TABLE RONDE « RESSOURCES ET INGENIERIE LINGUISTIQUES »

Animateur : Jean-Pierre Colson (Université de Louvain, Belgique)

Le problème des ressources linguistiques se pose de manière cruciale dans diverses branches de la linguistique, et notamment en sociolinguistique, en phraséologie, en linguistique de corpus et en linguistique informatique.

A travers les diverses contributions de cette table ronde, plusieurs pistes de réflexion se dégagent. La collecte manuelle des données doit-elle être complétée par une approche automatisée ? Quels sont les avantages et inconvénients des deux méthodes ? La constitution de corpus pour étudier la variation linguistique et en particulier la variation phraséologique présente-t-elle des difficultés spécifiques ? La confrontation de méthodologies issues de la sociolinguistique, de la phraséologie et de la linguistique informatique permet-elle d'envisager un enrichissement mutuel ?

Addelhak Razky (Université de Brasília / Université Fédérale du Pará) tracera les différentes étapes de la constitution des ressources du projet Atlas Linguistique du Brésil (ALiB). Le projet ALiB voit le jour concrètement en 1996 au sein du Centre de dialectologie et de géographie linguistique dirigé par Suzana Cardoso et Jacyra Mota à l´Université Fédérale de Bahia. Un comité national se constitue durant le premier Workshop de L´ALiB intégrant, par la suite, d´autres universités brésiliennes ayant une expertise dans le domaine de la géographie linguistique. Une équipe, la famille ALiB représentant les auteurs d´Atlas linguistiques régionaux déjà publiés, met en place une méthodologie pour la collecte et l´analyse in vivo du corpus linguistique. Les ressources de l´ALiB proviennent de l´application de questionnaires linguistiques à 1.100 informateurs représentant 250 localités au long de 8.515.767 km de terres brésiliennes parcourues par l´ensemble des enquêteurs. Les ressources de L´ALiB sont ainsi organisées en fonction du volet phonétique et phonologique (QFF), lexico-sémantique (QSL), morphosyntaxique (QMS), questions de pragmatique, thématiques de discours semi-dirigé, questions de métalinguistique et un texte de lecture. Trois dimensions sont impliquées dans le traitement des ressources de l´ALiB : l´organisation générale du corpus (fichiers de transcription, et fichiers audio), la gestion des cartes linguistiques, la mise en place du système ALiBWeb, et les ressources du dictionnaire dialectal (Machado Filho 2015). Des 6 volumes de publications, prévus pour rendre compte de la grande partie du domaine linguistique, les deux premiers ont été publiés en 2014 (Cardoso et al. 2014).

Pierre-André Buvet (Université de Paris 13) discutera de la nature des ressources linguistiques dédiées à la compréhension automatique des textes. Elles sont de trois sortes : dictionnaire électronique, grammaire locale et ontologie. Un dictionnaire électronique spécifie des informations métalinguistiques sur des mots en rapport avec leurs morphologies, leurs syntaxe, leurs sens et leurs conditions d'emploi. Ces connaissances sont organisées à partir des formes des unités lexicales qu'elles décrivent. Une grammaire locale décrit le cotexte d'une unité lexicale donnée en tant qu'ensemble de configurations de mots. Une ontologie est une spécification plus ou moins explicite d'une conceptualisation. Sur le plan formel, les

ontologies diffèrent des dictionnaires par leur mode d'organisation du matériel lexical. Dans les dictionnaires, les mots, ou leurs lemmes lorsqu'ils ne sont pas invariables, constituent autant de points d'entrée ; ils sont les points d'ancrage des descriptions lexicographiques. Dans les ontologies, une grande partie des mots constituent les points d'arrivée ; les entrées sont les descriptions conceptuelles. Les ontologies sont des ressources linguistiques de nature pragmatique qui sont complémentaires des dictionnaires et des grammaires locales. Pierre-André Buvet discutera plus particulièrement des ressources linguistiques du point de vue des phraséologismes. Il évoquera également les corpus, dans la mesure où ils permettent d'enrichir automatiquement les dictionnaires et les grammaires locales lorsqu'ils sont correctement profilés.

Luis Meneses-Lerín (Université d'Artois) présentera le projet ECOS-Nord entre la France et le Mexique autour de l'axe « constitution de ressources linguistiques ». Deux procédés de traitement se dégagent : un traitement linguistique et un autre de nature informatique. D'une part, le traitement linguistique a pour objectif de procéder à la délimitation des phraséologismes, faire la collecte de ces derniers, établir une typologie basée sur la variation linguistique (les marqueurs typiques de chaque variété), proposer une typologie à partir de leurs caractéristiques syntactico-sémantiques et finalement procéder à la description lexicographique. Toutes ces opérations ont pour finalité pratique de retenir les éléments pertinents pour une reconnaissance automatique ou semi-automatique des séquences figées et pour la gestion informatisée des ressources. D'autre part, le traitement informatique a permis l'élaboration d'outils de collectes de données et de reconnaissance des séquences. Parmi les résultats escomptés, nous pouvons mentionner des bases de données textuelles, des ressources lexicographiques phraséologiques et des outils (programmes) informatiques appropriés. La constitution de ces ressources a obéi aux exigences suivantes : une taille significative, une représentativité assurée (notamment pour l'espagnol du Mexique) et une conception des dictionnaires bien élaborée. Au-delà de ressources, nous retrouvons des enjeux théoriques, applicatifs et en matière de formation à la recherche, ces enjeux seront abordés tout au long de cette intervention.

Jean-Pierre Colson (Université de Louvain) abordera la problématique des ressources linguistiques du point de vue de la phraséologie informatique et de la linguistique de corpus. Depuis de nombreuses années, un débat est ouvert entre les partisans de recherches basées sur les corpus (en anglais « corpus-based ») et dérivées des corpus (« corpus-driven »). Même si plusieurs auteurs soutiennent que ces deux manières d'utiliser les ressources linguistiques présentent des différences minimes, la question mérite d'être examinée plus en détails dans le cas de la variation phraséologique. Le choix des unités phraséologiques retenues pour mesurer le degré de variation diatopique peut en effet varier selon la méthodologie retenue : une approche basée sur les corpus partira d'unités souvent idiomatiques, extraites du dictionnaire ou de l'usage, dont l'on vérifie la fréquence ou les contextes d'utilisation dans des variantes d'une langue ou même dans des langues apparentées ; à l'inverse, une approche dérivée des corpus permet d'utiliser directement la sélection automatisée des unités phraséologiques au sens large, et mesure globalement le profil phraséologique des variantes ou langues concernées. Les deux méthodologies seront illustrées par des exemples. La constitution de

corpus ciblés pour étudier la variation phraséologique sera également abordée. Cette dernière nécessite en effet des corpus de taille importante, car la plupart des unités phraséologiques offrent moins d'une occurrence par million de mots (tokens). Cette constatation mène naturellement à la création de corpus issus de la Toile, à partir d'outils qui ont aujourd'hui fait leurs preuves. Le traitement de telles données par des algorithmes pose toutefois des questions pratiques mais aussi théoriques.