Le transcripteur transcrit : retour d'expérience à partir du corpus des ESLO

Linda Hriba, Olivier Baude, Céline Dugua

Le transcripteur transcrit : retour d'expérience à partir du corpus des ESLO

Hriba, Linda[1], Baude, Olivier[2] & Dugua, Céline[1]

[1] LLL UMR7270, Université d'Orléans ; [2] MoDyCo UMR7114, Université Paris Nanterre

Hriba, Linda

LLL UMR7270, Université d'Orléans linda.hriba@yahoo.fr

1 Les enquêtes sociolinguistiques à Orléans : un très grand corpus variationniste

Les Enquêtes sociolinguistiques à Orléans (ESLO) forment un grand corpus oral constitué de deux enquêtes réalisées à deux périodes distinctes. La première enquête ESLO1 (1968-1971) est un corpus clos de 470 enregistrements, soit 318 heures d'enregistrements qui représente – selon l'estimation de l'époque – 4,5 millions de mots. La seconde enquête (ESLO2), commencée au début des années 2000 et toujours en cours de réalisation, affiche un objectif de plus de six millions de mots pour 450 heures d'enregistrements.

ESLO ne constitue pas seulement un corpus de masse de données, il s'agit d'un réservoir de corpus conçu dans un souci de représentativité des pratiques linguistiques d'une communauté d'auditeurs dans une ville donnée et à des moments distincts. La prise en compte de la variation, et de toutes les variations est au cœur du projet et guide à la fois les choix méthodologiques qui ont été réalisés dès les premières étapes de la constitution du corpus, les regards que nous porterons sur les analyses, et également la question de la transcription.

2 Procédure de transcription des ESLO

Depuis 2003, le LLL (Orléans) s'est donné pour objectif de transcrire et rendre disponible l'intégralité du corpus ESLO. Face à l'ampleur de la tâche et soucieux de rendre rapidement accessible le corpus, la transcription repose sur des conventions minimales. Il s'agit de répondre à un simple objectif de navigation dans le corpus. Ces premières contraintes nous ont, par ailleurs, orientés vers le logiciel de transcription *Transcriber* qui permet de réaliser les alignements/synchronisations son-transcription très facilement et qui, grâce à son interface simple, constitue le meilleur outil pour réaliser des transcriptions « au kilomètre ».

Afin de définir nos conventions de transcription, nous avions entrepris une comparaison des pratiques au sein de grands projets (CLAPI, DELIC etc.) travaillant sur l'oral, ce qui avait permis la mise en évidence des principes généralement partagés par tous ces projets, principes que nous avons adoptés pour ESLO. Ces derniers reposent sur une transcription orthographique standard qui rend compte des phénomènes spécifiques de l'oral (répétitions, amorces etc.), avec une segmentation en tours de parole (les choix opérés sont disponibles dans le « Guide du transcripteur » sur le site ESLO). Ainsi pour toute analyse ultérieure, une reprise de la transcription avec des conventions répondant aux cadres théoriques du chercheur et/ou des niveaux d'annotations sont indispensables.

Le LLL a également porté une attention particulière aux travaux qui, depuis 1970, ont montré la nécessité de relire à plusieurs reprises les transcriptions Fillol et Mouchon (1977). La relecture n'est efficace que si elle est réalisée par une autre personne que le transcripteur (Lahire, 1981). En partant de ce constat, le LLL a donc décidé de recourir à trois « écouteurs » (Blanche-Benveniste & Jeanjean, 1987) distincts pour la transcription des ESLO. La transcription se fait alors en trois étapes successives :

- une version A (VA) qui correspond à une première version « brute » de transcription, la priorité est donnée à la synchronisation de la transcription avec l'enregistrement,
- une version B (VB) dans laquelle il s'agit de vérifier l'orthographe et le respect des conventions de transcription de la version A,
- et une version C (VC) qui est la relecture de la version B.

On obtient ainsi trois versions de transcription pour un même enregistrement avec trois transcripteurs différents.

Ainsi, la transcription n'est plus conçue uniquement comme le préalable à une étude sur corpus oraux, elle est une façon de mettre en perspective les conditions de productions des données. En ce sens, elle constitue une étape qui reflète le champ de la linguistique, les théories, l'inscription du chercheur dans son domaine, et également les « attitudes » et les représentations des transcripteurs.

3 Les transcriptions : un observatoire des variations

Cette procédure, qui consiste à disposer pour chaque enregistrement de trois versions de transcription, nous a permis de relever d'importantes divergences entre ces étapes. En nous appuyant sur un corpus constitué de 20 enregistrements (60 fichiers de transcription), nous avons, à l'aide d'un logiciel spécialisé *Beyond Compare*, comparé les différentes versions obtenues en les confrontant deux à deux (VA *vs* VB et VB *vs* VC). L'extraction et l'analyse des différences ont révélé l'impossible stabilisation d'une version définitive de transcription. Il s'avère qu'en moyenne, 330 interventions sont nécessaires pour passer d'une VA à une VC ; ces dernières concernent trois grandes catégories de variations :

- des variations graphiques. Elles regroupent l'ensemble des erreurs qui, d'une part, correspondent aux fautes induites par les outils de saisie (clavier) et d'aide à la transcription (*Transcriber*) et qui, d'autre part, sont le reflet d'un non-respect d'une norme orthographique ou de codage (cf. conventions de transcription propres au projet) ;
- des variations de segmentation. Elles concernent l'alignement temporel, la segmentation en sections (ils s'agit de types de questions ou de thématiques), en tours de parole ainsi que les pauses ;
- et des variations de perception manifestant des divergences d'écoute.

