ESTHER BAIWIR
Chargée de recherches
FNRS - Université de Liège
Belgique
ebaiwir@ulg.ac.be

Lexicografía e informática: sí, pero... / Lexicographie et informatique: oui, mais...

1 Introduction

En matière de dictionnaire informatisé du français, le *Trésor de la Langue française*, le grand dictionnaire de la langue du 19e et 20e siècle, publié entre 1971 et 1994, fait sans doute partie des pionniers. En effet, si l'ouvrage est accessible en ligne depuis 2002,[1] la conception de son informatisation date de 1995. Jean-Marie Pierrel, cheville ouvrière de cette informatisation, explique dans l'introduction à la version électronique que «[l]'évolution des techniques informatiques permet [...], à travers l'informatisation d'un dictionnaire, de découvrir des usages nouveaux et des parcours véritablement novateurs qui s'affranchissent des aspects essentiellement séquentiels de la lecture et de la recherche dans les textes imprimés» (TLFi, Introduction, 2004).

Nous avons donc mis cette assertion à l'épreuve d'une question de linguistique historique. Dans ces quelque pages, nous exposerons notre ambition puis nous tenterons de voir en quoi l'outil «TLFi» nous a concrètement permis d'atteindre notre objectif – réunir un corpus pertinent pour traiter une question particulière. Ce parcours nous permettra de souligner tous les dangers d'une utilisation non raisonnée des dictionnaires numérisés et de rappeler l'importance de la dimension historique dans les sciences humaines: ces conversions informatiques sont l'œuvre d'humains, qui ont posé des choix en fonction de critères qu'il importe de connaître.

Notre question consiste en l'examen des traces que la langue arabe a laissées dans le français. Une première objection est la suivante: quel *français* ? Nul besoin, en effet, de s'appesantir sur l'illusion que constitue la perception de la langue comme une, du lexique comme définitif. D'où le choix d'un corpus facilement appréhendable et délimité, mais que nous envisagerons dans sa totalité: le *Trésor de la Langue française*.

Seconde objection: quelles *traces*, mais aussi quelle *langue arabe* ? Devant la difficulté de répondre rapidement à cette question, nous nous devons de poser: toutes les traces, toutes les langues arabes. Concrètement, il s'agissait donc de classer les mots d'origine arabe du TLF(i), en fonction de critères historiques et en s'appuyant sur un corpus restreint. Si nous limitons le corpus à la langue française générale, telle qu'elle est décrite dans le TLF(i), nous n'écartons en revanche aucune époque d'emprunt (médiéval ou plus récent), aucune langue-source (arabe «standard», écrit, dialectal).

[1] V. http://atilf.atilf.fr/tlf.htm.

2 Définition et délimitation du corpus grâce au moteur de recherche du TLFi

Dans l'immense corpus du TLF(i), nous avons donc effectué un premier tri afin de ne conserver «que» les quelque centaines mots passés de l'arabe au français, quels que soient leur origine, leur parcours, leur sens et les éventuelles langues intermédiaires par lesquelles ils ont transité. Nous intégrons donc dans notre corpus tous les mots qui ont existé dans une partie du monde arabophone au moins, et postérieurement dans le français, ou du moins dans l'état de langue que prétend représenter le *Trésor de la langue française*.

Très concrètement, l'option qui semblait la plus pertinente pour recueillir le corpus était l'exploitation des outils «intelligents» présentés sur le site du TLFi, via l'onglet «recherche complexe». Il est ainsi possible de sélectionner l'option «langue empruntée» et d'inscrire «ar.» dans le champs (ou «arabe», les résultats sont les mêmes, signe d'une construction raisonnée du moteur de recherche). Le chiffre obtenu est celui de 216 résultats, ce qui semble faible, au vu des chiffres avancés traditionnellement dans la littérature sur le sujet (entre 508 chez Walter et Baraké et 546 chez El Houssi[2]). En outre, aucun mot commençant par *a-* n'apparaît dans le corpus; pour une langue dont l'article est *al*, on ne peut que s'en étonner.

Nous avons donc tenté une autre approche, plus automatique, en émettant la requête d'isoler toutes les occurrences de «ar.», dans tout le TLFi. Des résultats apparaissent en effet, mais le chiffre donné semble tout aussi extravagant: 883 résultats. Si ce résultat est tout aussi douteux que le précédent, il a le mérite d'avoir employé un filet aux mailles serrées. Nous pouvons donc supposer qu'aucun lexème ne nous a échappé. Ne reste plus qu'à trier cette pêche abondante en la dépouillant systématiquement.

Un premier élément important vient gonfler notre chiffre: la numérotation par occurrence et non par entrée. Par exemple, l'article *abdalas* «compte pour deux», car l'étiquette apparaît à deux reprises dans la rubrique «étymologie».

