Les dictionnaires multilingues: de la tradition lexicographique à l'ère digitale[1]

Marius-Radu CLIM[*]

Keywords: *dictionary*; *neologism*; *lexicography*; *terminology*; *multilingual online resources*; *corpora*; *online multilingual dictionary*

1. Le rôle des dictionnaires et le problème du néologisme

Pendant les dernières décennies, le domaine de la lexicographie connaît une évolution remarquable, motivée par le développement social, par la globalisation, par l'agrandissement des aires de recherches philologiques et bien sûr par le contexte de l'informatisation et du développement des technologies digitales.

Les dictionnaires ont représenté toujours un instrument d'éducation culturelle dont l'utilité ne cesse d'agrandir. Dans une étude sur les tendances actuelles de la lexicographie française, Alain Rey souligne que l'émergence des dictionnaires a été motivée par des besoins sociaux, la preuve étant l'antériorité des dictionnaires bilingues par rapport aux dictionnaires monolingues. Il considère que la très longue tradition lexicographique est un frein dans l'adaptation à de nouvelles conditions théoriques: « l'ancienneté et le caractère conventionnel de la pratique lexicographique expliquent la lenteur d'une adaptation à des conditions théoriques nouvelles et encore indécises » (Rey 1970 : 471). Par conséquent, il propose d'analyser les dictionnaires comme des objets métalinguistiques qui peuvent mettre en évidence l'attitude sociale concernant la langue parlée :

> il faut insister sur l'intérêt d'une étude concernant le dictionnaire en tant qu'objet métalinguistique. Cette étude stimulerait la réflexion sémantique et serait révélatrice des attitudes sociales vis-à-vis du langage. Une science lexicographique serait ainsi... une contribution à la socio-linguistique (Rey 1970 : 473).

Ces œuvres lexicographiques ont été toujours un outil de standardisation, mais ils ont eu aussi comme but de représenter l'évolution culturelle et la conception des utilisateurs concernant la vie et le monde: « au-delà de l'"outil", le dictionnaire est l'un des reflets de la culture d'un pays: derrière un dictionnaire, en effet, il y a une langue, une communauté linguistique, une civilisation » (Ridel 2009 : 2). En analysant les dictionnaires, Ridel considère qu'ils ont comme but, d'une part, de faire

[1] Cet article a été financé par un projet de l'Autorité Nationale Roumaine pour la Recherche Scientifique et Innovation, CNCS–UEFISCDI, numéro de projet PN-II-RU-TE-2014-4-0195.

[*] Institut de Philologie Roumaine „A. Philippide", Académie Roumaine, Filiale de Iasi, Roumanie.

connaître une langue nationale et, d'autre part, de la standardiser. Parce que les dictionnaires mettent en évidence les cultures des nations, ils constituent aussi un instrument de conservation, non seulement de diffusion, car « le dictionnaire bilingue représente une forme de reconnaissance des langues nationales: la reconnaissance d'une identité linguistique et culturelle propre à chaque pays » (Ridel 2009 : 2). En plus, R.P.K. Hartmann et G. James soulignent l'autorité finale des dictionnaires en ce qui concerne l'utilisation et le sémantisme des mots, mais aussi en ce qui concerne la protection de l'identité d'une langue :

> The dictionary is supposed to represent some form of final authority in matters of lexical meaning and use. The academy dictionaries typically exert considerable influence… in protecting a language from what are perceived as unacceptable or corrupting pressures, for example, excessive borrowing from other languages (Hartmann, James 1998: IX).

On mentionne aussi le fait que les dictionnaires ne représentent pas une liste exhaustive des mots d'une langue, mais, selon le type de dictionnaire ou selon les destinataires sélectionnés, on peut choisir une partie du lexique d'une langue (Guilbert 1975 : 38). Mais ces œuvres sont premièrement des modèles pour l'utilisation correcte d'une langue, ayant une fonction descriptive-normative. La même idée est soulignée par Aïno Niklas-Salminen, dans son étude dédiée à la néologie, dans laquelle elle insiste sur le rôle des dictionnaires d'offrir une norme idéale qui devient représentative pour tous les locuteurs : « Le dictionnaire, comme tous les ouvrages de l'enseignement, vise à donner une image de l'homme, norme idéale à laquelle doivent se conformer tous les locuteurs » (Niklas-Salminen 2001 : 118).

Le domaine de la néologie en tant que recherche scientifique suppose l'analyse théorique, la description des innovations lexicales, le processus de formation des unités lexicales, des activités de politique linguistique, d'analyse, de standardisation et de promotion des néologismes, la création des terminologies et des dictionnaires spécialisés pour différents domaines techniques. Même si personne ne connaît exactement comment les langues changent (Coşeriu 1997 : 206) et ne peut prévoir les changements linguistiques, la néologie est « un indicador de l'estat d'una llengua » (Llengua 2004 : 32) en mettant en évidence le niveau de vitalité d'une langue, ayant une applicabilité immédiate dans les domaines de la lexicographie, de la terminologie et de la planification linguistique.

