

Géraldine WALTHER

LLF, Université Paris 7 & CNRS, et ALPAGE, INRIA Paris–Rocquencourt & Université Paris 7, France

geraldine.walther@linguist.jussieu.fr

Benoît SAGOT

ALPAGE, INRIA Paris–Rocquencourt & Université Paris 7, France

benoit.sagot@inria.fr

PROBLÈMES D'INTÉGRATION MORPHOLOGIQUE D'EMPRUNTS D'ORIGINE ANGLAISE EN FRANÇAIS

Résumé

Nous proposons une étude morphologique de l'emprunt, en particulier verbal, d'origine anglaise en français. À partir de données extraites d'un corpus volumineux, nous étudions les procédés morphologiques d'intégration des nouvelles unités lexicales (sous leur forme graphémique) et les problèmes qu'ils posent notamment en termes d'instabilité orthographique ou de mécanismes dérivationnels. Cette étude fournit ainsi une première approche théorique du phénomène morphologique de l'emprunt. Elle devra ensuite servir de support théorique à un traitement automatique des emprunts.

Mots clés : Néologie, Emprunt, Morphologie, Lexique, Intégration morphologique et lexicale

1. Introduction

La présence de mots inconnus dans les corpus est un problème important pour le traitement automatique des langues, notamment pour la constitution de lexiques ou l'extraction d'information à partir de textes produits en temps réel, comme les corpus journalistiques. Parmi les mots inconnus, ou plus précisément les *tokens* inconnus, on peut identifier des composants d'entités nommées (nom propres, dates, URL, sigles...), des fautes d'orthographe et des néologismes (Blancafort *et al.*, 2010). Ces derniers peuvent être des créations lexicales productives (p.ex. des dérivés comme *bravitude*), des jeux lexicaux ou langages secrets (verlan, javanais...) ou encore des emprunts.

Dans cet article nous étudions le problème des *emprunts*, que nous définissons comme des *mots hérités d'une langue étrangère mais ayant depuis fait l'objet d'une intégration morphologique*, c'est-à-dire d'un processus permettant leur ajout dans le lexique de la langue cible. Nous centrons ici plus particulièrement notre étude sur le cas des emprunts du français à l'anglais au cours des années 2000 concernant les unités lexicales verbales et les noms d'agent et d'action associés, leurs propriétés morphologiques et leur mode d'intégration dans le lexique. Dans la mesure où notre travail part d'une problématique de traitement automatique de données textuelles, nous étudions ici le lexique dans sa version graphémique. La stabilisation orthographique des emprunts constitue ainsi un enjeu important pour notre étude. Notre ambition est de contribuer à une première approche théorique de la morphologie des emprunts anglais en français. Cette approche théorique pourra ensuite servir à améliorer le traitement par des outils de traitement automatique des emprunts, difficiles en ce que, par leur actualisation permanente, ils constituent des unités lexicales potentiellement inconnues des lexiques disponibles. L'objet de notre travail, qui repose sur des données extraites de corpus, est l'étude de l'intégration de ces nouvelles entrées lexicales.

Nous décrivons dans un premier temps l'origine et le mode d'obtention sur corpus des données analysées (section 2), puis les notions employées pour l'analyse (section 3). En section 4, nous présentons les différents modes d'intégration morphologique des emprunts et les problèmes qu'ils posent, avant de conclure en section 5.

2. Obtention des données

L'extraction automatique d'emprunts à partir de corpus n'est pas une tâche simple. Premièrement, peu de corpus librement accessibles et de taille importante satisfont les critères nécessaires :

- Ils doivent contenir un nombre important d'emprunts récents, adaptés ou non : ceci exclut des corpus comme les corpus journalistiques, dans lesquels les emprunts sont restreints à quelques noms issus de l'actualité comme *vuvuzela*), mais également des corpus comme Wikipedia, dont le langage reste soutenu bien que parfois technique (ce qui apporte des emprunts nominaux de base tels que *firewall* mais peu de verbes ou de noms d'action) ;
- Ils doivent être d'une qualité orthographique suffisante : ceci exclut, du moins dans un premier temps, les corpus de type forums ou blogs.

