Synergies Italie n° 17 - 2021 p. 7-15

GERFLINT
ISSN 1724-0700
ISSN en ligne 2260-8087

Présentation

Annick Farina
Université de Florence, Italie
annick.farina@unifi.it
Valeria Zotti
Université de Bologne, Italie
valeria.zotti@unibo.it

Les industries de la langue ont connu une forte croissance depuis les années quatre-vingt-dix du siècle dernier, aussi bien au sein des institutions européennes et internationales, dans le cadre de programmes de promotion de la diversité linguistique, que dans le secteur privé qui a vu se multiplier le nombre de sociétés actives dans le périmètre de l'ingénierie linguistique, dégageant des chiffres d'affaires de plusieurs millions d'euros (Technolangue, 2007). Les industries de la langue constituent aujourd'hui une grappe industrielle en plein essor et sont des facteurs de croissance économique de plus en plus importants. L'usage du « langage naturel » au sein des systèmes d'information a été reconnu comme une source de productivité certaine, non seulement en termes économiques, mais aussi en termes de quantité de temps investi et de qualité du produit final. Ces vingt dernières années, la frénésie du marché de l'Internet et de la communication multilingue et la progression de l'usage des NTIC ont engendré un élargissement du marché des outils technolinguistiques vers le grand public. Le désir d'utiliser une interface conviviale en langage naturel, qui propose une traduction en ligne, fait désormais partie des exigences banales des internautes (Dabbadie, 2004).

Ce numéro de *Synergies Italie* est l'occasion de présenter quelques recherches en cours en France et en Italie dans le domaine linguistique qui ont des applications concrètes au sein des industries des langues, ainsi que d'analyser les produits (corpus, bases de données terminologiques, outils de traitement automatique) existants, un quart de siècle après l'essor de ces industries. Nous nous sommes intéressées aux retombées que l'emploi généralisé des technologies du traitement automatique des langues (TAL) a eues dans ces deux pays sur le grand public, sur les professionnels de la langue (traducteurs, terminologues, lexicographes) et dans le monde de la recherche universitaire. Nous nous sommes interrogées en particulier sur la présence de la langue française et de la langue italienne (et de la traduction de l'une à l'autre) à l'intérieur des nouveaux instruments développés.

Fortes d'une expérience partagée tant pratique que théorique dans le domaine de la lexicographie monolingue francophone et bilingue franco-italienne et d'un intérêt sans cesse renouvelé pour les nouvelles technologies reliées à l'analyse du lexique, nous nous sommes naturellement tournées vers les méthodes les plus récentes de traitement automatique de la langue. Depuis 2013, nous nous sommes engagées dans un projet interuniversitaire commun, le *Lessico dei Beni Culturali*, visant à la réalisation d'un nouveau dictionnaire plurilingue, entièrement basé sur corpus, qui devrait permettre aux professionnels de la langue et du tourisme d'accéder à une description riche et fiable du lexique du patrimoine artistique. Dans le cadre de ce projet, basé à l'Université de Florence et impliquant plusieurs Universités italiennes et étrangères, nous avons réalisé différents outils qui sont aujourd'hui disponibles en accès libre sur notre site lessicobeniculturali.net pour plusieurs langues (allemand, anglais, chinois, français, espagnol, italien, portugais, russe) : des corpus comparables (Billero, Farina, Nicolas, 2020 ; Cetro et Zotti, 2020), des lexiques fondamentaux du patrimoine (Cetro et Zotti, à paraître), et, depuis peu, un nouveau volet qui porte sur la création de corpus parallèles de traduction (voir Zotti, 2017).

Ces différents outils répondent, selon nous, aux besoins actuels des industries de langues, en particulier pour ce qui concerne l'amélioration des outils de TA qui se nourrissent d'énormes quantités de données linguistiques et textuelles et comblent une des lacunes principales que nous avons pu remarquer dans les ressources terminologiques existantes : reliées, comme nous l'avons déjà dit, à des intérêts économiques précis, elles n'ont pas pris en compte certains domaines, comme la terminologie artistique ou la langue littéraire, qui restent pourtant fondamentaux dans les entreprises du tourisme et dans le monde de l'édition internationale.