2. Les tendances de la lexicographie contemporaine

Une tendance définitoire pour le développement de la lexicographie et pour l'évolution des dictionnaires est celle nommée « user perspective », c'est-à-dire ce qui est dominant dans la réalisation et la diffusion des dictionnaires, voir l'utilisateur : quels sont ses besoins, ses raisons pour ouvrir un dictionnaire et dans quels contextes il y fait appel. Cette orientation a déterminé la création de dictionnaires de plus en plus diversifiés, qui correspondent aux différents profils d'utilisateurs, aux contextes dans lesquels ils les utilisent, aux compétences qu'ils possèdent ou acquièrent et aux moyens dont ils disposent.

La lexicographie contemporaine traverse maintenant une nouvelle étape de développement, qui est définitoire pour l'évolution de cette discipline et qui a

produit des changements significatifs dans ce domaine. Cette évolution est motivée par le phénomène de la globalisation et du développement digital. Cette étape est représentée par le processus de digitalisation des moyens d'étude lexicographique et aussi des dictionnaires. Ainsi, depuis les années 90 on remarque une préoccupation évidente des lexicographes roumains pour mettre en format numérique autant de dictionnaires que possible.

Ces préoccupations pour digitaliser les langues ont conduit à la création des organisations internationales qui rassemblent des lexicographes, des lexicologues, des linguistes, des informaticiens qui ont tous comme but d'atteindre le plus haut niveau de technologie pour autant de langues que possible à travers le monde. Cette tendance confirme la remarque d'Alain Danzin : « à l'ère électronique, il est essentiel pour la survie d'une langue qu'elle soit utilisée dans les systèmes d'information électroniques » (*apud* Tufiş, Filip 2002 : 137). Ainsi, en 1975 on a vu paraître le Dictionary Society of North America (DSNA), European Association for Lexicography (EURALEX) en 1983, Australasian Association for Lexicography (AUSTRALEX) en 1990, African Association for Lexicography (AFRILEX) en 1995 et Asian Association for Lexicography (ASIALEX) en 1997, ayant comme but de coordonner plus efficacement les activités et les projets de digitalisation des langues. En 2016 on a fondé le GLOBALEX pour faciliter les contacts entre les membres des associations continentales.

Les avantages des dictionnaires électroniques, disponibles pour le grand public, sont énormes. Tout d'abord, ils représentent une ressource très importante pour la recherche en matière de technologie linguistique, en raison du vaste contenu de l'information d'histoire linguistique, de sémantique lexicale, etc. Ensuite, informatiser un dictionnaire apporterait de nombreux avantages par rapport aux éditions imprimées, y compris: la possibilité de mettre à jour, avec le rythme d'évolution de la langue, la recherche automatique des phénomènes linguistiques (l'évolution des néologismes, par exemple), l'extraction automatique des dictionnaires (étymologique, phraséologique, néologique, etc., en particulier du dictionnaire thésaurus d'une langue), l'exploitation de la collection des collocations des significations des mots pour perfectionner les programmes de désambiguïsation sémantique. Mais pour qu'un tel instrument puisse être fait, il est nécessaire, tout d'abord, de constituer un corpus de textes qui devra inclure des sources (textes écrits ou audio) nombreuses et variées qui serviront de source et d'exemple dans la rédaction des mots et, d'autre part, créer des programmes qui permettent l'analyse et la reconnaissance des mots et des traits phonétiques, morphologiques, syntaxiques et sémantiques de ceux-ci.

3. La création des terminologies scientifiques

Le développement technologique a imposé la création des terminologies internationales. Les scientifiques ont la tendance « îndreptăţită din punct de vedere teoretic, de a întrebuinţa o terminologie convenţională, care în cea mai mare parte e latino-grecească » (Puşcariu 1940: 381). En plus, les échanges commerciaux ont favorisé la circulation non seulement des produits, mais aussi des termes assignés.

Cela a déterminé l'apparition de plusieurs organismes internationaux préoccupés par l'analyse et l'adaptation des néologismes et des terminologies scientifiques, comme par exemple, *Realiter* ou *le Réseau panlatin de terminologie*[2]. Ce réseau, qui fait partie de l'Union Latine, a été créé en 1993 à Paris et comprend des personnes, des institutions et des organismes de pays de langues néolatines actifs en terminologie. Cette collaboration a comme objectifs :

a) « établir des principes méthodologiques communs applicables à la réalisation des produits élaborés conjointement;

b) mener des recherches en commun et créer des outils susceptibles de favoriser le développement des langues latines;

c) mener des travaux terminologiques conjoints multilingues dans des domaines d'intérêt commun touchant la société;

d) mettre en commun les matériaux de référence documentaires;

e) favoriser la formation réciproque à travers les échanges de formateurs, d'experts, d'étudiants et de matériaux didactiques »

(http://www.realiter.net/presentazione?lang=fr).