Verbe candidat	Noms d'agent et d'action ayant induit la construction du candidat
RAPPER	(<i>au, avec, de, du, en, le, les, un, une</i>) <i>rappeur</i> ; (<i>aux, de, des, du, les, pour, vers</i>) <i>rappeurs</i> ; (<i>de, la, une</i>) <i>rappeuse</i> ; (<i>de, des, les</i>) <i>rappeuses</i>
LABELLISER	(<i>de, la, le, une, à</i>) <i>labellisation</i> ; (<i>aux, de, les</i>) <i>labellisations</i>
SCANNER	(<i>de, du, le, par, un</i>) <i>scannage</i> ; (<i>de, du, un</i>) <i>scanneur</i> ; (<i>de, des, les</i>) <i>scanneurs</i>
MAPPER	(<i>de, du, le, un</i>) <i>mappage</i> ; (<i>de, des, les</i>) <i>mappages</i> ; (<i>le</i>) <i>mappeur</i> ; (<i>aux, des</i>) <i>mappeurs</i>
SLAMMER	(<i>du, le</i>) <i>slammeur</i> ; (<i>de, les</i>) <i>slammeurs</i> ; (<i>les</i>) <i>slammeuses</i>
JOGGER	(<i>de, du, le, un</i>) <i>joggeur</i> ; (<i>des, les</i>) <i>joggeurs</i> ; (<i>la</i>) <i>joggeuse</i>
SPAMMER	(<i>le, un</i>) <i>spammeur</i> ; (<i>aux, de, des, les</i>) <i>spammeurs</i>

Tableau 1. Exemples de candidats lexèmes emprunts obtenus à partir de noms d'action et d'agent inconnus

Nous avons porté notre choix sur la version du corpus Wikipedia qui inclut les discussions entre rédacteurs à propos des articles¹. En effet, ces discussions, qui manifestent une créativité lexicale plus riche, complètent ainsi utilement la Wikipedia proprement dite. Nous avons converti ce corpus, originellement au format Wikipedia, en un corpus au format texte contenant 441 millions de tokens dont 4,7 millions de tokens distincts.

Deuxièmement, l'identification automatique d'emprunts n'est pas aisée. Pour extraire des néologismes empruntés au cours des années 2000, nous avons tout d'abord établi une liste de tokens considérés comme connus, en fusionnant la liste de toutes les formes fléchies du lexique Lefff (Sagot, 2010) et la liste des tokens faisant partie de livres publiés pendant les années 1990 et numérisés par Google². Nous avons dans un premier temps extrait des noms d'action et des noms d'agent faciles à reconnaître : nous avons cherché toutes les occurrences de motifs composés d'un token susceptible d'être un déterminant³ ou une préposition suivi d'un token inconnu se terminant en *-eur(s)*, *-euse(s)*, *-trice(s)*, *-age(s)*, *-isation(s)*, *-ification(s)*. Nous avons alors remplacé les suffixes précédents par *-er*, *-iser* ou *-ifier* suivant les cas, en ajoutant quelques variantes possibles (*-ateur/-ateuse/-atrice* > *-er* en plus de *-ater*) et en appliquant certaines règles grapho-phonologiques standard (par exemple, *-cage* > *-quer*). Ceci nous a permis de construire 5 418 néologismes verbaux candidats, auxquels nous avons associé le cumul des nombres d'occurrences des motifs impliquant des noms ayant conduit à leur construction. Ainsi, selon cette métrique, le néologisme verbal candidat le mieux classé est RAPPER, avec un score de 2 363, créé à partir des quatre tokens *rappeur(s)* et *rappeuse(s)*.

¹ <http://dumps.wikimedia.org/frwiki/latest/frwiki-latest-pages-meta-current.xml.bz2>

² <http://ngrams.googlelabs.com/datasets>

³ Nous n'avons pas pris en compte les déterminants complexes, en raison de leur fréquence moindre et de la plus grande difficulté de leur identification.

