

Avant-propos

Ce numéro de *Studia Romanica Posnaniensia* intitulé « Recensement et description des données en vue des applications lexicographiques » est consacré au domaine en constante évolution. D'un côté, la lexicographie est une discipline ancienne, si les premiers glossaires visant à expliquer des termes de la Bible se développent au VI^e siècle, les plus vieux prototypes des dictionnaires bilingues sont des listes lexicales remontant au 24^e siècle av. J.-C., alors que des listes monolingues auraient été même plus anciennes¹. Une multitude de recueils monolingues et plurilingues ont vu le jour depuis, en contribuant au développement des recherches dont le but est de cataloguer, expliquer et structurer le lexique.

La lexicographie, ayant une longue tradition, est en même temps très moderne car surgissent de nouvelles conceptions des dictionnaires informatisés, ainsi que des possibilités efficaces de collecte des données. La question peut, toutefois, se poser si et comment les plus récentes avancées technologiques vont altérer ce domaine.

Longtemps, le développement de la lexicographie résultait des besoins très concrets liés à la fois à l'apprentissage des langues étrangères et à la traduction des documents. Remarquons qu'à l'heure actuelle, seulement à la Commission Européenne, la loi impose la traduction de plusieurs millions de pages de texte par an. Par chance, l'informatique facilite la tâche des linguistes en leur fournissant des outils de recherche automatique et des corpus électroniques. Dans ces nouvelles conditions, la lexicographie se présente comme une discipline ayant pour but de recenser, de décrire et de classer des mots selon les règles bien définies et rigoureuses, assurant le bon fonctionnement des systèmes informatiques.

Un facteur important qui contribue toujours à la mise en valeur des travaux lexicographiques est lié à la traduction automatique des textes. C'est le domaine des textes spécialisés qui semble particulièrement privilégié. Pour améliorer et accélérer ce processus, on crée des bases de données électroniques destinées au traitement

¹ Voir C. Boisson, P. Kirtchuk, H. Béjoint (1991). Aux origines de la lexicographie: les premiers dictionnaires monolingues et bilingues. *International Journal of Lexicography*, 4/4, 261-315.

informatique par les systèmes dépourvus d'intuition humaine. Par conséquent, comme description des significations on propose différents modèles et formats en fonction des besoins des utilisateurs des dictionnaires, qu'il s'agisse des humains ou des systèmes informatiques. Ces formalismes, étant divers et plus ou moins opérationnels dans le passé, reviennent renouvelés dans des ressources linguistiques construites pour surmonter tout type de difficultés, p. ex. : traitement des unités composées, recensement des mots des langues moins documentées, *etc.*

La lexicographie donne aussi accès aux langues moins connues, régionales ou minoritaires, et aide ainsi à préserver la diversité linguistique et culturelle du monde. Seulement en Europe, il y a 24 langues officielles et une soixantaine de langues régionales. Pour ces différentes langues étant toutes des trésors de l'humanité, il faut construire des dictionnaires de référence, traditionnels ou informatisés, en tant que moyen de les documenter, et sauvegarder ainsi le patrimoine linguistique et culturel, non seulement européen.

On trouvera dans ce volume un ensemble de contributions reflétant les directions mentionnées.

Le premier texte, proposé par Zaida Bartolomé-Díaz, présente une méthodologie pour la conception d'une ressource lexicale bilingue français-espagnol dédiée à l'architecture contemporaine. La proposition repose sur un modèle intégrant l'ensemble des activités lexicographiques : à partir de la construction du corpus à l'encodage final, en passant par l'extraction, l'alignement, la définition et l'enrichissement, tout en tirant parti des apports de l'intelligence artificielle. Les modèles statistiques et neuronaux sont utilisés pour automatiser des étapes traditionnellement longues et exigeantes, telles que l'identification des termes, le regroupement sémantique, la génération d'un projet de définition ou la détection de variantes. Cette automatisation reste encadrée par une validation humaine systématique, indispensable pour garantir l'exactitude conceptuelle et la cohérence terminologique.

Lian Chen propose une étude comparative du développement de la lexicographie informatique en France et en Chine, en analysant leurs trajectoires historiques, cadres institutionnels et orientations technologiques. En France, la discipline s'inscrit dans une tradition académique structurée autour de grandes infrastructures telles que l'ATILF, le TLFi, Frantext et le CNRTL, avec un fort accent sur les humanités numériques, la normalisation des données (XML/TEI, LLOD) et l'interopérabilité, notamment à travers des outils et modèles tels que NooJ et OntoLex-Lemon. En Chine, le développement, plus récent mais rapide depuis les années 1980, repose sur la constitution de corpus massifs, des bases sémantiques comme HowNet, et l'intégration croissante de l'intelligence artificielle et des grands modèles de langage dans la production lexicographique. La comparaison met en évidence deux approches complémentaires : une tradition patrimoniale et méthodologiquement rigoureuse en France, et une dynamique pragmatique et technologique en Chine, ouvrant la voie à une lexicographie numérique, multilingue et interconnectée.