Une préoccupation particulière constitue l'assurance d'une évolution commune de la néologie scientifique et technique dans les langues néolatines. Cela a conduit au développement d'un programme pour créer des dictionnaires multilingues dans différents domaines d'intérêt: internet, informatique, nanotechnologie, génétique, bioéthique, e-commerce, etc. Les ouvrages réalisés jusqu'à présent (dont on peut mentionner *Lexique panlatin de l'internet*, *Léxico panlatino de terminologia do ambiente*, *Neologismos económicos en las lenguas románicas à través de la prensa*) comprennent des termes des sept langues néolatines (le catalan, l'espagnol, le français, le galicien, l'italien, le portugais, le roumain) et des termes anglais, quand ils représentent des notions de référence dans le domaine. Pour la langue romaine, les préoccupations concernant les terminologies sont mises en évidence par la création de la *Banque de données terminologiques* (BDT), réalisée par l'Association Roumaine de Terminologie, TermRom. A l'aide de Realiter, ces données ont été améliorées et ont été faites publiques, accessibles à tous les chercheurs de l'espace latin (Tufiş, Filip 2002 : 29).

Dans ce réseau on analyse non seulement les langues littéraires, mais aussi les régionalismes. Les dictionnaires et les autres travaux réalisés sont disponibles sur le site du réseau. Ce site vise à rassembler toutes les ressources électroniques liées aux langues néolatines: dictionnaires anciens et modernes, grammaires, terminologies, outils d'apprentissage des langues, enregistrements audio et vidéo, etc. Ce moyen électronique cherche à mettre en évidence le patrimoine de ces langues et permet l'accès rapide à ces ressources.

Dans le réseau REALITER on a initié un projet de recherche intitulé NEOROM, qui propose l'analyse contrastive des néologismes des langues romanes. La première étape a été la création d'une base de données néologiques et l'analyse contrastive de 1000 néologismes de la presse des langues néolatines. Cela a

[2] Pour une présentation plus détaillée concernant l'adaptation des terminologies voir Clim (2012 : 140–147).

impliqué différents groupes de chercheurs[3] des pays néolatins, qui ont cherché à mettre en évidence les ressources internes des langues, ainsi que les moyens spécifiques d'enrichissement du lexique.

Terminometro, également créé au sein de l'Union latine, est un autre outil important de terminologie. *Terminometro*[4] était à l'origine un bulletin de référence, mais aujourd'hui il est devenu un portail multilingue sur la terminologie, l'ingénierie linguistique, la traduction scientifique et technique, les dictionnaires, la documentation, la rédaction technique. Ses objectifs sont, d'une part, de présenter dans la version électronique les nouvelles des domaines mentionnés et, d'autre part, de donner accès à toutes les informations contenues sur tous les sites Web latino-américains et de permettre des recherches simultanées dans plusieurs banques de données terminologiques. En outre, le site contient de nombreux articles sur les dictionnaires, des outils de traduction automatique et linguistique, des résumés de diverses conférences, des périodiques, etc., rédigés dans les langues du portail.

4. Des ressources multilingues digitales

Le traitement automatique du langage naturel ou de la langue naturelle (abr. TALN) ou des langues (abr. TAL) est une discipline à la frontière de la linguistique, de l'informatique et de l'intelligence artificielle, qui concerne l'application de programmes et techniques informatiques à tous les aspects du langage humain (Charniak, McDermott 1985 : 2).

Le champ du traitement automatique du langage couvre de très nombreuses disciplines de recherche qui peuvent mettre en œuvre des compétences diverses, comme par exemple : la syntaxe (lemmatisation, étiquetage morphosyntaxique, analyse syntaxique), la sémantique (traduction automatique, résumé automatique, désambiguïsation lexicale), le traitement du signal ou l'extraction d'informations.

La version électronique d'un dictionnaire peut offrir de nombreuses possibilités pour les applications de traitement du langage naturel, telles que: la

[3] Du projet NEOROM font partie:

1. *Observatori de Neologia* (OBNEO) de l'Institut Universitaire de Linguistique Appliquée de l'Université Pompeu Fabra, Barcelone, qui étudie les langues catalane et castillane;
2. *Observatorio de Neoloxia* de l'Université de Vigo, qui étudie la langue galicienne.
3. *Osservatorio neologico della lingua italiana* (ONLI) de l'*Istituto per il lessico intellettuale europeo e storia delle idee* (ILIESI-CNR) et l'Université La Sapienza de Rome, qui étudie la langue italienne.
4. *Observatoire de Néologie du français de France* (NEOFRAN), de l'Université Paris VII-Jussieu din Paris, qui étudie la langue française. Les responsables de projet sont John Humbley et Jean-François Sablayrolles;
5. *Observatório de neologismos do português contemporâneo do Brasil*, de l'Université de São Paulo, qui étudie la langue portugaise parlée au Brésil;
6. *Observatório de Neologia do Português* (ONP), de l'Instituto de Linguística Teórica e Computacional (ILTEC) de Lisboa;
7. *Observatoire de Néologie du français du Québec*, de l'Université Laval et de l'Office Québecois de la Langue Française;
8. *Observatoire de Néologie du français de Belgique*, de l'Institut Marie Haps de Bruxelles;
9. *Observatorul neologic român* (ONEROM), de l'Institut de Linguistique „Iorgu Iordan – Al. Rosetti" de Bucarest. Responsable de projet: Ioana Vintilă-Rădulescu.