Nous avons annoté manuellement les 789 candidats dont ce score est de 7 ou plus, en identifiant ceux qui relèveraient effectivement de l'emprunt. Ces derniers sont au nombre de 97, plus 3 calques. À ce stade, ces 97 verbes sont des candidats dont l'attestation en corpus n'a pas encore été vérifiée. Un sous-ensemble en est montré au tableau 1.

Par ailleurs, nous avons cherché, dans ce même corpus, des unités lexicales complexes de type verbe support + nom prédicatif emprunté. Pour cela, nous avons cherché les occurrences du motif *Vsup Det Npred*, où *Vsup* est l'une des formes fléchies de l'un des 37 verbes supports possibles que nous avons retenus (*être exclu*)⁴, *Det* est un token susceptible d'être un déterminant et *Npred* est un token inconnu au sens défini ci-dessus. Une fois le déterminant éliminé, les 31 072 occurrences de ce motif donnent 18 090 séquences distinctes pour 15 129 noms prédicatifs différents possibles. Après élimination des hapax, les 3 655 séquences distinctes restantes ont été classées manuellement afin d'identifier celles qui sont effectivement des constructions à verbe support et dont le nom prédicatif est un emprunt. Nous avons ainsi identifié 167 séquences distinctes mettant en jeu 89 noms prédicatifs, au singulier ou au pluriel, et stocké leurs nombres d'occurrences respectifs (cf. tableau 2).

<i>Vsup</i>	<i>Npred</i>	Occ	Verbe simple	<i>Vsup</i>	<i>Npred</i>	Occ	Verbe simple
avoir	spin	41	SPINER	prendre	gameplay	23	SPINER
faire	remix	41	REMIXER	posséder	spin	20	REMIXER
avoir	gameplay	40	GAMEPLAYER	effectuer	heel	19	GAMEPLAYER
avoir	blog	28	BLOGUER	avoir	flashbacks	14	BLOGUER
avoir	remix	25	REMIXER	faire	buzz	13	REMIXER
faire	heel	25	HEELER	lancer	blog	13	HEELER

Tableau 2. Exemples de lexèmes emprunts candidats obtenus à partir de noms d'action et d'agent inconnus

Verbe	« Occ »	Verbe	« Occ »	Verbe	« Occ »
REMIXER	1634	REMASTERISER	192	MASTERISER	120
REMASTERISER	599	CATCHER	187	TAGUER	85
HACKER	473	SKATER	137	RAPPER	85
SAMPLER	356	MAPPER	125	SCORER	76

Tableau 3. Emprunts verbaux simples les plus fréquents extraits du corpus

Nous avons alors complété notre première liste de candidats verbes simples empruntés en la complétant comme suit : à partir de la forme singulier des 89 noms prédicatifs ci-dessus, nous avons construit un verbe candidat par ajout de *-er* (ou simplement de *-r* lorsque le nom se termine par *-e*). Nous avons ainsi construit 98 lemmes verbaux (cf. tableau 2), dont 7 communs avec les 97 verbes construits plus haut (SPINER, SNOWBOARDER, BLOGUER, STRIPTEASER, REMIXER, SLAMER, DEALER). L'union de ces deux listes, qui contient donc 188 verbes, a alors été fléchié selon la classe flexionnelle des verbes standard du premier groupe, qui est celle de la quasi-totalité des néologismes (Bonami et Boyé, 2003). Parmi les formes fléchies obtenues, celles correspondant à des tokens connus ont été éliminées, ainsi que 95 formes qui auraient conduit à des erreurs (exemples : *cora*, nom d'une marque, *game*, emprunt nominal, et d'autres). Nous avons extrait du corpus le nombre d'occurrences des 434 formes restantes, et avons alors calculé un nombre d'occurrences approché (noté « Occ ») pour les 188 verbes, en additionnant les contributions de celles de leurs formes qui ont été

⁴ Il s'agit des verbes suivants : *accorder, allonger, allouer, apporter, asséner, attribuer, avoir, bourrer, commettre, crépiter, distribuer, donner, décocher, déployer, effectuer, émettre, engager, éprouver, faire, filer, flanquer, infliger, jeter, lancer, livrer, mener, mettre, nourrir, partager, perpétrer, porter, posséder, pousser, prendre, recevoir, ressentir, subir.*

conservées. Seuls 106 des candidats verbes sont alors considérés comme attestés, les 12 plus fréquents étant indiqués au tableau 3. Ce sont ces candidats verbes qui ont fourni les données dont l'intégration morphologique est étudiée en 4.

