Corpus de langue des signes : premières réflexions sur leur conception et leur représentativité

Jérémie Segouat, Annelies Braffort, Annick Choisier

Travaux linguistiques du Cerlico « L'exemple et le corpus, quel statut ? » N° 23, PUR

Corpus de langue des signes : premières réflexions sur leur conception et leur représentativité

Jérémie Segouat, Annelies Braffort et Annick Choisier
LIMSI-CNRS/WebSourd, LIMSI-CNRS, LIMSI-CNRS

Résumé

Cet article propose une réflexion autour de la notion de corpus de langue des signes, de la méthodologie de constitution de ces corpus et de la notion de représentativité dans ces corpus. En effet, le corpus est utilisé depuis peu pour la recherche sur les langues des signes, en linguistique comme en informatique. Cette réflexion est menée par comparaison avec le domaine des langues vocales, où ces notions sont étudiées pour les corpus oraux et écrits. Les différences et les points communs constatés nous permettent d'identifier les pistes à approfondir, tant sur les définitions que sur les méthodologies.

Introduction

Nos recherches se situent dans le contexte du traitement automatique des langues des signes (TALS), domaine à la frontière de la linguistique et de l'informatique. Nous effectuons des analyses de la langue en nous appuyant sur des études linguistiques, pour parvenir à mettre en place des modèles pour le TAL et à les implémenter et les évaluer au sein de diverses applications informatiques. Nos analyses sont faites sur des corpus vidéo de langue des signes (LS), qui constituent ainsi le socle des études que nous menons.

Il y a de plus en plus de recherches dans le domaine des LS qui s'appuient sur des corpus. Mais qu'est-ce que qu'un corpus en LS ? Cette notion est-elle empruntée aux langues vocales (LV) ou y a-t-il des spécificités ? Les méthodes pour les constituer suivent-elles les mêmes principes que celles des corpus de LV ? Les réponses à ces questions ont un impact sur les traitements qui seront réalisés sur les corpus, leurs analyses, etc. : est-il possible de suivre les mêmes principes que pour les LV ou faut-il en envisager de nouveaux ?

Nous posons ici les bases d'une réflexion sur la notion de corpus en LS, ainsi que sur les méthodologies existantes pour construire ces corpus. Nous nous proposons d'établir une comparaison avec les LV afin de déterminer les aspects communs de ces notions, et les spécificités. Nous étudions de manière plus particulière la notion de représentativité, et présentons la méthodologie qui nous permet de contrôler précisément cet aspect dans le domaine des corpus de LS.

Cet article se structure comme suit : dans la première section nous introduisons la langue des signes et les théories linguistiques associées ; en section 2 nous établissons un parallèle entre les LS et les LV sur les méthodologies de constitution des corpus ; en section 3 nous mettons à jour quelques spécificités pour la constitution de corpus de LS ; en section 4, nous nous focalisons sur la notion de « représentativité » et nous explicitons la méthodologie que nous avons mise en place pour la garantir ; enfin, nous présentons une application de notre méthodologie pour la constitution d'un corpus, en section 5.

I. La langue des signes et les théories linguistiques associées

La langue des signes est une langue naturelle utilisant une modalité visuo-gestuelle. Elle s'exprime à l'aide de gestes manuels et non manuels, incluant le regard, les expressions du visage, les mouvements du haut du corps, des bras et de la tête. Elle se réalise dans un espace à trois dimensions situé devant le locuteur (appelé « signeur »), dans un « espace de signation » illustré (figure 1). Le signeur réalise les unités gestuelles constituant les énoncés, encore appelés « signes » dans l'espace : il « positionne » des entités de son discours, pour ensuite y faire référence (par un geste déictique de pointage, un regard, un mouvement d'épaule…) et les mettre en relation.

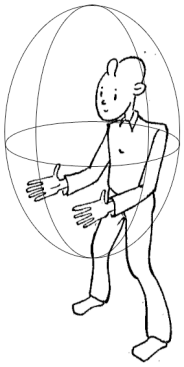

Figure 1 : représentation schématique de l'espace de signation (Guitteny 2004).

Les LS ont longtemps été vues comme étant un agrégat de pantomimes sans aucune structure linguistique propre. Stokoe (1960) a le premier démontré que la langue des signes américaine (ASL) possédait un système linguistique complet, avec son lexique et sa grammaire.