[4] http://www.unilat.org/DTIL/Terminologie/Terminometro/ro.

désambiguïsation des sens, l'annotation sémantique, l'apprentissage automatique, le traitement du discours, la traduction automatique. Le développement de ces technologies représente la condition de base pour l'intégration de la langue roumaine dans le grand groupe des langues informatisées (Vazquez et alii 2015: 96–97). C'est pourquoi, dans la plupart des langues, il existe une préoccupation pour la digitalisation d'un grand nombre d'œuvres lexicographiques.

Mais les préoccupations ne se limitent pas à des dictionnaires en ligne, elles vont aussi vers la création de ressources linguistiques pour autant de langues, par des normes communes et leur traitement digital (cf. Tufiş, Filip 2002: 138). Tenant compte de notre société multilingue, ces ressources se sont révélées indispensables pour au moins deux raisons: d'abord elles sont indispensables pour préserver les langues et les cultures présentes, et, d'autre part, toujours indispensables pour la recherche dans le domaine des ressources multilingues.

La plus importante ressource multilingue c'est WordNet. Ce projet a été développé depuis 1985 à l'Université Princeton, et a été coordonné par le psycholinguiste George Miller. WordNet est une base de données lexicales et son but est de répertorier, classifier et mettre en relation de diverses manières le contenu sémantique et lexical de la langue anglaise. Le lexique de WordNet est structuré en quatre grandes sur-catégories lexicales : les noms, les verbes, les adjectifs et les adverbes. La composante atomique sur laquelle repose le système entier est le synset (*synonym set*), un groupe de mots interchangeables, dénotant un sens ou un usage particulier. Selon le modèle d'un dictionnaire traditionnel, WordNet offre, pour chaque mot, une liste de synsets correspondant à toutes ses acceptions répertoriées. Mais les synsets ont également d'autres usages : ils peuvent représenter des concepts plus abstraits, d'un niveau plus haut que les mots et leurs sens, qu'on peut organiser sous forme d'ontologies. Une ontologie est un système de catégories permettant de classifier les éléments d'un univers. Le système de catégorisation correspond aux relations sémantiques. Ceci permet de regrouper de manière cohérente toutes les composantes d'un univers linguistique tel que les mots, les sens ou bien les concepts.

La relation sémantique servant de critère pour l'agrégation d'un groupe de concepts définira le type de l'ontologie. WordNet répertorie ainsi une grande variété de relations sémantiques permettant d'organiser le sens des mots (et donc par extension les mots eux-mêmes) en des systèmes de catégories qu'on peut consulter de manière cohérente et uniforme : antonymie/synonymie, hyperonymie/hyponymie, homonymie/méronymie etc.

WordNet en tant que ressource publique répertorie plus de 200 000 mots de classes ouvertes, pour lesquelles l'ajout d'éléments lexicaux est possible, ainsi que plus de 115 000 synsets. L'influence de ce projet a été énorme dans le domaine de la technologie du langage et les bénéfices en sont évidents ! La Commission Européenne a financé pendant 1996 et 1998 un projet similaire, EuroWordNet. Ce projet a permis la constitution de ressources similaires à Wordnet pour l'anglais, le hollandais, l'italien, l'espagnol, l'allemand, le français, le tchèque et l'estonien. Mais ce projet ne vise pas seulement à créer des wordnets pour ces langues, mais aussi à lier ces réseaux lexicaux de sorte que le synset d'une langue soit lié à son équivalent dans les autres langues. En outre, EuroWordNet propose un nombre beaucoup plus élevé de relations, à savoir 90, telles que les relations thématiques de

type causal (Agent, Patient, Instrument, Localisation, Direction), ou celles qui corrèlent les significations des dérivés lexicaux. Pour la corrélation de ces réseaux, un index interlingual, indépendant du langage (ILI), a été défini. Chaque sens dans l'une des langues représentées dans le réseau sémantique multilingue est généralement conformé à un concept unique de l'indice interlingual. Ces correspondances sont faites à l'aide de 20 types distincts de relations binaires. Les synsets (les séries synonymes) dans deux langues ou plus qui correspondent au même concept d'ILI sont considérés comme équivalents de traduction (Tufiş, Filip 2002: 140).

Un autre projet a été lancé en 2001, comme un prolongement d'EuroWordNet, à savoir Balkanet, qui vise six langues des Balkans: le bulgare, le tchèque, le grec, le roumain, le serbe et le turc. Comme dans le projet précédent, les ontologies lexicales monolingues sont corrélées avec une multitude de concepts interlinguaux, les correspondances s'établissant au moyen de relations d'équivalence complexes (*eq-synonymy*, *eq-near-synonymy*, *eq-has-hyperonym* etc.). Pour le roumain, un corpus de plus de 100 millions de mots a été utilisé dans une collection de romans et de textes journalistiques gratuits de web. Ce corpus a ensuite été analysé selon les 4000 concepts interlinguaux sélectionnés dans ce projet.