3. Définition des notions sous-jacentes à l'analyse

Notre étude se situe dans une *approche lexématique* de la morphologie (Matthews, 1974) : le *lexème* est considéré comme une unité lexicale abstraite définie par sa phonologie et/ou sa graphie, son sens et sa catégorie morphosyntaxique⁵. Notre approche s'inscrit par ailleurs dans l'approche *Mot et Paradigme (Word and Paradigm)* dans laquelle la notion de *paradigme* défendue par (Hockett, 1954) et (Robins, 1959) constitue une notion centrale. Ainsi, nous désignons par *paradigme* l'ensemble des formes d'un même lexème. Notre vision du *paradigme* s'appuie par ailleurs sur les notions de *radicaux* et d'*exposants* (Matthews, 1974)⁶ ; les formes concrètes sont construites à partir des lexèmes abstraits à l'aide de *règles de réalisations* définies dans des *modèles inférentiels réalisationnels* comme PFM (*Paradigm Function Morphology* (Stump, 2001) ou *Network Morphology* (Corbett & Fraser, 1993)), produisant ainsi les formes effectivement observables dans la langue.

En morphologie lexématique, la création d'un nouveau lexème ne peut se faire que par des *règles de création de lexèmes (LCR)* qui spécifient, pour chaque nouveau lexème créé, sa phonologie, son sens et sa catégorie morphosyntaxique. Les LCR peuvent être de plusieurs types, les deux types majeurs étant les règles de composition et les règles de dérivation. Les premières mettent en jeu la combinaison de deux lexèmes existants, la seconde modifie un lexème unique (Fradin, 2003 ; Tribout, 2010). Parmi ces dernières, celles qui nous intéressent ici sont les règles de *dérivation affixale* et les règles de *conversion*.

Les LCR de **dérivation affixale** du français peuvent être définies comme faisant intervenir une base et un *affixe dérivationnel*. La combinaison entre base et affixe est soumise à des contraintes sémantiques du lexème base⁷. Nous définissons la dérivation affixale (en français) par

1. un ensemble de conditions sur l'input de la règle indiquant notamment la (sous-)catégorie morphosyntaxique du lexème, sa structure argumentale et des contraintes sémantiques ;
2. les traits sémantiques réalisés par la règle (cf. nom d'agent pour les noms dérivés en *-eur* comme *danseur*) — ces traits sémantiques constituent ainsi des conditions sur l'output ;
3. les propriétés morphosyntaxiques associées à l'output — déterminant notamment la catégorie morphosyntaxique et la flexion du lexème dérivé ;
4. la dérivation affixale proprement dite définie par des opérations phonologiques (affixales) appliquées au lexème base — et potentiellement assorties de contraintes phonologiques.

Les lexèmes issus d'un processus de **conversion** sont appelés *converts*. Nous adoptons les critères définis par Tribout (2010), à une exception près⁸ :

- Il existe au moins une forme du paradigme du convert phonologiquement/graphiquement identique à une forme du paradigme du lexème base ;
- la LCR a pour effet un changement de catégorie ou de sous-catégorie ;
- le convert possède toutes les propriétés attendues de sa (sous-)catégorie. Il ne peut ainsi y avoir de défektivité comme pour *le vrai/*les vrais*.

⁵ Nous considérons que des propriétés syntaxiques (structure argumentale, etc.) sont spécifiées dans le lexique.

⁶ Le radical est la partie d'une forme après suppression des marques flexionnelles (à savoir des exposants).

⁷ Fradin et Kerleroux (2003) notent que le verbe VOLER 'dérober' peut être dérivé en son nom d'action VOLEUR, mais pas VOLER dans le sens de 'se déplacer dans l'air' malgré des propriétés morphologiques identiques.