Les propositions présentées dans ce numéro de *Synergies Italie* sont de différentes natures : les auteurs y exposent de manière réflexive leurs projets de recherche (Marzi, Bisiani), leurs expériences didactiques (Vezzani, Dankova) mais des approches plus théoriques avec des retombées appliquées sont également présentes (De Giovanni, Orlandi et Fasciolo). Les six contributions couvrent les domaines, très féconds dans les industries de la langue, de la traduction automatique, de la linguistique de corpus, de la terminographie, de la lexicographie, ainsi que de l'enseignement des langues et de l'expertise linguistique.

La première partie de ce numéro est consacrée à l'axe thématique « Industrie de la traduction ». Nous avons accueilli trois contributions qui se penchent sur les sujets suivants : les biais de genre dans les outils de traduction automatique neuronale (Marzi), le rôle des dispositifs terminologiques et de traduction automatique dans la diffusion et dans la mise en circulation de la terminologie institutionnelle du droit

pénal (Bisiani), et la conception et à la mise en œuvre d'un nouveau produit linguistique pour la construction de terminologies visant à soutenir les apprenants-traducteurs et les futurs professionnels des langues de spécialité (Vezzani).

La deuxième partie de ce numéro, intitulée « Industries de la langue et traitement automatique des langues (TAL) », est centrée sur deux autres volets des industries des langues qui ont un intérêt théorique et une répercussion pratique à la fois : l'apport des corpus numériques issus du Web pour la recherche terminologique dans le domaine du bien-être animal (De Giovanni), la création de corpus textuels spécialisés pour la traduction de la terminologie de l'urbanisme et l'aménagement du territoire à travers une expérience didactique documentée (Dankova) et, pour finir, l'application d'approches théoriques différentes mais complémentaires, l'une lexicographique, l'autre basée sur corpus, du traitement automatique de la langue (Fasciolo et Orlandi).

Dans la contribution qui ouvre ce numéro, Eleonora Marzi aborde le sujet de l'intelligence artificielle au service de la traduction automatique, un domaine clé de l'industrie de la traduction qui a connu un grand essor dans les dix dernières années. Marzi se penche sur l'analyse de trois outils de traduction automatique neuronale généralistes, disponibles gratuitement en ligne, *Google Translate*, *Microsoft Translator* et *DeepL*, non dans le simple but d'en évaluer et comparer les performances, mais pour éclairer la nature des stéréotypes présents dans les données d'entrées qui nourrissent ces logiciels, un problème majeur dont on a pris conscience **récemment**. De fait, comme le précise l'auteure, de nombreuses études supposent que l'existence du biais de genre dans la traduction automatique (TA) serait due à une biodiversité insuffisante des données d'entrées. Marzi s'intéresse aux erreurs de traduction en termes de genre qui sont générées par ces systèmes de TA, dans le but de comprendre si ces erreurs sont simplement aléatoires ou bien révélatrices de stéréotypes provenant d'un imaginaire sous-jacent. À partir d'une expérience basée sur un corpus de phrases ayant la même structure syntaxique, traduites en français et en italien dans les deux directions, Marzi dévoile la nature des stéréotypes présents dans les bases de données statistiques des logiciels de TA, en donnant l'exemple d'une liste de 73 noms de métiers déclinés au masculin et au féminin en français et en italien. À travers l'analyse d'un échantillon représentatif, elle confirme son hypothèse de départ, à savoir que la présence dans ces phrases de deux adjectifs appartenant à deux champs sémantiques différents, celui de l'apparence (*beau*/*belle*) et celui de la compétence (*intelligent*/*intelligente*), influence l'exactitude des traductions, le taux d'erreurs étant sensiblement plus élevé pour les noms au féminin, ce qui implique l'existence d'une représentation stéréotypée des genres présents dans la langue que les logiciels de TA reproduisent de manière systématique et incontrôlée.