En France, Cuxac (2000) considère que l'iconicité est intrinsèque à la langue des signes et qu'elle est le principe fondateur à toute description. Pour lui, l'iconicité prend place dans le « processus d'iconicisation » par lequel le signeur va rendre iconique son expérience vécue ou imaginée. Il considère que ce processus aboutit à l'utilisation d'une « visée illustrative », qui consiste à « dire en montrant » et qui n'exclut pas une « visée non-illustrative », qui consiste à dire sans montrer. Lors de l'expression d'un énoncé, il y a une combinaison, alternée ou simultanée, des visées illustrative et non-illustrative. Le discours à visée non-illustrative correspond à l'utilisation des signes lexicalisés (signes du vocabulaire de la langue), des pointages et de la dactylologie (code gestuel représentant l'alphabet écrit, du français dans le cas de la LSF). Le discours à visée illustrative est exprimé grâce à des « structures de grande iconicité ». Ces structures combinent trois formes principales : les transferts de taille et de forme (pour décrire plus en détail des entités), les transferts de situation (pour exprimer des relations spatiales entre entités ou déplacements) et les transferts personnels (le signeur prend le rôle d'une des entités du discours) (Sallandre 2003). Cuxac définit les transferts comme des opérations cognitives qui permettent de transférer des expériences réelles ou imaginaires dans l'espace de signation.

De nos jours, il existe deux courants de pensée majeurs quant à la structure linguistique des LS : un qui considère que l'iconicité n'appartient pas à la langue, et un autre qui considère que l'iconicité en fait partie. Nous positionnons nos études dans la perspective des travaux de Cuxac et de son équipe. Cette base de départ n'est pas sans conséquences : le positionnement scientifique a une influence importante sur la constitution du corpus. Ainsi, nous considérons

l'iconicité comme principe organisateur de la langue, donc les corpus que nous étudions se doivent d'intégrer ces structures linguistiques nommées « transferts » par Cuxac.

II. Les méthodologies de constitution de corpus

Un corpus au sens général est un ensemble de données, qui peut se présenter sous diverses formes, sélectionnées selon des critères précis définis suivant l'objectif de l'étude. Pour Thoutenhoofd (2007), « un corpus de langue des signes est un ensemble multimédia contenant des vidéos de Sourds signants, ainsi que des annotations qui facilitent les recherches sur les données »[1]. Les corpus de LS sont créés avec des objectifs de conservation patrimoniale, ou des objectifs de recherche linguistique, sociologique et, depuis peu, informatique.

Il est possible de définir des méthodologies générales de constitution de corpus au niveau de la collecte des données, des aspects techniques, des élicitations, etc. Cependant, il n'existe pas de méthodologie permettant de s'assurer des propriétés (comme la représentativité, cf. infra) que doit vérifier le contenu d'un corpus. Notre objectif dans cette partie est de mettre à jour des grands principes de critères à respecter lors de la création d'un corpus en langue vocale, orale ou écrite, et d'analyser ce qui peut être mis en parallèle dans le domaine des LS.

a) Méthodologie de constitution de corpus de textes

Dans la littérature, il est fait état de différents critères à respecter afin de constituer un ensemble de données qui puisse être dénommé corpus. Ainsi, dans le domaine des corpus de texte, Bommier-Pincemin (1999), pose le principe que c'est le traitement sur le corpus qui va guider la conception du corpus. Il y a donc constitution d'un ensemble de textes suivant des critères de rapport entre quantité et qualité, de contraintes techniques (de formats ou de mémoire, entre autres), et d'adaptabilité aux traitements ultérieurs. Benzécri (1981) met en avant que bien qu'en théorie il faille établir une sélection des constituants du corpus, en pratique il s'agit de se contenter de ce qui est disponible. Bommier-Pincemin (1999) revient par ailleurs sur les trois types de conditions que doit vérifier un corpus pour être nommé comme tel : « des conditions de signifiance, des conditions d'acceptabilité, et des conditions d'exploitabilité ».