⁸ Tribout (2010) considère que la conversion change nécessairement la catégorie morphosyntaxique. Dans la mesure où nous considérons les propriétés morphosyntaxiques comme inhérentes à la définition d'un lexème, nous considérons qu'un changement de sous-catégorie (ayant par exemple un effet sur la structure argumentale du convert) est un critère suffisant. En ce sens, notre définition rejoint ainsi celle de (Mel'čuk, 1996).

Lors du processus de *lexicalisation* (Corbin, 1992) initié par une LCR, une nouvelle entrée est introduite dans le lexique. Nous définissons ici le *lexique* comme l'ensemble des *lexèmes appartenant à une langue*. Dans cet article nous nous intéressons aux LCR spécifiques à la création de lexèmes-emprunts. Nous définissons les *emprunts* comme des *mots hérités d'une langue étrangère mais ayant depuis fait l'objet d'une intégration morphologique*. Cette intégration morphologique se traduit par l'existence de paradigmes propres à ces nouveaux lexèmes de la langue cible.

4. Processus et degrés d'intégration morphologique

4.1 Stabilisation de l'orthographe des emprunts morphologisés

L'enrichissement du lexique d'une langue par emprunts passe dans un premier temps par une adaptation phonétique et/ou graphémique de mots (formes) issus de la langue source en séquence de signes phonétiques ou graphiques intégrables dans la langue cible. Ces séquences adaptées fonctionnent alors comme des bases pour la morphologie constructionnelle. Elles peuvent constituer l'input de LCR dont l'output fournit une nouvelle entrée lexicale dans la langue cible. La première difficulté dans la stabilisation des nouvelles entrées lexicales réside ainsi dans les processus d'adaptation phonétique et/ou graphémique des formes empruntées à la langue source.

Ainsi, la coexistence de séquences graphiques telles *déboguer*, *debugger*, *debuguer*, *débugguer* et *debugguer* rend difficile la reconnaissance des entrées lexicales nouvelles créées par les processus d'emprunt. La difficulté de stabilisation orthographique apparaît dès lors que les conventions orthographiques de la langue source et de la langue cible diffèrent. Dans le cas du mot anglais *debug*, la création d'un verbe en *-er* est compliqué par les conventions orthographiques du français permettant de transcrire le son [g] : en français <g> devant <e> se prononce [ʒ], des stratégies de restitution du son [g] sont nécessaires. Pour les emprunts, il y a concurrence entre la stratégie de la langue cible de l'insertion d'un <u> entre <g> et <e> (cf. *tangage* → *tanguer*) et celle de la langue source par redoublement de la consonne finale devant voyelle (cf. *debug* + *-ing* → *debugging*).

Forme	Lemme	Occ	Forme	Lemme	Occ	Lemme	Occ
<i>déboguer</i>	DÉBOGUER	44	<i>débogue</i>	DÉBOGUER	2	DÉBOGUER	59
<i>debugger</i>	DÉBUGGER	11	<i>débugué</i>	DÉBUGUER	1	DÉBUGGER	16
<i>déboguée</i>	DÉBOGUER	4	<i>débuguée</i>	DÉBUGUER	1	DÉBUGUER	6
<i>debuguer</i>	DÉBUGUER	4	<i>debugguer</i>	DÉBUGGUER	1	DÉBUGGUER	1
<i>débogué</i>	DÉBOGUER	4	<i>déboguées</i>	DÉBOGUER	1	DEBBUGUER	1
<i>debuggé</i>	DÉBUGGER	3	<i>déboguaient</i>	DÉBOGUER	1		
<i>debuggée</i>	DÉBUGGER	3	<i>déboguant</i>	DÉBOGUER	1		
<i>débogués</i>	DÉBOGUER	2	<i>debuggué</i>	DEBBUGUER	1		

	-gage	-ggage	-guage	-geur	-gueur	-ggeur	-ggueur
<i>debo-</i>	4			1			
<i>debu-</i>	1	6					
<i>débo-</i>	115	2	24	1	109	3	1
<i>débu-</i>	12	8			2	8	1

Tableau 4. Cas d'instabilité orthographique pour l'emprunt construit à partir de *debug*

Les formes appartenant aux paradigmes des cinq infinitifs ci-dessus observables dans les corpus étudiés sont listées avec leur nombre d'occurrences dans le tableau 4, de même que différentes variantes du nom d'action et du nom d'agent correspondants. Sont également indiquées pour les verbes les nombres d'occurrences de chaque entrée lexicale (lemme)

potentielle en additionnant les occurrences de toutes ses formes (calcul différent de celui de la section 2). Le tableau 4 montre que la stabilisation orthographique de l'emprunt du mot anglais *debug* n'est pas achevée. Mais le nombre d'occurrences permet déjà d'identifier des préférences orthographiques émergentes.