De tels projets nécessitent une mise à jour et une maintenance continue, car le wordnet est à la fois un dictionnaire et un corpus annoté au niveau du sens. En 2006, le wordnet roumain avait atteint 33421 synsets, et son développement continue, car il représente l'une des ressources lexicales les plus importantes pour la langue roumaine, étant indispensable dans les applications du traitement du langage naturel (RLIPLR 2006: 17). Les moyens par lesquels cette ontologie lexicale peut se développer sont des corpus multilingues. Un tel corpus est JRC-Acquis, qui englobe les lois et les traités communs à tous les États membres de l'Union Européenne. C'est l'un des plus grands corpus parallèles, et le grand nombre de langages constitutifs en fait l'outil parfait pour valider une ontologie lexicale multilingue.

À la fin du projet BalkaNet, ona discuté la possibilité d'améliorer ces ressources en ajoutant des concepts ou des réalités propres à la région des Balkans. Ainsi, pour la langue roumaine, les résidences de toutes les regions du pays ont été ajoutés en assignant un synset séparé avec un format spécifique (Mitocariu et alii 2013 : 109). Le WordNet roumain (RoWN) est continuellement développé par édition manuelle et par enrichissement semi-automatique. Une telle application semi-automatique permet d'accéder à RoWN ainsi qu'à des opérations plus complexes, en corrigeant certaines structures et en réalisant des recherches plus profondes (Mitocariu et alii 2013 : 117).

De telles listes liées sémantiquement, comme WordNet, sont utilisées dans de nombreux processus de traitement automatique du langage. De nombreuses approches visent actuellement à analyser l'attitude ou les sentiments des auteurs de documents, à travers lesquels les documents peuvent être classés. Nancy Ide insiste sur la nécessité d'améliorer les types de relations sémantiques existentes dans WordNet en créant automatiquement des listes de mots alignés au sens (Tufiş, Forăscu 2010: 225). Afin d'analyser l'attitude, il est nécessaire non seulement d'aligner les mots, mais il faut que le corpus analysé soit aligné au niveau du sens. L'alignement manuel de tels corpus nécessite des coûts très élevés, et c'est pourquoi, selon l'auteur, il est nécessaire de regrouper les significations dans

WordNet en des catégories moins granulaires, nommées catégories sémantiques grossières. Et ça parce que les erreurs des processus de désambiguïsation sémantique sont données par la granularité des significations dans WordNet. Par conséquent, un inventaire facile à utiliser pourrait aider à automatiser et mieux organiser les processus de désambiguïsation sémantique.

Un autre type de corpus multilingue c'est Wikipédia. Wikipédia est un projet d'encyclopédie collective en ligne, universelle, multilingue, qui a pour objectif d'offrir un contenu librement réutilisable, objectif et vérifiable, que l'utilisateur peut modifier et améliorer. Selon les informations présentées sur le site du projet (https://fr.wikipedia.org/wiki/), il a été créé par Jimmy Wales et Larry Sanger le 15 janvier 2001. L'encyclopédie est en libre accès, en lecture comme en écriture, c'est-à-dire que n'importe qui peut, en accédant au site, modifier la quasi-totalité des articles publiés sous licence CC-BY-SA 3.0. En octobre 2017, Wikipédia était le cinquième site le plus fréquenté au monde, constituant le plus grand et le plus populaire des ouvrages de références générales d'Internet.

Wikipédia a pour slogan : « Le projet d'encyclopédie librement distribuable que chacun peut améliorer ». Ce projet est décrit par son cofondateur Jimmy Wales comme « un effort pour créer et distribuer une encyclopédie libre de la meilleure qualité possible à chaque personne sur la planète dans sa propre langue ». Ainsi, Jimmy Wales proposa comme objectif que Wikipédia puisse atteindre un niveau de qualité au moins équivalent à celui de l'*Encyclopædia Britannica*. C'est un exemple de collaboration massive à but non lucratif.

En revanche, Wikipédia n'a pas pour but de présenter des informations inédites, et ne vise donc qu'à exposer des connaissances déjà établies et reconnues. Historiquement, l'anglais a été la principale langue utilisée, avant qu'une multitude de sites ne soient ouverts dans d'autres langues (283 langues en 2016). Le Wikipédia en anglais a toujours conservé cette importance relative : le nombre d'articles s'élève à plus de 5 millions. Tous les articles de Wikipédia évoluent en permanence, grâce à des milliers de contributeurs bénévoles. Wikipédia propose des articles sur un maximum de sujets, en principe en s'appuyant sur des sources externes sérieuses, dont ils offrent la synthèse. Pour chaque article, les références devraient être accessibles par les annotations et la bibliographie. Ceci permet au lecteur de pouvoir utiliser son esprit critique en ayant à tout moment une idée sur la fiabilité du contenu, ou lui offrant des références qui lui permettent d'aller compléter ses recherches. Wikipédia est réalisé d'une manière collaborative, sur Internet, via un « réseau coopératif » auto-organisé et sans frontières linguistiques. Le système wiki de Wikipédia permet la création et la modification immédiates des pages par tous les visiteurs, même sans inscription. Wikipédia fut la première encyclopédie généraliste à ouvrir, grâce à ce système, l'édition de ses articles à tous les internautes. Aucun article n'est considéré comme achevé et Wikipédia se présente comme un projet en amélioration continue. La constante surveillance des modifications est également ouverte à tous à travers le système wiki. Il n'y a aucun système hiérarchique de validation ; aussi l'encyclopédie est-elle l'objet de nombreuses incompréhensions et critiques quant à la qualité et à la fiabilité de son contenu, et l'objet d'études sur sa fiabilité en anglais, la langue la plus développée.