O'Donnell (2008) préconise de plus, vu les contraintes à respecter et les conditions à vérifier, qu'il faut réfléchir en trois étapes lorsqu'il est question de construire un corpus. La première consiste à réutiliser l'existant, voire à combiner plusieurs corpus ou à n'en retenir qu'une partie. Si cela ne convient pas, il faut essayer de redéfinir l'existant suivant l'étude qu'on souhaite poursuivre. Enfin, si cela ne convient toujours pas, il faut construire son propre corpus. Cette réflexion met l'accent sur le fait que dans le domaine textuel il ne faut pas hésiter à utiliser les corpus existants avant de penser à en constituer un nouveau.

Il nous semble intéressant de noter dans les travaux précédemment cités que, d'une manière générale, même si la nature des données va dépendre de l'étude menée, la première question à se poser lors de la constitution du corpus est la suivante : « pour quels objectifs ces données sont-elles recueillies ? ». De la réponse à cette question découlent d'autres problématiques comme les contraintes en présence (techniques, par exemple), les différentes conditions (signifiance, acceptabilité, exploitabilité) à vérifier, l'optique choisie (quantité vs. qualité), etc.

[1] Citation originale : *« A sign language corpus is a multimedia repository that contains digital movies of signing Deaf people, along with annotations that facilitate searching through the data. »*

En résumé, un corpus de textes est constitué dans un but précis, suivant des contraintes qui peuvent ne pas être liées au corpus lui-même (limite de mémoire informatique, par exemple), et doit vérifier des conditions a posteriori.

b) Méthodologie de constitution de corpus de parole

Pour O.Baude (Baude, 2006), dans le domaine des langues orales, le rôle d'un corpus est de répondre à un des deux objectifs suivants : être exhaustif (i.e. restituer « la totalité des données sonores produites à l'occasion d'un événement, c'est-à-dire avec un commencement et une fin définis ») ou être représentatif (i.e. répondre « à la question de la représentativité : comment rendre compte d'une langue, d'un dialecte, d'un parler ? » en tant qu'« échantillon d'une langue »). La méthodologie mise en œuvre pour la constitution d'un corpus de langue orale découlera du choix de l'objectif ce qui intrinsèquement entraînera des contraintes quant aux données et à leur sélection.

D'autre part, l'auteur met l'accent sur quatre aspects des corpus oraux : « les types de données et de locuteurs, la dimension du corpus, les transcriptions ». Les données sont sollicitées, dans le sens où elles ne viennent pas toutes seules au chercheur, mais celui-ci doit créer les conditions dans lesquelles il va les recueillir. Les locuteurs sont choisis, ainsi que les situations d'enregistrement, en fonction des objectifs de départ. La taille du corpus et des éléments qui le compose sont eux aussi liés aux objectifs d'utilisation du corpus : le choix se limite généralement à l'exhaustivité ou à la représentativité. Les transcriptions effectuées sur les corpus sont de nature très différentes d'une étude à une autre, en fonction des objectifs mais aussi des domaines scientifiques d'utilisation du corpus : les transcriptions ne seront pas les mêmes en informatique, linguistique, sociologie, etc.

La constitution d'un corpus oral est ici décrite aussi bien lors de sa conception théorique qu'au moment de son traitement : en amont il s'agit de sélectionner les locuteurs, les données et une limite de taille, en aval il faut tenir compte du fait que les transcriptions font partie du corpus et les y inclure.

c) Méthodologie de constitution de corpus de langue des signes

Les premiers corpus de LS ont été conçus dans une optique de linguistique descriptive, pour mettre en lumière le fonctionnement de la langue : c'est la démarche typique de recherche « corpus-based ». Tout comme les corpus textuels et oraux, certaines recherches sur la LS considèrent le corpus comme « un observatoire d'une théorie a priori », tel que le décrit Mayaffre (2005) pour les LV. Ainsi, Sallandre (2003) a démontré le bien fondé, qualitativement et quantitativement, de la théorie de Cuxac. D'autres envisagent le corpus comme « un observé dynamique qui permet de décrire puis d'élaborer des modèles a posteriori », toujours selon Mayaffre (2005) pour les LV, ainsi que l'a fait Cuxac pour construire sa théorie, ou les partenaires du projet LS-COLIN (cf. infra) : c'est ici la démarche de recherche « corpus-driven ».