D'après nos données, de tels cas de flottements orthographiques dus au redoublement de la consonne finale concernent plus généralement les mots se terminant par une consonne finale unique précédée par une voyelle courte en anglais. Ces mots redoublent leur consonne finale devant tout suffixe commençant par une voyelle. Cette règle n'existe pas de façon native en français, de sorte qu'on observe pour les emprunts des flottements entre l'utilisation de bases d'emprunt à consonne finale redoublée et de bases finales simples, cf. RAP(P)ER (< *rap*), SCAN(N)ER (< *scan*) ou NOVEL(L)ISER (< *novel*).

4.2 Types de LCR à partir de séquences adaptées de mots d'une langue source

Après l'adaptation des mots empruntés, l'intégration morphologique passe par des LCR propres aux emprunts. Sur le modèle de la dérivation et de la conversion définies ci-dessus, nous définissons deux types de LCR applicables aux séquences en langue étrangère adaptées :

1. Les LCR à structure dérivationnelle pour lesquelles une structure en base (*empruntée*) + affixe dérivationnel (propre à la langue cible) est reconnaissable ;
2. Les LCR de type conversion où au moins une forme du paradigme ou un radical du lexème créé est homophone/homographe avec la séquence adaptée de l'input.

LCR à structure dérivationnelle affixale C'est le cas pour les exemples (1) et (2).

- (1) Anglais *remix* → base : *remix* + suffixe *-age* → lexème : REMIXAGE
- (2) Anglais *pixel* → Français : base : *pixel* + suffixe(s) *-isation* → lexème : PIXELISATION

LCR de type conversion C'est le cas pour tous les verbes construits à partir de séquences adaptées de mots anglais se terminant par *-e* comme dans les exemples (3) et (4).

- (3) Anglais *parse* → Français : forme : (*il/elle*) *parse* → lexème : PARSER
- (4) Anglais *impulse* → Français : forme : (*il/elle*) *impulse* → lexème : IMPULSER

Ces procédés peuvent être réguliers, mais parfois des cas particuliers se distinguent.

4.3 Ambiguïtés dans l'identification des bases dérivationnelles

Lors de l'intégration des nouvelles unités lexicales dans la langue cible par des LCR à structure dérivationnelle affixale, la frontière entre base et affixes n'est pas toujours clairement identifiable. Nous avons notamment observé cette difficulté pour les verbes dont la forme source vient d'un verbe anglais se terminant en *-ate* comme *populate*. Comme expliqué dans la section 2, nous avons, pour ces données, produit deux lemmes candidats, doublons provenant de généralisations excessives sur des procédés analogiques observables en français, cf. exemples (5) et (6).

- (5) *acclimateur* et *acclimater*, mais pas **acclimer*
- (6) *exploitation* et *exploiter*, mais pas **exploiter*

Ainsi avons-nous produit **gentrificater* et *gentrifier* à partir de *gentrification* et **populater* et *populer* à partir de *populage* et *populateur*. L'obtention du radical *popul-* plutôt que *populat-* semble s'expliquer par une réanalyse du mot anglais en analogie avec le mot français *population* décomposé en radical *popul-* + suffixe *-ation*. La LCR permettant d'obtenir le lexème POPULER passe donc par une analogie avec un mot similaire déjà présent dans le lexique français.

4.4 Surdifférenciation lexicale

Enfin, l'intégration des emprunts permet de créer des différenciations dans la langue cible qui n'existaient pas dans la langue source. Un tel cas est représenté dans l'exemple (7).

