L'apport du numérique pour la linguistique de corpus

Stéphane Robert

Stéphane Robert

L'apport du numérique pour la linguistique de corpus Contribution du

Réponse du Comité de Pilotage de l'IRCOM pour le dossier thématique consacré aux consortiums de la TGIR préparé par Huma-Num
[16 octobre 2013]

(dans notre texte, nous avons gardé les questions suggérées lors de la demande)

En quoi le numérique permet-il de renouveler vos objets de recherche, suscite de nouvelles questions et hypothèses, transforme vos méthodologies ?

A la différence de disciplines où l'ordinateur est apparu comme un outil au service de pratiques établies dans un environnement peu technique, dans les sciences du langage, l'utilisation du numérique a prolongé des techniques d'enregistrement de corpus oraux et des protocoles expérimentaux qui avaient déjà intégré, dans leurs pratiques et leur raisonnement, l'usage d'appareils. C'est le cas notamment pour la linguistique de terrain, qui traite le plus souvent de langues à tradition orale, mais également, dans le domaine du traitement de la parole, pour l'analyse du signal appliquée à des phénomènes discrets de bas niveau (phonologie ou phonétique) ou à des phénomènes suprasegmentaux plus complexes (étude de la prosodie par exemple).

Mais si l'introduction du numérique dans ces pratiques a été faite dans la continuité de certaines pratiques antérieures, elle a néanmoins permis un saut qualitatif remarquable pour ces sous-disciplines, grâce à la qualité des enregistrements, à la maniabilité des appareils (petits enregistreurs numériques *vs* anciens appareils à bandes) et surtout en raison des possibilités considérables offertes par le numérique pour l'exploitation informatique des données recueillies : il est, par exemple, désormais possible, à partir d'un enregistrement audio-visuel, de combiner le puissant logiciel ELAN qui permet d'aligner l'image, le son, la transcription du texte et ses différentes annotations, avec PRAAT pour l'analyse phonétique et TOOLBOX pour l'interlinéarisation des gloses et la constitution automatique de lexiques, ce qui permet un large éventail d'analyses linguistiques sur les différentes composantes du langage. La mise à disposition de corpus transcrits, traduits et glosés, incluant son et image, a ainsi fait émerger une nouvelle organisation du travail et la constitution de communautés de chercheurs en réseau. Dans le même temps, la nature même des ressources et leur mode de présentation ont décidé de méthodes de travail qui privilégient le retour sur les interrogations et la pluralité des requêtes à une échelle autrefois inaccessible.

La transformation a été encore plus radicale dans la constitution et l'exploitation de masses de données qui sont venues confirmer ou infirmer les hypothèses constituées sur des échantillons limités. L'augmentation de la taille des corpus a donné une dimension nouvelle à la statistique linguistique, présente dès les années 1950, qu'il s'agisse de raisonner sur l'évolution des langues ou encore de recourir à des calculs de fréquences pour simuler des états de langue ou encore pour établir la réalité d'un phénomène ou d'une construction dans une langue en synchronie.

Dans ce développement, en même temps que les méthodes de travail des linguistes changeaient et que leurs études trouvaient des applications d'usage courant (reconnaissance vocale, correcteurs), de nouveaux questionnements ont surgi avec des formes inédites de communication rapide (Chat, SMS) qui modifient les usages langagiers. Ces (r)évolutions technologiques et méthodologiques se sont produites relativement rapidement et supposent à la fois un effort considérable de formation

de la part des linguistes pour s'approprier les différents outils liés au numérique et une réflexion scientifique sur l'utilisation raisonnée de ces nouveaux moyens d'investigation. C'est pour organiser et accompagner la réflexion et la formation de la communauté des producteurs et utilisateurs de corpus oraux et multimodaux qu'a été créé le Consortium IRCOM qui n'est pas une nouvelle structure disposant de moyens spécifiques mais un consortium réunissant l'ensemble des chercheurs et ingénieurs impliqués dans ces travaux. L'IRCOM a ainsi mis en place un site web (http://ircom.corpus-ir.fr) qui présente les différentes activités du consortium (formations, groupes de travail sur divers aspects concernant les corpus), ainsi qu'un recensement des corpus oraux existants ou en cours de constitution en France, un volet sur les ressources existantes (outils, logiciels, normes et formats) et un important glossaire concernant les principaux termes ou notions utilisés à propos des corpus oraux et multimodaux et destiné à aider les linguistes.