L'Union européenne, qui a fait du multilinguisme l'un de ses principes fondateurs, a commencé à intervenir depuis une dizaine d'années pour faire face au problème de l'inégalité linguistique engendrée par l'intelligence artificielle. Plusieurs mesures ont été adoptées, comme la création du réseau d'excellence META (Alliance de la Technologie pour une Europe multilingue), le lancement du programme ELE (Égalité des Langues en Europe) qui vise à atteindre la complète égalité linguistique d'ici la fin de 2030 et, le 21 avril dernier, la publication d'une proposition pour la régulation de l'approche de l'Union Européenne face à l'intelligence artificielle.

Dans sa contribution, Francesca Bisiani aborde justement la complexité du multilinguisme au sein de l'Union européenne et nous invite à réfléchir sur les écarts interprétatifs qui se produisent lorsque les termes et le discours circulent dans l'espace européen. Elle se penche surtout sur les difficultés liées à l'aménagement linguistique et à la terminologie des concepts du droit européen dans les différentes langues-cultures. Dans sa recherche, l'auteure examine, selon une approche discursive de la terminologie et à travers une démarche à la fois quantitative et qualitative, les variantes dénominatives qui se manifestent dans les versions linguistiques en français, italien, anglais et espagnol des actes européens contraignants qui concernent le traitement des données personnelles en matière pénale (2016-2019). Elle se concentre sur les termes exprimant l'objet de la démarche juridique, c'est-à-dire le champ d'application de la décision-cadre et des directives concernées, en analysant la séquence « prévention + détection + enquêtes + poursuites ».

L'émergence de désalignements conceptuels dans les versions linguistiques des documents étudiés, notamment l'existence de deux séquences terminologiques différentes pour chaque langue, révèle que les décalages conceptuels qui ressortent de la traduction sont des symptômes de discordances idéologiques. La comparaison de ces résultats dans la base de données terminologique de l'UE (*IATE*) et dans deux outils de traduction automatique en libre accès (*DeepL* et *Google Translate*) confirme que les désalignements observés dans les traductions du corpus d'analyse se reproduisent aussi dans les dispositifs terminologiques et dans les outils de traduction automatique, ce qui confirme que ces instruments puisent dans les bases de données documentaires et terminologiques européennes. Ce constat amène l'auteure à souligner le rapport entre la terminologie, la traduction spécialisée et les outils d'aide à la traduction : la masse terminologique qui est employée par les traducteurs, les terminographes, mais aussi par les juristes linguistes et les experts au sein des institutions, alimente les logiciels de traduction automatique ou les concordanciers en ligne. Les termes utilisés par les traducteurs se propagent ainsi

non seulement dans les textes publiés, mais aussi dans les mémoires de traduction, ce qui a pour effet de fabriquer des segments prêts à être réemployés et d'amplifier ce qui a été dit par les instances énonciatrices. Pour cette raison, l'auteure insiste sur la nécessité de revenir de manière critique sur les choix dénominatifs effectués en amont, au moment de la production ou de la traduction de la terminologie institutionnelle juridique et politique.