Un des principes fondateurs de Wikipédia est la neutralité du point de vue. Le projet se veut universel, en traitant tous les domaines de la connaissance, y compris la culture populaire, multilingue et gratuite dans sa version en ligne, afin de favoriser l'accès sans limitations à la connaissance.

Wikipédia est disponible sous licence libre, ce qui signifie que chacun est libre de la recopier, de la modifier, et de la redistribuer gratuitement et onéreusement.

Autour de ce réseau se sont formés plusieurs, plus petits et plus spécifiques, comme par exemple : Wiktionary, un dictionnaire et thésaurus créé le 12 décembre 2002 ; Wikiquote, un recueil de citations (27 juin 2003) ; Wikibooks, un annuaire de livres électroniques destinés aux étudiants (10 juillet 2003) ; Wikisource, un recueil de textes dans le domaine public (23 novembre 2003) ; Wikinews, un site d'informations (décembre 2004) ; Wikispecies, un répertoire du vivant (2004) ; Wikiversity, une communauté pédagogique créée en 2006 ; et Wikivoyage, un guide touristique en ligne (octobre 2012). Créé en 2001, Meta-Wiki est un wiki utilisé pour coordonner tous ces projets et servir à la communication entre les communautés linguistiques de Wikipédia, celles des projets frères, et la Wikimedia Foundation.

Depuis sa création, le contenu de l'encyclopédie Wikipédia n'a cessé de grandir, en quantité, dans des proportions difficilement imaginables à ses débuts. Le 19 janvier 2015, le nombre total d'articles de l'ensemble des éditions de Wikipédia était 38 229 930.

Aujourd'hui le plus grand dictionnaire multilingue est BabelNet. Sur le site de ce corpus (http://babelnet.org/about) BabelNet est à la fois un dictionnaire encyclopédique multilingue, avec une couverture lexicographique et encyclopédique des termes, et aussi un réseau sémantique qui relie les concepts et les entités nommées dans un très grand réseau de relations sémantiques, composé d'environ 14 millions d'entrées appelées synsets de Babel. Chaque synset de Babel représente une signification donnée et contient tous les synonymes qui expriment ce sens dans une gamme de langues différentes.

BabelNet 3.7 couvre 271 langues et il est obtenu à partir de l'intégration automatique de:

a) WordNet (version 3.0);

b) WordNet multilingue ouvert, une collection de wordnets disponibles en différentes langues (téléchargé en août 2015);

c) Wikipédia, la plus grande encyclopédie Web multilingue, collaborative (décharge de novembre 2014);

d) OmegaWiki, un grand dictionnaire multilingue collaboratif (décharge de juillet 2015);

e) Wiktionary, un projet collaboratif pour produire un dictionnaire multilingue à contenu libre (décharge d'août 2014);

f) Wikidata, une base de connaissances gratuite qui peut être lue et éditée par les humains et les machines (décharge de novembre 2014);

g) Wikiquote, un recueil en ligne gratuit de citations provenant de personnalités et d'œuvres créatives dans toutes les langues (décharge de mars 2015);

h) un Lexique de verbe basé sur les classes (version 3.2);

i) un ensemble de terminologies pouvant être utilisées pour développer des versions localisées d'applications (décharge de juillet 2015);

j) une base de données géographiques libre, couvrant tous les pays et contenant plus de huit millions de noms de lieux (décharge d'avril 2015);

k) une traduction française automatique améliorée, élargie et évaluée de WordNet (version haute précision, téléchargée en août 2015);

l) une base de données lexico-sémantiques développée dans le cadre de deux projets de recherche différents: EuroWordNet et SI-TAL (téléchargés en décembre 2015);

m) une base de données d'images organisée selon la hiérarchie WordNet (version 2011);

n) une base de données lexicale de l'anglais à la fois lisible par l'homme et la machine (version 1.6);

o) les mappings générés automatiquement parmi les versions de WordNet (version 2007).

BabelNet 3.7 contient plus de 13 millions de synsets et à peu près 745 millions de sens (indépendamment de la langue). Chaque synset de Babel contient en moyenne 2 synonymes par langue. Le réseau sémantique comprend toutes les relations lexique-sémantiques de WordNet (hyperonymie et hyponymie, méronymie et holonymie, antonymie et synonymie, etc., pour un total d'environ 364 000 arcs relationnels) ainsi qu'une relation générique de corrélation par Wikipédia (pour un total 380 millions d'arcs). La version 3.7 fournit aussi 11 millions d'images associées à des synsets de Babel.