(7) *babysitter* (M/F) → BABYSITTEUR (M/F) → BABYSITTEUR (M) vs. BABYSITTEUSE (F)

Ainsi le mot *babysitter* en anglais désigne une personne qui garde un enfant, qu'elle soit un homme ou une femme. Lors de la création de l'emprunt, une réanalyse des composants morphologiques du mot anglais conduit à identifier, par analogie, une base verbale *babysit-* et le suffixe dérivationnel de type nom d'agent en *-er* qui est rendu par *-eur* en français. Le français possédant par ailleurs une alternance régulière entre les suffixes *-eur* pour les noms d'agent de sexe masculin et *-euse* pour ceux de sexe féminin, nous obtenons à partir d'un lexème unique dans la langue source, deux lexèmes dans la langue cible. Ces deux lexèmes introduisent à partir d'un même mot emprunté une nouvelle différenciation propre à la langue cible que nous pouvons qualifier de *surdifférenciation* par rapport à la langue source⁹.

4.5. Liens de dérivation empruntés

La majorité des emprunts se font mot pour mot. Par ailleurs, la plupart des lexèmes empruntés de l'anglais en français est constituée de noms. Ces noms peuvent alors à leur tour servir de base à la construction d'un verbe par des procédés de dérivation propres à la langue cible (cf. SCAN > SCANNER). Le lien de dérivation entre SCAN et SCANNER se construit ici par une LCR de dérivation interne.

Néanmoins, dans certains cas, le français peut emprunter plusieurs lexèmes qui en anglais étaient formés sur la même base. C'est par exemple le cas du verbe PARSE et du nom d'action PARSING. Les deux lexèmes sont les résultats de LCR d'emprunts séparées. Pourtant, les locuteurs reconstituent dans le lexique d'arrivée les liens de dérivations existant dans la langue source : PARSE et PARSING sont clairement perçus comme issus de la même base dérivationnelle, y compris en français. On peut alors parler d'emprunts de liens de dérivation qui s'ajoutent au cas d'emprunts lexicaux. Ces liens de dérivations empruntés coexistent avec les liens de dérivation internes construits à partir de lexèmes empruntés (CRACKER > CRACKAGE), au point que dans certains cas, il y a concurrence directe entre dérivations internes et liens de dérivation empruntés (cf. MAPPER vs. MAPPING et MAPPING).

5. Conclusions et perspectives

À partir de données extraites de corpus volumineux, nous avons étudié les emprunts verbaux, et leurs noms d'agent ou d'action correspondants, d'origine anglaise en français et étudié les modes d'intégration morphologique de ces entrées dans le lexique du français. Nous avons identifiés un certain nombre de problèmes liés à la stabilisation des mots nouveaux et à leur réinterprétation en termes d'analogie morphologique avec les processus existant en français. Nous avons également montré que le phénomène morphologique de l'emprunt ne se limite pas à l'intégration d'unités lexicales, mais qu'il peut également concerner des cas d'intégration de relation morphologiques comme des liens de dérivation.

Pour l'instant nous avons laissé de côté les calques (cf. exemple (8)) qui constituent une version particulière de l'emprunt. En effet, dans ce cas, l'emprunt ne porte pas sur une séquence phonologique/graphémique, mais est construit par analogie avec des mots existants dans la langue source et dont les constituants ont été identifiés et traduits dans la langue cible. Par ailleurs, l'intégration morphologique n'est pas la seule façon pour une unité lexicale verbale d'intégrer le lexique d'une langue : l'intégration syntaxique le permet également, notamment au moyen de constructions à verbe support (Danlos, 1992). Nous ne sommes pas pour l'instant en mesure de proposer des critères généraux permettant de prédire si un

⁹ D'autres cas de ce phénomène de surdifférenciation sont illustrés par GOLFEUR/GOLFEUSE (< *golfer*), BOXEUR/BOXEUSE (< *boxer*) ou HACKEUR/HACKEUSE (< *hacker*).

emprunt sera intégré plutôt morphologiquement ou syntaxiquement¹⁰, mais quelques tendances peuvent déjà être discernées :