C'est sur le besoin de développer des ressources plus riches et fiables pour la traduction spécialisée que se penche la contribution de Federica Vezzani. L'auteure propose une ressource, FAIRterm, disponible gratuitement en ligne, qui fournit une méthode pour la formation des futurs traducteurs technico-scientifiques. Ce produit a été conçu comme une base de données multilingue qui collecte des fiches terminologiques structurées et normalisées à des fins didactiques, de traduction spécialisée et de professionnalisation. L'auteure a réalisé un modèle de fiche qui permet aux apprenants-traducteurs de réfléchir au comportement morphosyntaxique, sémantique et phraséologique du terme technique analysé et du candidat terme équivalent. Cette ressource permet en outre d'importer les données structurées obtenues dans les différents systèmes de traduction assistée par ordinateur (TAO) permettant de bénéficier, pendant le processus de traduction spécialisée, des analyses menées pour chaque terme. L'auteure détaille une expérience en traduction spécialisée active dans le domaine œnologique, pour le couple de langues italien-français, qui a permis de valider la méthodologie didactique sous-jacente à l'utilisation de l'application FAIRterm. La perspective est de valider également la fonctionnalité de cette ressource dans d'autres domaines de spécialisation pour pouvoir la mettre à la disposition d'un public large, ce qui répond à un besoin croissant de la communauté scientifique de disposer de données de la recherche accessibles, interopérables et réutilisables.

La deuxième partie de ce numéro s'ouvre avec la contribution de Cosimo De Giovanni, qui se propose de démontrer l'apport des corpus numériques à la recherche terminologique par un cas d'étude. L'auteur analyse la circulation de deux termes, *abattoir mobile* pour le français et *unità mobile di macellazione* pour l'italien, dans deux différents corpus : un corpus parallèle, formé de textes émanant de l'Union européenne et repérés dans le site *EUR-Lex*, qui réglementent le domaine du bien-être animal (BEA), et un corpus composé des textes législatifs promulgués en France et en Italie, les deux analysés entre 1991 et 2009. En inscrivant son étude dans la perspective d'une terminologie communicationnelle, l'auteur montre que l'interprétation des contextes d'apparition des candidats termes se révèle fondamentale pour l'analyse de deux différents points de vue (PdV), un premier concernant les choix faits par le terminologue et le législateur, et un second qui

concerne les choix linguistiques opérés par un groupe de locuteurs sur la scène discursive. L'auteur démontre que l'utilisation du corpus numérique, à la fois du *corpus-web* (un corpus constitué à partir des données repérées sur le *Web* à l'aide de logiciels) et du *web-corpus* (un corpus constitué des données brutes extraites du *Web* à usages linguistiques), s'est avéré essentielle pour préciser la description des variantes dénominatives et conceptuelles des deux termes dans le domaine examiné et pour vérifier le comportement des termes en contexte. La comparaison des deux types de corpus numériques montre que le *web-corpus*, qui se prête bien à des analyses en termes quantitatifs et qualitatifs, pose des problèmes concernant la volatilité des données collectées et la rapidité avec laquelle elles peuvent varier. Il est évident, cependant, que les corpus numériques, sources d'enrichissement et d'intégration de données déjà collectées à partir de corpus traditionnels, peuvent être utiles pour la construction de nouvelles banques de terminologie.

Une constante dans toutes les contributions est la réflexion sur la nature et la qualité des sources employées par les outils technologiques et l'absence de grands corpus textuels informatisés pour certaines langues. Ces sources ne sont pas représentatives de certains domaines ni de certaines langues, et la qualité des données n'est pas toujours fiable, ce qui fait que les traducteurs et terminologues sont amenés à construire leurs propres corpus spécialisés. C'est justement la question abordée par Klara Dankova, qui propose un parcours didactique destiné à des étudiants italophones en traduction FR-IT de niveau de français avancé, en attirant l'attention sur les avantages que l'emploi de corpus peut apporter pour la traduction spécialisée. Des études récentes montrent en effet que, dans l'espace francophone, les corpus ne représentent pas encore des outils d'aide à la traduction utilisés fréquemment et de façon adéquate par les traducteurs professionnels (Loock, 2016 : 1-2). Le parcours proposé par l'auteure se base précisément sur la construction et l'exploitation, à l'aide de l'outil *Sketch Engine*, de corpus FR-IT comparables et parallèles qui concernent l'urbanisme et l'aménagement du territoire. Les sources de documentation sélectionnées couvrent la période 2015-2020 et contiennent de la sorte la terminologie de référence la plus actuelle possible de ce domaine. À l'heure actuelle, la taille modeste des corpus ne permet que des recherches limitées, mais la base textuelle est destinée à être enrichie dans l'avenir. Le parcours didactique, qui accorde une attention particulière à la traduction de la terminologie, est complété par une série d'exercices proposés à partir des besoins de traduction de textes concrets. Les étudiants et futurs traducteurs sont ainsi sensibilisés à l'importance de s'en servir dans la traduction de textes relevant d'un domaine spécialisé, tout en étant informés de leurs limites. La prudence s'impose notamment au niveau terminologique, car, comme l'auteure