BabelNet a été utilisé pour la réalisation d'un système de désambiguïsation et de liaison des entités, Babelfy, qui – grâce à l'intégration entre les sens lexicographiques et les entités encyclopédiques en un seul réseau sémantique – obtient des performances à l'état de l'art en utilisant des algorithmes sur graphes.

Par conséquent, les corpus parallèles sont les outils les plus appropriés pour de nombreux processus automatisés d'apprentissage des langues naturelles, comme par exemple: l'analyse des attitudes et des opinions à partir de différents textes, la classification des documents, l'occurrence des termes utilisés, l'apprentissage automatique et autres.

5. Le rôle des dictionnaires multilingues à l'ère digitale

C'est intéressant le fait que de nombreux corpus parallèles sont également basés sur des dictionnaires, des lexiques ou des thésaurus. Mais les dictionnaires multilingues ne sont pas seulement des moyens pour réaliser des corpus en ligne, ils sont le résultat du développement et du traitement automatique de ces corpus.

Une pratique souvent utilisée c'est l'extraction des dictionnaires bilingues ou multilingues à partir des équivalents lexicaux extraits des corpus parallèles, alignés au niveau de la phrase. Les corpus représentent en fait un texte dans une langue, qui est traduit dans une ou plusieurs langues. Grâce à ce processus, les corpus parallèles sont convertis dans un nouveau corpus et alignés au niveau du mot. Cet alignement est accompli par divers algorithmes basés, en grande partie, sur la relation entre les termes et les occurrences d'éléments lexicaux dans une langue. Les corpus parallèles, composés d'un texte original et des versions traduites, permettent une analyse linguistique de nombreux phénomènes de traduction (Vazquez et alii 2015 : 31). Parmi d'autres applications digitales possibles, l'exploitation de ces corpus

permet la recherche d'informations multilingues, l'extraction des terminologies bilingues et même des lexiques spécialisés.

La préoccupation pour des ressources en ligne, pour des corpus ou pour des analyses automatiques du langage naturel, n'a pas diminué l'intérêt pour les dictionnaires multilingues. C'est tout à fait vrai qu'on assiste à un changement des moyens de travail, et aussi des moyens d'utilisation des dictionnaires. Par conséquent, aujourd'hui on fait distinction entre un dictionnaire imprimé, qui reste une autorité en ce qui concerne le lexique, et un dictionnaire en ligne qui peut être dynamique – c'est-à-dire qu'il peut présenter l'évolution de la langue dans une manière continue, mais il lui manque l'autorité académique, parce qu'il est actualisé par les utilisateurs – ou semi-dynamique, comme par exemple OED (Oxford English Dictionary), qui est mis à jour périodiquement par les lexicographes (Vazquez et alii 2015 : 460). En outre, les possibilités offertes par un dictionnaire en ligne sont de plus en plus bénéfiques pour l'utilisateur, qui sera toujours tenté de choisir un dictionnaire mis à jour avec différentes options multimédia, plutôt qu'un dictionnaire imprimé (Vazquez et alii 2015 : 297).

Les dictionnaires multilingues ne sont pas seulement un moyen pour construire des corpus de textes, mais, dans une certaine manière, ils sont aussi le moyen de rendre les ressources numériques disponibles. Ces ressources lexicographiques, ainsi que les multiples corpus parallèles, se trouvent à la base des moteurs de recherche virtuels (google, yahoo, baidu, aol, etc.) et sont indispensables à l'utilisateur contemporain.

En raison de ces dictionnaires multilingues et des corpus parallèles, un moteur de recherche comme Google peut fournir à un utilisateur des informations sur une recherche particulière dans plusieurs langues que celle dans laquelle la recherche a été générée. Et si l'utilisateur n'a pas les compétences linguistiques pour lire toutes les informations fournies, le moteur de recherche a la possibilité de traduire cette page dans la langue souhaitée, précisément à cause de ces dictionnaires multilingues. En d'autres termes, même si l'évolution digitale a conduit à une baisse des apparitions des dictionnaires imprimés, les dictionnaires multilingues ont connu au cours de la dernière période une évolution et un développement sans précédent. Bien que ce développement ne soit pas si visible, l'internaute bénéficie de ces ressources, car elles sont indispensables à la transmission d'informations et à leur traitement pour les besoins de l'utilisateur.

6. Conclusions

Ainsi, à l'ère numérique, les dictionnaires multilingues connaissent un développement continu et dynamique et deviennent indispensables pour l'environnement virtuel. D'un certain point de vue, tout internaute doit utiliser un dictionnaire multilingue, même s'il n'en a pas connaissance. Comme on peut observer, les dictionnaires multilingues peuvent être mis à jour automatiquement, mais ils nécessitent aussi une contribution humaine bien informée. Afin d'assurer l'exactitude d'une traduction, il est nécessaire non seulement d'avoir des corpus parallèles, mais aussi une supervision académique. Ces deux aspects sont absolument indispensables à la réalisation de dictionnaires multilingues sur l'autorité desquels on peut compter.