Par la suite, les corpus ont été utilisés et conçus par des informaticiens pour des recherches sur la reconnaissance automatique ou le traitement d'images. Dans la plupart des recherches, les informaticiens ont collaborés avec les linguistes pour constituer leurs corpus. Il est important de noter que les linguistes, et donc les informaticiens, se sont quasiment toujours assurés de la collaboration de la communauté sourde, afin de disposer de données valables.

De manière analogue aux corpus de langue vocale (que ce soit sous forme écrite ou sous forme orale), Thoutenhoofd (2007) exprime l'idée générale que ce qu'il faut inclure dans les

corpus de LS est fonction d'un équilibre entre représentativité et quantité, des ressources disponibles, de la possibilité ultérieure de comparer avec des données existantes, et des motifs linguistiques.

d) Exemples de corpus de LS

Des corpus vidéo de LS disponibles actuellement, nous n'en citerons que quelques-uns. Nous avons retenu ceux qui selon nous présentent les dernières recherches, celles qui se basent sur des « corpus » et pas seulement sur des ensembles de données.

- Le corpus Auslan (Johnston 2009), porte sur la langue des signes australienne (Auslan). Les données ont été recueillies par élicitation et expression « spontanée » (terme mis entre guillemets parce que le sujet devait répondre à des questions ou s'exprimer sur un sujet donné, dans un contexte de laboratoire, ce qui ne permet pas selon nous de véritable spontanéité). L'objectif de ce corpus est double : tout d'abord créer une référence de l'Auslan afin de protéger cette langue considérée « en danger », ensuite construire un corpus utilisable pour des études linguistiques. Les différents domaines d'utilisation de ce corpus sont aujourd'hui : linguistique (analyse de la grammaire et du lexique, accès à des données inédites), collaboratif (partage des données), sémiogénétique (évolution des signes dans le temps), et politique (faire pression sur les instances dirigeantes en faveur de l'Auslan).
- Le corpus NGT (Crasborn 2009), porte sur la langue des signes des Pays-Bas (NGT). Là encore les méthodes choisies pour le recueil des données sont l'expression « spontanée » et l'élicitation. L'objectif de ce corpus est d'établir un instantané des variétés de NGT en fonction de l'âge, de la région d'habitation, et du genre de discours. Ce corpus a été conçu à destination des linguistes, sociolinguistes, médecins, personnels enseignants (sourds, interprètes, professeurs de NGT), et du public sourd en général.
- Le corpus LS-COLIN (Cuxac 2001), porte sur la LSF. Les méthodes de recueil par élicitation et expression « spontanée » ont été utilisée dans un double objectif : ce corpus sert aux linguistes pour analyser la langue, et aux informaticiens pour mener des recherches sur l'analyse et la synthèse de LSF.
- Le corpus SNCF (Segouat 2008), porte sur la LSF sous forme vidéo et 3d. En effet, ce corpus est composé d'enregistrements vidéo de phrases et de vocabulaire ciblés utilisés dans les gares qui servent de support à la création d'animations 3d. Ces animations sont utilisées pour diffuser des informations en LSF, à l'instar du système de diffusion d'information audio par synthèse vocale. Les données collectées sont des énoncés et signes de la LSF issus de la traduction des énoncés vocaux du système existant. Il ne s'agit donc ni d'élicitation ni d'expression « spontanée ». L'objectif de ce corpus est double : être utilisé en informatique dans le domaine de la synthèse de LSF, et permettre l'analyse du phénomène linguistique de « coarticulation ».

Tous ces corpus ont pour objectif commun l'analyse linguistique des données. Certains ont en plus pour but d'être utilisés en sociologie, en informatique, etc. Ils ont tous été conçus dans un souci de représentativité. Les uns s'en sont assurés en essayant d'être le plus exhaustifs possible (Auslan, NGT), les autres par des considérations linguistiques en amont (LS-COLIN : les linguistes ont sélectionnés les locuteurs et les situations en fonction des théories qu'ils voulaient vérifier et ont essayé de s'assurer d'une certaine exhaustivité ; SNCF : le domaine dans son ensemble a été couvert de sorte que tous les éléments sont présents).