Dans quelle mesure, le numérique favorise l'émergence de questions pluridisciplinaires ?

Le numérique a redessiné les frontières des disciplines en confortant, dans les sciences du langage, la part de l'oral et désormais du multimodal (gestes, postures, mimiques), qui avait longtemps été freinée par les difficultés techniques que posait l'étude de matériaux non scripturaux. Les questions spécifiques liées aux corpus multimodaux font d'ailleurs l'objet de réflexions et de propositions en particulier par les formations proposées par le groupe 'Multimodalité et modalité gestuo-visuelle' du consortium IRCOM. Par ailleurs, l'étude de la production orale bénéficie désormais également de l'utilisation des techniques d'imagerie médicale (e.g. échographie) et les collaborations avec les neurosciences se sont développées dans le cadre d'expériences d'imagerie par résonance magnétique fonctionnelle (IRMf) ou d'électro-encéphalographie (EEG). Les ouvertures vers la physique (traitement du signal), voire la biomécanique, ont également gagné en abstraction et en généralité grâce à l'adaptation et à l'enregistrement des données et des métadonnées (voir *métadonnées* dans le glossaire, http://ircom.corpus-ir.fr/site/glossaire.php#metadonnees) dans des formats configurés par l'informatique. Il en a résulté une autre relation aux mathématiques où, en plus des statistiques, les linguistes peuvent aussi avoir recours à la théorie des graphes, à la modélisation… Dans le spectre des études linguistiques, c'est la linguistique de corpus qui a souvent servi de banc d'essai à ces interactions.

Parallèlement, il existe d'autres exploitations écrites (correction, traduction) dont il est seulement fait état pour mémoire dans ce document du fait qu'elles se situent en dehors de l'aire de compétence de l'IRCOM.

Selon quels modes et quelles voies les chercheurs de votre/vos discipline(s) se l'approprient ?

Ces changements de méthodes et de pratiques liés aux potentialités du numérique ont engendré de nouvelles contraintes pour les linguistes qui s'y intéressent. L'appropriation de ces outils est d'autant moins évidente que leur standardisation n'est pas encore achevée. Si, encore aujourd'hui, une majorité de chercheurs semblent avoir conservé pour centre d'intérêt des thématiques déjà éprouvées alors même qu'ils accèdent à certaines potentialités offertes par le numérique, on peut distinguer, au-delà d'effets générationnels marqués, (i) ceux qui ont entrepris l'aggiornamento de leurs méthodes de travail à partir de l'utilisation d'un ou plusieurs outils dont ils ont trouvé l'application sur leur objet d'étude et (ii) ceux qui ont infléchi leur approche à partir d'une réflexion sur les données, un cas particulièrement sensible en linguistique de corpus.