Ensuite, un tri rapide – mais manuel – permet d'écarter celles ne correspondant pas à «arabe», ainsi que celles ne figurant pas dans le commentaire étymologique de l'article. Pour quelques cas, l'étiquette «ar.» n'apparaît que pour réfuter une étymologie (ex. *bagage*, *bahut*, *charabia*); ces cas ont également été écartés. Dans d'autres cas, l'étymologie n'est pas assurée et l'origine arabe n'est qu'une proposition parmi d'autres (*barbacane* et *tambour* sont issus soit de mots arabes, soit de mots persans, etc.). Signalons encore qu'on n'a pris en compte que les mots simples; lorsqu'un dérivé apparaît, construit sur un mot déjà français, il a été écarté. C'est le cas, par

[2] Nous ne tenons pas compte d'ouvrages a-scientifiques comme celui de Dîb Farâdj, qui dénombre pas moins de 950 entrées.

exemple, de *émirat* (issu de *émir*), qui semble évident. D'autres le sont moins: *fardeau* provenant de *farde* ou *abrine*, issu de *abre* et ce, bien que ce dernier ne constitue pas une entrée du TLF.[3] Ce travail, fastidieux, semble bien être incontournable; en effet, le balisage, effectué secondairement par rapport à l'ouvrage, ne permet pas d'obtenir des résultats fiables.

Restent 460 cas, que l'on peut soumettre à différentes analyses. Pour ce faire, nous avons passé chaque article au crible afin de sélectionner les informations intéressantes pour notre propos. En effet, le premier objectif est de réduire la matière afin de pouvoir la mettre en perspective. Il faut donc transformer les textes suivis que constituent les rubriques «Étymologie et Histoire» du TLF(i) en données balisées et comparables.

Dès cette étape, le travail d'élaboration du corpus ne bénéficie plus que très peu de l'informatisation de l'ouvrage, si ce n'est par sa facilité d'emploi. Il n'est en effet pas possible actuellement d'isoler, au sein d'une première sélection, un sous-groupe présentant telle ou telle particularité.

Il convient donc d'examiner une à une les notices, afin d'assurer le statut des lexèmes. Dans le cadre d'une étude sérieuse des arabismes, il conviendra d'écarter les arabismes indirects, pour ne pas verser dans l'*etimologia remota*. Des mots tels que *guitare* ont en effet abouti en français *via* l'espagnol; l'arabe lui-même l'avait emprunté au grec. Dès lors, lui conférer un statut d'arabisme n'a pas plus de sens que de lui attribuer le statut d'hellenisme.

Ne restent plus, dès lors, que 241 cas d'emprunts apparemment directs. Mais sur base de l'examen attentif de quelques exemples (*henné*, emprunté par l'intermédiaire du latin écrit, *antari*, issu d'un anthroponyme, *fez*, apparemment issu d'un toponyme), nous montrerons la fragilité de ces décomptes trop péremptoires. Tous ces cas «méritent»-ils leur statut d'arabisme ? C'est une question qui doit être posée et qui, en tout cas, invite à la prudence; des concepts simplistes ou trop monolithiques sont souvent à remettre en question.

Mais si l'on revient à notre propos de base, à savoir tenter de débusquer les arabismes du français dans le TLF, deux autres éléments sont à interroger: qu'est-ce que *le* français, d'une part (nous évoquerons le cas des terminologies, mais également l'aspect diachronique; si le TLF conserve des mots vieillis, sa nomenclature n'accueille pas les emprunts les plus récents), et quel est exactement l'état de la science que donne à voir le TLF. En effet, cet ouvrage colossal est une synthèse magnifique d'un état ancien de la recherche. Dès lors, certaines notices mériteraient d'être réexaminées. C'est l'ambition du projet TLF-Etym, qui a pour objectif de revoir certaines des notices «étymologie et histoire» du *Trésor de la langue française* (v. http://www.atilf.fr/tlf-etym/) afin, d'une part, d'intégrer les nouveaux matériaux dont disposent les lexicographes (éditions de textes, glossaires,

[3] Ce choix permet d'éviter de «gonfler» les chiffres et de surpondérer ainsi certaines catégories de mots.

études sur un lexème, etc.) et d'autre part, d'appliquer aux matériaux anciens un traitement répondant aux méthodes actuelles.

Divers outils de recherche, tels que les moteurs de recherche de *Google*, permettent aujourd'hui de trouver d'anciennes attestations. Enfin, l'étymologie ne se conçoit aujourd'hui que pourvue d'une véritable histoire du mot, de sa naissance à ses derniers développements (syntagmes ou sens nouveaux). Dans ce cadre, nous examinerons à nouveau les cas d'*antari* (qui se révèlera finalement être un vrai emprunt) ou de *fez* (qui, en fait, est un calque et non un emprunt).