Ces différentes recherches ont pour but d'analyser la langue : elles se basent sur des théories linguistiques qui influent sur les données contenues dans le corpus. Certaines théories, comme nous l'avons vu, considèrent que les « structures de grande iconicité » (SGI), telles que définies par Cuxac, ne sont pas comprises dans la linguistique de la LS : ils n'en tiennent donc pas compte lors de la constitution de leur corpus, et lors de leur analyse. D'autres mettent l'accent sur l'importance de collecter des éléments issus de ces structures de grande iconicité. Ainsi, il est délicat de dire d'un corpus qu'il est correctement constitué selon qu'il contient ou pas des SGI. Il conviendra mieux de dire qu'il est constitué suivant telle théorie linguistique et de ce fait il est logique que les SGI soient prises en compte ou non.

III. Constitution de corpus : quelle(s) spécificité(s) de la langue des signes ?

De cette présentation des différents critères qui garantissent la représentativité des corpus en LV ou en LS, nous pouvons faire ressortir certains points communs mais également quelques spécificités en ce qui concerne les LS.

Un premier point commun, de considération générale, est que ce sont majoritairement des chercheurs natifs d'autres langues (le français) qui font des recherches sur la LS ; des Sourds commencent à rejoindre les EEPS (enfant entendant de parents sourds) qui font des recherches sur la LS et constituent des corpus, mais c'est une petite minorité. Ce nombre grandissant de locuteurs natifs de LS participant aux recherches peut avoir un impact sur la conception des corpus : il est possible qu'il y ait des changements dans la manière de capter cette langue, aussi bien au niveau technique qu'au niveau linguistique. Qu'en est-il dans les autres langues ? Les langues minoritaires sont majoritairement étudiées par des non natifs de la langue, pour des raisons principalement socio-économiques : les chercheurs qui ont les moyens et les connaissances peuvent faire des recherches sur leur langue maternelle mais aussi sur d'autres langues. Les personnes qui n'ont pas de moyens et pas de connaissances ne peuvent faire ni l'un ni l'autre.

Un second point commun concerne le « paradoxe de l'observateur », décrit par Labov (1973) et définit comme suit : pour recueillir les données les plus pertinentes nécessaires aux théories linguistiques, il faudrait observer la parole des personnes lorsqu'elles ne sont pas observées[2]. Ce problème se pose de manière générale lors de la collecte de phénomènes linguistiques, que ce soit pour les corpus de LV ou de LS : au même titre que la présence du micro, ou de quelque élément de l'installation de captation, la présence de la caméra va avoir une influence sur la locution captée.

Un dernier point commun qu'il nous semble intéressant de mettre en avant est que, comme le dit Mondada (2005), « de nombreux chercheurs en linguistique travaillent sur des données qu'ils n'ont pas recueillies eux-mêmes (…) ou qu'ils collectent sans nécessairement effectuer une enquête de terrain ou contacter des participants ».

Une première spécificité des LS est qu'il est très rare de pouvoir réutiliser des corpus déjà constitués. Généralement, les chercheurs recréent un corpus lorsqu'ils entament une nouvelle étude, au motif que d'une part il y a peu ou pas de corpus disponible, et d'autre part la réutilisation de corpus existant nécessite que la méthodologie de constitution et les caractéristiques du corpus soient explicitées. Contrairement aux LS, en ce qui concerne les corpus de texte ou oraux, il y a fréquemment réutilisation de données existantes, ainsi que le

[2] Citation originale : « *To obtain the data most important for linguistic theory, we have to observe how people speak when they are not being observed*. »

démontre le conseil donné par O'Donnell (2008) : lorsqu'on souhaite effectuer une recherche à partir d'un corpus, il faut réfléchir en, trois étapes. La première consiste à réutiliser l'existant, voir à combiner plusieurs corpus ou à n'en retenir qu'une partie. Si cela ne convient pas, la seconde étape préconise de redéfinir de l'existant suivant l'étude qu'on souhaite poursuivre. Enfin, si cela ne convient toujours pas, la dernière proposition est de construire son propre corpus. Nous voyons bien au travers de cette réflexion qu'il ne faut pas hésiter à réutiliser les corpus existants avant que de penser à en constituer un nouveau.