L'article de Joanna Cholewa et Vita Valiukienė a le caractère confrontatif. Les auteures étudient les structures françaises à *arriver*, à savoir : N[abstract] *arrive* (à N[human]), N[human] *arrive* à Vinf et *Il arrive que* + Prop / *Il arrive* à N[human] *de* Vinf, tout en cherchant les équivalents les plus courants de ces usages en polonais et en lituanien. L'objectif de cette analyse est de démontrer que le verbe *arriver* est en cours de grammaticalisation et qu'il est nécessaire que la pratique lexicographique suive cette évolution. L'étude s'appuie sur un corpus parallèle trilingue français-polonais-lituanien. La fréquence d'usage des structures analysées démontre leur émergence aux côtés de l'usage locatif et illustre un écart par rapport aux dictionnaires bilingues existants.

La contribution d'Antonia Ciolac est de nature descriptive et a pour objectif de présenter trois dictionnaires inédits, actuellement en cours d'élaboration, qui diffèrent les uns des autres sur le plan thématique, mais qui ont pour point de référence commun la langue roumaine. Il est question d'un petit dictionnaire roumain-français québécois comportant des unités québécoises d'origine dialectale ou des archaïsmes, d'un dictionnaire des gallicismes romans et d'un dictionnaire des xénismes (daco-) roumains. Ces trois dictionnaires se distinguent des inventaires standards et normatifs par le fait qu'ils se concentrent avant tout sur des particularismes lexicaux. Ces particularismes sont pourtant bien attestés dans l'usage courant.

L'étude de Francis Grossmann et Anna Krzyżanowska s'inscrit dans un programme de recherche plus vaste en lexicographie visant à décrire des formules conversationnelles impliquant des fonctions pragmatiques. Les auteurs considèrent que la glose définitionnelle reste un élément essentiel du travail lexicographique et qu'une telle entreprise est particulièrement difficile lorsqu'on l'envisage d'un point de vue contrastif. À l'exemple des expressions françaises *je suis désolé(e)* / *désolé(e)* et de leur équivalent polonais *przepraszam*, les auteurs présentent la méthodologie de rédaction de gloses définitionnelles pour ce type de formule.

François Nemo et Binène Horchani proposent une étude sur les signes infralexicaux affixaux (du type *trag-* de *tragique*). Les auteurs accordent une place spéciale à ce type d'éléments, en considérant qu'ils constituent une part importante des constituants minimaux d'une langue. L'article présente les critères de définition, ainsi que la démarche conduisant à l'inventaire exhaustif de telles bases en français, en prenant comme exemple le traitement des adjectifs. Sont ensuite discutées les difficultés rencontrées à chaque étape, notamment en ce qui concerne la polymorphie des bases, et comment y faire face ; la question de l'intégration d'un dictionnaire de bases reliées en français dans les dictionnaires électroniques existants ; la question de la nature de la description sémantique que de telles bases peuvent recevoir, compte tenu de l'impossibilité de définir ce à quoi elles feraient référence ; et la question de savoir quelles informations chaque entrée doit fournir pour être complète. Il est montré notamment que l'élément central de la description lexicographique de ces bases est la description de leurs conditions d'utilisation, et donc des présupposés hérités des mots dans lesquels elles sont utilisées.

L'article de Verónica Del Valle Cacela se concentre sur les difficultés rencontrées par les traducteurs des textes juridiques, surtout quand il s'agit des langues moins utilisées que l'anglais. L'auteure fait une riche revue des ressources lexicographiques juridiques dont dispose aujourd'hui le traducteur espagnol-polonais et constate que, malgré les dernières évolutions consistant en création de nouveaux outils qui sont nombreux et variés, la situation ne s'est pas améliorée parce que ces outils, ne tenant pas compte du profil du traducteur, restent peu fonctionnels. C'est la raison pour laquelle les traducteurs sont souvent obligés de créer leurs propres ressources.

Daniel Vojtek présente une contribution ayant pour objet l'élaboration du corpus de terminologie grammaticale multilingue : slovaque, française, anglaise et latine, dans la perspective synchronique et diachronique. La méthode de l'élaboration du corpus se base sur trois points : le choix des ressources lexicographiques actuelles, le traitement multiparamétrique du matériel lexical au sein d'un fichier commun incluant les quatre langues traitées et l'application d'un seul critère équivalentiel qui est la correspondance conceptuelle découlant des caractéristiques définitionnelles de chaque terme. Il s'agit d'une tentative de systémiser l'ensemble de la terminologie grammaticale sur un plan quadrilingue.

Agnieszka Kaliska et Grażyna Vetulani

Poznań, le 21 septembre 2026