe *SH* (*ashkénaze, crash, shochu*)
* le digraphe vocalique *EA* quand il n'est pas précédé par les graphes *G-* ou *-G* ou pas suivi par le graphe *-u* (*skinhead, jean, deleatur*)
* le digraphe *GH* (*ghetto, copyright*)
* le trigraphe *RRH* (*cynorrhodon, logorrhée, sialorrhée*)

etc.

2.3.2.3. Principe de position

Certains graphèmes n'apparaissent jamais à certaines positions dans le mot. Nous donnons ci-dessous quelques-uns des exemples les plus fiables pour l'analyse de graphèmes marqués :

● en position initiale :
  * le graphème *X-* (*xylophone, xénophobie*)
  * le digraphe *PS-* (*psittacidé, psoque, psoralène*)
  * le redoublement consonantique (*ll-* par ex. : de l'esp. : *llanos*, de l'anglais : *lloyd*)

● en position finale :
  * les consonnes géminées (*bluff, edelweiss, hall, jazz, stress*)
  * le graphème *-A* (*coryza* > grec, *pas(s)ionaria* < esp., *polenta* < ital., *chapska* var. *schapska* < polonais, *vodka* < russe)
  * le graphème *-H* excepté dans les interjections telles que *ah, eh...* (*casbah, bobsleigh...*)
    CH en fin d'unité lexicale se rencontre essentiellement dans des mots empruntés à l'anglais (*coach*, *punch, putsch, ranch*)
  * le graphème *-Y* et les digraphes *-EY, -AY* (*bey, derby, trolley, tramway, spray, faraday* < NP)
  * le digraphes *-EA* (*althaea*)
  * le digraphe *-EM* (*harem, requiem...*)

2.3.3. Structures morpho-syllabiques particulières

Bien que ce point n'ait jusqu'à présent pas suffisamment fait l'objet de recherches dans le cadre de nos travaux, nous pouvons déjà utiliser le décompte des syllabes pour reconnaître un certain

nombre d'unités lexicales construites à partir notamment de formants grecs. En effet, plus une unité lexicale comporte de syllabes, plus elle a de chances de correspondre à une unité composée de plusieurs formants, l'ensemble pouvant lui-même être dérivé. Le nombre d'éléments entrant dans le processus de composition en français est limité et permet, souvent associé à d'autres critères, de reconnaître des termes tels que *hexachlorophène*, *photophosphorylation*, *dichlorodiphényltrichloréthane*, etc. comme des formes empruntées.

Si le redoublement syllabique (une ou plusieurs syllabes) n'est pas un critère toujours suffisant pour reconnaître un emprunt (ainsi dans le domaine du langage général il peut s'agir d'un mot enfantin), il peut jouer le rôle de filtre supplémentaire : *aye-aye*, *ylang-ylang*, *youyou* (dans cette série la présence d'un *y* et d'un redoublement consonantique permettent de repérer de manière fiable un mot emprunté).

2.3.4. Morphèmes marqués

* certains suffixes anglais (*-ing, ity, -ful,- less* ...)
* certains formants grecs (*anthro-, baty-, (-)rhiz, philo-, poly-*, *-rrh-*…).

Conclusion

La reconnaissance des emprunts en TAL constitue une partie de l'analyse morphologique comprise comme une analyse qui doit transmettre aux niveaux « supérieurs » des informations même rudimentaires. Mais il n'est pas impensable ailleurs que ce traitement fasse ultérieurement apparaître des propriétés spécifiques à cette sous-classe du lexique.

Cette conception de l'emprunt en TAL diffère donc nettement de celle exposée précédemment, qui se place dans la perspective de la néologie et de l'identification des matrices à l'œuvre dans la production de néologismes à un moment donné de l'histoire de la langue. Ces deux approches sont linguistiquement pertinentes et instructives à condition toutefois de bien les distinguer sous peine, sinon, de tomber dans des confusions dommageables. Cette confusion entre l'aspect dynamique (processuel) et l'aspect statique (résultatif) n'est pas propre à l'emprunt : elle est partagée par les mots *dérivé*, *composé*, etc. Cette confusion est cependant peut-être plus facile à faire dans le cas de l'emprunt du fait que les deux conceptions présentent des similitudes. Dans les deux approches l'emprunt a un statut singulier qui l'oppose globalement au reste du lexique de la langue. La première dichotomie au sein du tableau des matrices, entre les matrices internes et la matrice externe qu'est l'emprunt, n'est pas sans rappeler le recours à des anomalies dans leur reconnaissance en TAL, même si les deux ensembles de données délimités par ces deux approches ne coïncident pas exactement (ils se chevauchent cependant largement). Là où cette différence des deux approches de l'emprunt se manifeste le plus fondamentalement, c'est sans doute dans le mode de leur singularité. Si tout néologisme constitue peu ou prou un événement (certains peuvent néanmoins passer inaperçus), l'emprunt, comme processus d'introduction d'un corps étranger dans un système, l'est encore plus. En revanche, c'est moins d'événement dont il est question dans l'identification en TAL, mais de repérage de faits saillants. Et tous les néologismes-événements, lors de leur introduction, ne présentent pas nécessairement de saillances, phonétiques ou morphographémiques. La distinction des deux approches est donc d'autant plus nécessaire qu'un certain nombre de données pourrait les faire confondre et conduire à conclure à une superposition totale des deux ensembles que ces approches déterminent.

Références bibliographiques

ANIS Jacques, [avec la collaboration de] CHISS Jean-Louis et PUECH Christian, *L'Ecriture : Théories et descriptions*, Bruxelles, De Boeck-Wesmael, 1988 (coll. Prismes, Problématiques ; 10).
BERGMANN, Rolf, « 'Europäismus' und 'Internationalismus'. Zur lexicologischen Terminologie », *Sprachwissenschaft* 20/3, Heidelbreg, Universitätsverlag C. Winter, 1995, pp. 239-277.
CATACH N., en collaboration avec GRUAZ C. et DUPREZ D., *L'orthographe française. Traité théorique et pratique*, Paris, Nathan, 1980.

COTTEZ Henri, *Dictionnaire des structures du vocabulaire savant. Éléments de formation*, Paris, Dictionnaires Le Robert, 1989 (coll. «Les usuels du Robert») (4e édition ; 1ère éd. : 1985).
JACQUET-PFAU Christine, « Du statut de l'emprunt en traitement automatique des langues », *Actes du Colloque international L'innovation lexicale, Université de Limoges, 1-3 février 2001*, Honoré Champion, 2003, pp. 79-97.
JACQUET-PFAU Christine, MOREAUX Marie-Anne, « Motivation et transparence des emprunts gréco-latins en français et en allemand », in A. Clas, S. Mejri et T. Baccouche (dir), *La mémoire des mots, Actes du colloque de Tunis,* Tunis, AUPELF-UREF, 1998, pp. 587-600 (coll. « Universités francophones »).
PRUVOST Jean et SABLAYROLLES Jean-François (2003), *Les néologismes*, Que sais-je ? n°3674, PUF.
SABAH Gérard (1989), *L'intelligence artificielle et le langage*, volume 2 : *Processus et compréhension*, Paris, Hermès.
SABLAYROLLES Jean-François (2000), *La néologie en français contemporain*, Champion.
SCHMITT Christian (1991), « L'Europe et l'évolution des langues de spécialité », *Terminologie et traduction*, n° 2, Commission des communautés européennes, pp. 115-127.
TOURNIER Jean (1985), *Introduction descriptive à la lexicogénétique de l'anglais contemporain*, Champion.