Les deux démarches sont souvent conjointes et, dans l'exploitation des données, les exigences concernant l'interopérabilité (voir glossaire, http://ircom.corpus-ir.fr/site/glossaire.php#i), l'indexation ou la pérennisation contribuent à un retour d'expérience qui est une dimension nouvelle des études linguistiques. La versatilité des logiciels et leur adaptabilité ont encouragé une certaine standardisation qui contribue à l'émergence d'un langage de représentation et de formalisation commun ; cette standardisation est cependant encore loin d'être accomplie et l'IRCOM a un rôle important à jouer pour aider la communauté à s'organiser autour de bonnes pratiques et de standards partagés, dans un contexte qui reste encore foisonnant. En principe, cependant, les mêmes formats de transcription (cf. http://ircom.corpus- ir.fr/site/glossaire.php#t) peuvent être exploités par des linguistes de terrain et par des psycholinguistes, par des phonéticiens et par des dialectologues. Le transfert de compétences et les échanges sur les moyens numériques à disposition (voir le débat sur la TEI, cf *Text Encoding Initiative*, http://ircom.corpus-ir.fr/site/glossaire.php#TEI) dessinent de nouveaux réseaux d'interaction transversaux aux thématiques et aux périmètres régionaux. A partir des usages et des besoins qui émanent des laboratoires, au sein du consortium IRCOM, un groupe de travail consacre ses travaux à améliorer l'interopérabilité entre les logiciels d'annotation et les données en respectant les normes et les standards internationaux (voir http://ircom.corpus- ir.fr/site/p.php?p=groupetravail2). Un autre groupe de travail est en cours de constitution afin de pouvoir présenter aux producteurs de corpus les divers circuits existant pour l'archivage pérenne, dans un paysage qui reste encore morcelé et peu lisible pour les utilisateurs.

En quoi la réutilisabilité des données de recherche peut revêtir une importance particulière dans votre/vos domaine(s) ?

Le recyclage des données par des programmes pour lesquels ils n'avaient pas été initialement conçus est une source importante d'économie. Alors que les corpus oraux et multimodaux représentent un coût élevé avant que ne soit obtenue une ressource fiable et quantitativement significative (temps d'enquête ou d'expérimentation, transcriptions, vérifications, annotations…), leur reprise, sous condition de contrôle et de validation assure un gain de temps et d'argent considérable, une part de travail invisible dans ce que livre un chercheur s'il n'est pas l'auteur de son corpus. Au demeurant l'archivage des langues non écrites ou non encore scriptibles telles que les langues signées constitue à ce jour le seul moyen de conserver un état de langue à valeur patrimoniale qui pourra être réinterrogé.

L'existence virtuelle d'un réexamen des ressources constitue une contrainte bénéfique sur la production des résultats en ce que la contrôlabilité représente la meilleure garantie de la véracité. Le principal résultat, familier aux sciences, est la cumulativité non seulement des acquis du savoir mais également des matériaux qui les certifient. La masse d'informations électroniques disponibles se rapproche de celles qui sont accessibles aux locuteurs, du moins pour les langues de grande diffusion qui bénéficient de larges corpus, l'écart entre la quantité de données archivées et celles dans lesquelles est immergé un auditeur se résorbant progressivement.

Comment ces nouveaux modes de production transforment vos pratiques de recherche ?

Les résultats scientifiques sont soumis à une condition de non contradiction avec l'expérience. Celle-ci a longtemps été confinée aux connaissances accumulées par un ou plusieurs experts, voire – dans un paradigme qui l'avait érigé en dogme – à l'intuition du locuteur natif. Le

traitement des données, leur accumulation comme leur mode d'analyse par des procédures automatisées, ont rendu au concept de démonstration sa pleine valeur, plus encore au concept de reproductibilité de l'expérience. Désormais l'accès aux corpus comprenant les données primaires et les annotations favorisera l'échange et la discussion scientifique entre chercheurs en permettant la vérifiabilité des analyses. Cet accès croissant des corpus oraux et multimodaux occupe une place grandissante dans l'élaboration et la validation des hypothèses linguistiques, posant sous un nouveau jour des questions méthodologiques et scientifiques. En outre, la finalité d'un corpus est évidemment interrogée lorsqu'il est mis à la disposition d'une communauté de chercheurs et, partant, la linguistique de corpus s'en trouve modifiée. Ces questionnements sont discutés dans le consortium par un groupe de linguistes qui s'attachent à mettre en évidence les nouvelles finalités scientifiques que posent ces modes de production et de diffusion des corpus (cf. http://ircom.corpus-ir.fr/site/p.php?p=groupetravail1). Les modifications de nos pratiques de recherche liées à l'utilisation du numérique, déjà sensibles, sont clairement amenées à se développer encore plus largement.