- certains noms prédicatifs empruntés sont phonologiquement difficiles à utiliser comme base pour la création d'un verbe simple : c'est par exemple le cas du verbe anglais *google* 'rechercher sur Google', prononcé [gu:gəl] en anglais et [gugœl] en français, qui donnerait régulièrement le verbe *googler* [gugle], attesté mais rare (du reste, on trouve également *googoliser* et *googliser* sur le Wiktionnaire) ;
- certains noms prédicatifs empruntés sont ressentis comme très similaires, en sens et/ou en phonologie, à des noms prédicatifs du français ; lorsque ces derniers n'ont pas d'équivalent verbal, il en est de même pour l'emprunt ; c'est par exemple le cas de *masterclass*, associé à *classe* : ce dernier n'ayant pas d'équivalent verbal à cause du blocage provenant de l'existence du verbe *classer* ayant un tout autre sens, le verbe *masterclasser* n'est pas attesté ;
- certains verbes simples, notamment s'ils proviennent d'un verbe anglais qui n'est pas lui-même un convert ou un dérivé, ne semblent pas avoir de contrepartie fonctionnant comme un nom prédicatif ; c'est par exemple le cas de *solver* 'résoudre', issu de l'anglais *solve*. Nous avons également l'intention d'approfondir ce travail en étudiant le rapport entre structure argumentale et préférence morphologique vs. syntaxique, cf. exemples (9) et (10).

(9) Intégration syntaxique : *Je vérifie ça et je te fais un mail.*

(10) Intégration morphologique : *Je te maile le document dès qu'il est prêt.*

Remerciements Travail réalisé au sein du projet ANR EDyLex (ANR-09-CORD-008)

Références

- Blancafort San José H., Recourcé G., Couto J., Sagot B., Stern R. & Teyssou D. 2010. Traitement des inconnus : une approche systématique de l'incomplétude lexicale. In *TALN 2010*, Montréal Canada.
- Bonami, O., Boyé, G. 2003. Supplétion et classes flexionnelles dans la conjugaison du français. *Langages*, **152**, 102–126
- Corbett G. G. & Fraser N. 1993. Network Morphology: a DATR account of Russian nominal inflection. *Journal of Linguistics*, **29**, 113–142.
- Corbin D. 1992. Hypothèses sur les frontières de la composition nominale. *Cahiers de Grammaire*, **17**, 25–55.
- Danlos L. 1992. Support verb constructions: linguistic properties, representation, translation. *Journal of French Linguistic Studies*, **2**(1), 1–32.
- Fradin, B. 2003. Nouvelles approches en morphologie. Paris, France : Presses Universitaires de France.
- Fradin B. & Kerleroux F. 2003. Troubles with lexemes. In G. Booij & A. R. Janet de Cesaris, Sergio Scalise, Eds., *Selected papers from MMM3*, Topics in Morphology, p. 177–196, Barcelona, Spain.
- Hockett C. F. 1954. Two models of linguistic descriptions. *Words*, **10**, 210–234.
- Matthews P. H. 1974. Morphology. Cambridge, UK: CUP.
- Mel'čuk I. 1996. Cours de morphologie générale. Troisième partie: Moyens morphologiques. Quatrième partie: Syntactiques morphologiques. Montréal, Canada: Presses de l'Université de Montréal.
- Robins R. H. 1959. In defense of wp. *Transactions of the Philological Society 1959*, p. 116–144.
- Sagot B. 2010. The *Lefff*, a freely available, accurate and large-coverage lexicon for French. In Proceedings of LREC 2010, La Valette, Malte.
- Stump G. T. 2001. Inflectional Morphology. A Theory of Paradigm Structure. Cambridge, UK: CUP.
- Tribout D. 2010. Les conversions de nom à verbe et de verbe à nom en français. PhD thesis, Univ. Paris 7, Paris.

¹⁰ Comparer systématiquement l'intégration morphologique et l'intégration syntaxique nécessiterait une extraction exhaustive des deux types d'emprunts. Ceci est impossible pour les emprunts de type verbes simples en raison de l'homographie de certaines de leurs formes avec des mots préexistants, voire de l'ensemble de leur paradigme (cf. l'emprunt *POSTER* (< angl. *post*) dans son emploi *publier un message sur un forum internet*).