L'incroyable évolution des moyens de communication électroniques et la nécessité d'une transmission rapide de l'information sans barrières linguistiques ont conduit à l'émergence d'une nouvelle phase de la lexicographie et au processus de création des dictionnaires multilingues. Dans l'environnement digital, l'utilité des dictionnaires a augmenté. Contrairement aux dictionnaires imprimés multilingues qui intéressent parfois le lecteur, la version des dictionnaires en ligne est indispensable à tout utilisateur de l'environnement virtuel.

Bibliographie

Charniak, McDermott 1985 : Eugene Charniak, Drew McDermott, *Introduction to Artificial Intelligence*, Boston, Addison-Wesley Longman Publishing Co.

Clim 2012 : Marius-Radu Clim, *Neologismul în lexicografia românească*, Iaşi, Editura Universităţii „Alexandru Ioan Cuza".

Coşeriu 1997 : Eugeniu Coşeriu, *Sincronie, diacronie şi istorie. Problema schimbării lingvistice*. Versiune în limba română de Nicolae Saramandu, Bucureşti, Editura Enciclopedică.

Guilbert 1975 : Louis Guilbert, *La créativité lexicale*, Paris, Librairie Larousse.

Hartmann, James 1998 : R.R.K. Hartmann, Gregory James, *Dictionary of Lexicography*, Londra, Editura Routledge.

Llengua 2004 : *Llengua catalana i neologia*. Part I *La creació lèxica en català*. Part II *Recull de neologismes.* Coordinadores: Judit Freixa i Elisabet Solé, Barcelona, Editorial Meteora.

Mitocariu et alii 2013 : Elena Mitocariu, Mihai Alex Moruz, Dan Cristea, Dan Tufiş, Marius Clim (eds.) *Proceedings of the 9th International Conference „Linguistic Resources and Tools for Processing the Romanian Language" 16-17 May 2013*, Iaşi, Editura Universităţii „Alexandru Ioan Cuza".

Niklas-Salminen 2001 : Aïno Niklas-Salminen, *Sur la néologie et la norme*, în volumul *La norme lexicale*. Etudes rassemblées par Gilles Siouffi et Agnès Steuckardt, Montpellier, Université Paul-Valéry Montpellier III, collection Dipralang (E. A. 739), p. 109–126.

Puşcariu 1940 : Sextil Puşcariu, *Limba română*, vol. I, *Privire generală*. Bucureşti, Fundaţia pentru literatură şi artă „Regele Carol II".

Rey 1970 : Alain Rey, *Sémantique et lexicographie (orientations actuelles des dictionnaires français)* în *Actes du X^e Congrès International des Linguistes, Bucharest, 28 août – 2 septembre 1967*, Bucureşti, Editura Academiei Române, p. 469–473.

Ridel 2009 : Élisabeth Ridel, *Réflexions autour des dictionnaires bilingues et multilingues. Introduction à la problématique*, article disponible en ligne à l'adresse http://unicaen.fr/recherche/mrsh/files/Intro-Ridel_O.pdf.

RLIPLR 2006 : Corina Forăscu, Dan Tufiş, Dan Cristea (eds.), *Lucrările atelierului „Resurse lingvistice şi instrumente pentru prelucrarea limbii române"*, Iaşi, Editura Universităţii „Alexandru Ioan Cuza".

Tufiş, Filip 2002 : Dan Tufiş, Florin Gh. Filip (coord.), *Limba română în Societatea Informaţională – Societatea Cunoaşterii*, Bucureşti, Editura Expert.

Tufiş, Forăscu 2010 : Dan Tufiş, Corina Forăscu (eds.), *Multilinguality and Interoperability in language processing with emphasis on Romanian*, Bucureşti, Editura Academiei Române.

Vazquez et alii 2015 : María José Domínguez Vásquez, Xavier Gómez Guinovart, Carlos Valcárcel Riveiro (eds.) *Lexicografía de las lenguas románicas*, vol. II, *Aproximaciones a la lexicografía moderna y contrastiva*, Berlin, Editura De Gruyter.

Multilingual Dictionaries: from Lexicographic Tradition to Digital Era

In the digital age, multilingual dictionaries have a continuous, dynamic development and become indispensable for the virtual environment. The incredible evolution of electronic means of communication and the need for rapid transmission of information by eliminating language barriers have led to the emergence of a new phase of lexicography and the process of making multilingual dictionaries. In the digital environment the usefulness of dictionaries has increased. They contribute to the creation of multilingual corpora of different types (such as WordNet, BabelNet, Wikipedia, or others) or have become indispensable to the Internet user to provide the required information. The virtual user has to use a multilingual dictionary, even if he is not aware of it, and through these tools he can benefit in the desired information regardless of the language in which it was primarily written.