le souligne, les termes utilisés dans les documents de l'UE ne sont pas nécessairement pertinents pour la traduction des textes rédigés dans un autre contexte.

La dernière contribution contenue dans ce numéro prolonge la réflexion sur l'apport du domaine de la linguistique de corpus mais en le comparant avec un autre domaine central dans les industries des langues, celui de la lexicographie, et cela avec une approche innovante qui combine la réflexion théorique à l'application de ces théories dans le cadre du traitement automatique des langues. Les deux auteurs, Adriana Orlandi et Marco Fasciolo, proposent de comparer la Théorie des Classes d'Objets (TCO), une approche lexicographique française conçue par Michel Mathieu-Colas et Gaston Gross comme un développement de l'idée de lexique-grammaire de Maurice Gross, avec la *Corpus Pattern Analysis* (CPA), une méthode d'analyse des patrons basée sur les corpus qui est à la base de la ressource linguistique T-Pas. En partant du constat que, dans les deux approches, les notions de *classes d'objets* et de *types sémantiques* d'une part, et les notions de *schéma prédicatif* et de *pattern* de l'autre semblent très proches, les auteurs démontrent qu'il existe des différences non marginales entre la *TCO* et la *CPA*. Ces différences ne concernent pas véritablement les résultats pratiques des deux approches, mais plutôt leurs présupposés théoriques et leurs implications méthodologiques. Les auteurs opposent *TCO* et *CPA* sur deux points spécifiques. Ils examinent tout d'abord la différence entre les notions de *schéma prédicatif* et de *pattern*, et observent qu'elles peuvent être apparentées à la dichotomie saussurienne *langue/parole*. En effet, la notion de *schéma prédicatif* relèverait du niveau de la *phrase-modèle*, alors que la notion de *pattern* relèverait du niveau de l'*énoncé*. Dans un second temps, les auteurs analysent une portion de la polysémie du verbe prédicatif *suivre/seguire*, notamment les acceptions de ce verbe en tant qu'« activité », et comparent les schémas prédicatifs du verbe *suivre* isolés suivant la *TCO*, et les *patterns* du verbe *seguire* dans la ressource T-Pas, isolés moyennant la *CPA*. Ils remarquent que les schémas prédicatifs sont plus nombreux que les T-Pas (*patterns*) et qu'ils sont souvent plus précis et plus spécifiques que ces derniers. Ils constatent ainsi que, sur le plan applicatif, la *CPA* pourrait tirer profit des travaux menés sur la TCO au sein du laboratoire LDI, la description des schémas prédicatifs étant plus fine dans la *TCO* que dans la *CPA*, ce qui augmenterait le pouvoir de désambiguïsation des *patterns*. Étant donné qu'au-delà des différences qui caractérisent les deux approches sur le plan théorique et méthodologique, la *CPA* et la *TCO* partagent une même visée applicative lexicographique, c'est-à-dire la réalisation de dictionnaires électroniques pour le traitement automatique du langage, les auteurs souhaitent en dernière instance que les approches et les méthodes existantes en matière d'analyse de corpus sortent des frontières nationales et entament un processus de

connaissance réciproque. L'adoption d'une approche « mixte » pourrait améliorer sensiblement les résultats obtenus et avoir des applications concrètes dans de nombreux domaines de l'industries des langues.