4 Conclusions

Combien d'arabismes le français compte-t-il ? On en dénombre 546 selon Majid El Houssi; 520 d'après Hassane Makki; 508, si l'on en croit Walter et Baraké (alors que dans un autre ouvrage, Henriette walter avançait le chiffre de 420). Ou encore, 950 chez Dîb Farâdj, même si Françoise Quinsat estime, dans son compte rendu, que «beaucoup ne sont pas des arabismes en réalité. Il y a parmis les entrées de ce répertoire seulement environ 160 véritables arabismes du français» (Aljamía 16, 2004, pp. 287-291). Quant au TLF, selon la finesse de la lecture, nous y avons dénombré d'abord 883, puis 216, puis 460, enfin 241, avant de souligner le peu de pertinence de ces chiffres.

Nous espérons avoir montré l'importance du travail méticuleux du linguiste qui, dans le cas du TLF, ne peut pas être remplacé par la machine. Il est parfois dangereux de faire confiance à un moteur de recherche. Et même, lorsqu'on compare les résultats entre la recherche «complexe» et la recherche «simple», on voit que les outils qui semblaient les plus «avancés» ou les mieux pensés sont inutilisables pour des recherches visant à une certaine exhaustivité, en tout cas en ce qui concerne le TLF.

Pourquoi ? Parce qu'il s'agit d'une entreprise conçue secondairement par rapport à la version «papier». Les balises ont été pensées, on l'a vu, mais il n'en reste pas moins dangereux de les utiliser sans contrôle. Nous tenons aussi à rappeler la date de 1995. Il est évident que depuis, de l'eau a coulé sous les ponts de la lexicographie informatique. Néanmoins, des problèmes similaires se poseront pour toute entreprise de rétroconversion, sauf à envisager son objet comme un objet *historique* et à chercher à distinguer les différentes strates chronologiques présentes dans l'ouvrage afin de les traiter séparément. Dans le cadre du TLF, il s'agirait peut-être de distinguer la partie synchronique de la partie diachronique, qui ont été conçues indépendamment l'une de l'autre, ce qui limiterait les risques d'interférences de recherche; dans le cadre d'autres ouvrages, il faudra considérer isolément les articles dus à des rédacteurs différents, etc.

Évidemment, c'est tout de même grâce à la version informatisée que nous avons pu réunir un corpus; elle est donc un outil éminemment précieux, dont nous ne voudrions pas nous passer. En revanche, s'il convient de se méfier des outils informatiques, on doit surtout se méfier de soi-même. Les chiffres ou les informations données par un ordinateur doivent toujours être interrogés, analysés et remis en cause; c'est seulement à ce prix que l'informatisation des données est une réelle avancée. L'informatique, pour complexe ou puissante qu'elle soit, doit rester subordonnée à l'analyse; elle n'est qu'un des outils du lexicographe.

Références

Arveiller, Raymond, 1999. *Addenda au FEW XIX (Orientalia)* (Max Pfister éd.), Beihefte zur Zeitschrift für Romanische Philologie, Band 298, Tübingen, Max Niemeyer.

Baiwir, Esther, à paraître. «Examen méta-lexicographique des arabismes du TLF(i): une tentative de classement chronologique des emprunts directs et indirects du français à l'arabe».

Buchi Éva, 2005. «Le projet TLF-Étym (projet de révision sélective des notices étymologiques du *Trésor de la langue française informatisé*)», in *Estudis romànics* 27, 569-571.

Dîb Farâdj Allâh Sâlih, 2001. *Mosaïque: des mots arabes dans la langue française, Muzawwaqāt, min kalām al-ʿarab fī al-luġah al-firansiyyah*, Beyrouth, Naufal.

El Houssi, Majid, 2002. *Les arabismes dans la langue française. Du moyen Âge à nos jours*, Torino - Paris, Harmattan.

Kiesler, Reinhard, 2006. «Sprachkontakte: Arabisch und Galloromania. Contacts linguistiques: arabe et Galloromania», in *Romanische Sprachgeschichte*, HSK 23.2., Berlin-New York, de Gruyter, 1648-1655.

Projet TLF-Étym: mise à jour des notices étymologiques du Trésor de la langue française informatisé. *Dossier de présentation*, Collectif, 2005. Nancy, ATILF/CNRS/Université Nancy 2/UHP.

Quinsat, Françoise, 2008. «Remarques sur le traitement des arabismes dans le TLF(i): premier bilan et perspectives», in *Zeitschrift für romanische Philologie*, vol. 124, n°3, 402-417.

TLF = Imbs, Paul / Quemada, Bernard (dir.), 1971-1994. *Trésor de la langue française. Dictionnaire de la langue du XIXe et du XXe siècle (1789-1960)*, 16 vol., Nancy, CNRS Éditions/Gallimard.

Walter, Henriette et Baraké, Bassam, 2007. *Arabesques: l'aventure de la langue arabe en Occident*, Paris, Point.

http://atilf.atilf.fr/tlf.htm

http://www.atilf.fr/tlf-etym/

http://www.atilf.fr/IMG/pdf/La_preface_du_TLFi_par_Jean.pdf