Différents paramètres acoustiques, linguistiques et sociologiques permettent d'expliquer ces variations de perception, dont en premier lieu les caractéristiques propres des transcripteurs. Sur ce dernier point, nous manquons d'informations sur nos transcripteurs et de données qui permettraient de mieux comprendre qui ils sont et quel est leur rapport à l'écrit.

4 Un nouveau module ESLO : entretiens avec les transcripteurs

Ce nouveau module, en cours de construction, consiste en des entretiens semi-directifs auprès des transcripteurs avec pour objectif de capter leurs représentations de la langue, de les situer au sein de pratiques sociales et de confronter ces éléments aux variations de transcription relevées dans le corpus. Ces entretiens s'organiseront en cinq grandes thématiques.

- la trajectoire des transcripteurs : depuis la période scolaire, puis universitaire jusqu'à leur activité professionnelle actuelle ;
- l'expérience de transcription : l'objectif sera de recueillir leurs ressentis sur cette activité, de faire émerger ce qui était le plus difficile et pénible mais aussi le plus plaisant ;
- des questions autour de leur rapport à la norme et à l'écrit, notamment à l'orthographe. Il s'agira par exemple de leur demander comment ils abordent l'écrit dans des nouveaux moyens de communication ;
- une thématique sur leur rapport à la lecture aujourd'hui, mais aussi durant leur apprentissage, au collège/lycée et à l'Université ;
- et enfin, des questions sur leurs pratiques d'écriture, et leurs pratiques culturelles.

L'analyse de ces entretiens se réalisera à trois niveaux. Nous souhaitons établir une échelle qui prenne en compte ces différents paramètres et qui éventuellement les pondère afin de proposer, pour chaque transcripteur, une catégorisation de ses pratiques et représentations. Nous réaliserons ensuite une analyse du contenu des entretiens et enfin nous utiliserons ces éléments dans l'analyse des variations de transcription de chacun des transcripteurs.

Cette approche réflexive de la phase de transcription du corpus des ESLO souhaite dépasser une simple description de la méthodologie de linguistique de corpus oraux pour atteindre une analyse linguistique fondée sur des données attestées et situées. La transcription d'un très grand corpus oral, pour peu que celui-ci soit documenté par suffisamment d'éléments permettant de situer socialement la parole captée, offre incontestablement un point de vue privilégié sur le passage de la perception acoustique à l'empreinte linguistique.

Références bibliographiques

Barras C., Adda, G., Adda-Decker, M., Habert, B., Boula de Mareüil, P, Paroubek, P (2004). Automatic audio and manual transcripts alignments, time-code transfer and selection of exact transcripts. *Actes de la Fourth International Conference on Language Resources and Evaluation (LREC)*, Lisboa, May 2004, vol. 3, pp 877-880.

Baude, O. Dugua C. (2011). (Re)faire le corpus d'Orléans quanrante ans après : quoi de neuf, linguiste ? *Corpus*, *10* Varia, 99-118.

Baude, O. (2006). *Corpus oraux, Guide des bonnes pratiques*. Paris, CNRS Editions.

Bilger, M. (ed) (2008). *Données orales, les enjeux de la transcription*. Les cahiers de l'Université de Perpignan.

Blanche-Benveniste, C., Jeanjean, C. (1987). *Le français parlé. Transcription et édition*. Paris, Inalf, Didier érudition.

Cappeau P., Gadet F. (2013). Quand l'œil écoute : que donnent à lire les transcriptions d'oral ? *Communication orale au CILPR*, Nancy.

Cappeau, P., Gadet, F., Guerin, E., Paternostro, R. (2011) « Les incidences de quelques aspects de la transcription outillée », *Linx* [En ligne], *64-65*, 85-100. Mis en ligne le 01 juillet 2014, consulté le 17 janvier 2017. URL : http://linx.revues.org/1403 ; DOI : 10.4000/linx.1403

Cappeau, P., Gadet, F., (2010). Transcrire, ponctuer, découper l'oral : bien plus que de simples choix techniques. *Cahiers de linguistique, 35/1*, 187-202.

Delais-Roussarie, E. & Yoon, H.-Y. (2011). Transcrire la prosodie : un préalable à l'échange et à l'analyse des données. *Journal of French Language Studies, 21*, 13-37.

Encrevé, P. (1977). Présentation: linguistique et sociolinguistique. *Langue Française, 34*, 3-16.

Falbo, C. (2005). La transcription : une tâche paradoxale. *The Interpreters' Newsletter, 13*. 25-38.

Fillol, F. & Mouchon, J. (1977). « Alors cet événement s'est passé » - Les éléments organisateurs du récit oral. *Pratiques*, *17*, 100-127.

Habert, B. (2005). Portrait de linguiste(s) à l'instrument. *Texto!* [en ligne], *vol. X*, n°4.

Koch, P. & Oesterreicher, W. (2001). Langage oral et langage écrit. *Lexicon der romanistischen Linguistik, 1-2*. Tübingen : Max Niemeyer Verlag, 584-627.

Mondada, L. (2000). Les effets théoriques des pratiques de transcription, *LINX*, *42*, 131-146.

Ochs, E. (1979). Transcription as theory, in *Developmental pragmatics*, ed. by E. Ochs & B. Schieffelin. New york: Academic Press, 43-72.

Corpus eslo : http://eslo.huma-num.fr/