En conclusion, ce numéro de *Synergies Italie* lance une réflexion sur plusieurs aspects des industries des langues sous le signe de l'interdisciplinarité et de la rencontre entre l'informatique et la linguistique. Les études exploratoires présentées ici par des chercheurs du monde universitaire touchent des questions qui sont d'une grande actualité et qui vont à l'encontre des préoccupations des professionnels de l'ingénierie linguistique, des terminographes, des lexicographes, des experts de traduction automatique ainsi que des traducteurs et des enseignants en langues étrangères. Nous espérons que cette réflexion sera enrichie progressivement par leur expertise pratique en la matière.

Bibliographie

Association Encrages 1986. *Les industries de la langue : enjeux pour l'Europe : actes du colloque de Tours*. Paris : Université de Paris VIII-Vincennes.

Auger, P. 1989. « Informatique et terminologie : revue des technologies nouvelles ». In : *Actes du colloque terminologie et industries de la langue. Meta : journal des traducteurs*, n° 34 (3), p. 485-492.

Baker, M., Francis, G., Tognini-Bonelli, E. 1993, *Text & Technology*, Amsterdam/ Philadelphie : John Benjamins.

Bédard, C. 2000. « Mémoire de traduction cherche traducteur de phrases ». *Traduire*, n°186, p. 41-49.

Bernardini, S., Ferraresi, A. 2013. *I corpora nella didattica della traduzione: Corpus Use and Learning to Translate*. Bologne : CLUEB.

Billero, R., Farina, A., Nicolas, C. (éds.) 2020. *I corpora LBC. Informatica Umanistica per il Lessico dei Beni Culturali*. Florence : Firenze University Press.

Cetro, R., Zotti, V. 2020. « Les corpus et la base terminologique LBC. Des ressources pour la traduction du patrimoine artistique ». In : Mangeot, M., Tutin, A. *Lexique(s) et genre(s) textuel(s) : approches sur corpus. Actes de la conférence 11e Journées du réseau « Lexicologie, Terminologie, Traduction »*. Paris : Éditions des archives contemporaines, p. 81-98.

Cetro, R., Zotti, V. (à paraître). « Le corpus LBC français : bases, développement et applications ». https://www.lessicobeniculturali.net/

Cresti, E., Panunzi A. 2013. *Introduzione ai corpora dell'italiano*. Bologne : Il Mulino.

Dabbadie, M. 2004. « Industrie de la langue, vous êtes plutôt TIL ou TAL ? ». *VEILLE. Le magazine des professionnels de l'information stratégique*, n° 73.

Délégation Générale à la Langue Française et aux Langues de France (DGLFLF) 2015. *Mieux comprendre les outils d'aide à la traduction*.

Groupe de Réflexion sur les Industries de l'Information et les Industries de la Langue (GRIIIL) 2005. *Livre blanc. Le traitement automatique des langues dans les industries de l'information*.

Kubler, N. (éd.) 2011. *Language Corpora, Teaching, and Resources: from Theory to Practice*. Berne : Peter Lang.

L'Homme, M.-C., Jacquemin, Ch., Bourigault, D. 2001. *Recent Advances in Computational Terminology*. Amsterdam/Philadelphie : John Benjamins.

Loock, R. 2016. *La traductologie de corpus*. Villeneuve d'Ascq : Presses Universitaires du Septentrion.

Technolangue 2007. *Technologies de la langue en Europe : marché et tendances* réalisée par le Bureau Van Dijk, à la demande du Ministère de la recherche dans le cadre du programme Technolangue. http://www.technolangue.net

Zotti, V. 2017. « L'integrazione di corpora paralleli di traduzione alla descrizione lessicografica della lingua dell'arte: l'esempio delle traduzioni francesi delle Vite di Vasari ». In: Zotti, V., Pano Alaman, A. (éds.). *Informatica umanistica. Risorse e strumenti per lo studio del lessico dei beni culturali*. Florence : Firenze University Press, p. 105-134.