Une autre différence entre LV et LS concerne la question de l'anonymisation. D'un point de vue juridique, le droit à l'image (rattaché à l'article 9 du code civil depuis 1998) est un droit absolu de s'opposer à l'utilisation non consentie de son image. Baude (2006) se pose la question de l'anonymisation des données recueillies, que ce soit du texte, de l'audio ou de l'image. L'auteur propose trois solutions pour rendre anonyme les images (dans le domaine des corpus oraux, il s'agit « des enregistrements vidéo, des photographies ou des captures d'écran ») : la « suppression » qui revient à couper des séquences lors du montage, le « remplacement par un brouillage du signal » utilisant des techniques de floutage, de contourage, ou autres applications de différents filtres, et enfin le « placement d'un bandeau noir » sur l'image à l'endroit souhaité. Ce problème est rédhibitoire en LS quand il s'agit de diffuser les résultats d'une étude, de mettre en commun avec d'autres partenaires (institutions, chercheurs, etc.) le contenu d'un corpus, de faire des présentations scientifiques des travaux, etc. Il est indispensable de s'assurer que les locuteurs, informateurs et participants dont la LS sera captée, autorisent explicitement l'utilisation de leur enregistrement dans les différents contextes précédemment cités. En effet, au contraire des corpus oraux, il n'est pas possible de mettre en œuvre les solutions proposées par Baude. La suppression élimine tout simplement la donnée, le brouillage du signal altère tellement la donnée qu'il ne sera pas possible d'en tirer quelque conclusion ou même de mener une analyse, et le bandeau noir occulte une partie non négligeable des phénomènes linguistiques. Il n'existe à ce jour aucune technique d'anonymisation pour les corpus de LS : des recherches sont en cours pour appliquer un filtre sur le visage du locuteur, qui permettrait de coller virtuellement un autre visage qui suivrait les déformations du premier ; d'autres recherches visent à remplacer l'intégralité de ce qui est filmé (le corps entier du locuteur) par un personnage de synthèse qui là aussi suivrait les mêmes mouvements que le locuteur original. Ces recherches permettront de déterminer jusqu'à quel point doit aller l'anonymisation : est-ce qu'il suffit de synthétiser seulement le visage, ou faudra t'il synthétiser également les mains, ou le corps dans son ensemble ?

Une dernière différence tient dans le contexte sociolinguistique de la LS. Elle a été interdite pendant un siècle, et l'éducation des jeunes sourds est toujours aussi chaotique : même s'il y a ponctuellement des améliorations, l'éducation des enfants sourds relève toujours majoritairement du ministère de la santé au lieu du ministère de l'éducation. Cette particularité pèse lourdement sur les parcours scolaires, puisque la rééducation orale prend une place importante et a par conséquent un impact non négligeable sur la relation qu'ont ces jeunes avec la LS. L'altération de cette relation, qui se traduit concrètement par une dévalorisation de la langue à leurs yeux, est en porte à faux avec la représentation valorisante qu'ont les plus anciens face à cette langue qui leur a été interdite et dont ils ont grand soif maintenant. Ce phénomène, bien plus complexe que ce que nous décrivons ici succinctement, est quelque chose qu'on ne retrouve pas à notre connaissance en LV actuellement. Cela peut influer de manière importante sur la manière de constituer un corpus : avec ces jeunes sourds, il s'agirait alors de mettre en place une méthodologie qui ne soit pas en contradiction avec l'image qu'ils ont de leur LS (par exemple, ne pas se focaliser sur la LS mais user de moyens détournés pour les faire s'exprimer en LS). En effet, ce n'est que récemment par la loi n°2005-102 du 11

février 2005 pour l'égalité des droits et des chances, la participation et la citoyenneté des personnes handicapées, que la LSF a été reconnue officiellement. Dans les institutions le règne de l'oralisme (éducation par l'oral : les sourds doivent apprendre à parler, cause d'un retard conséquent pour l'apprentissage des différentes matières et surtout du français écrit) perdure, et il existe peu d'écoles qui accordent une réelle place à la LSF dans l'enseignement. Pour ces jeunes, la LSF a peu de valeur, et ils communiquent via une langue métissée entre le français et la LSF : les faire s'exprimer en LS seulement nécessiterait donc une approche particulière.

Les spécificités des LS que nous avons pu relever portent donc sur : la motivation de la conception de corpus (pas de possibilité de réutiliser l'existant), l'impossibilité actuelle d'anonymiser les corpus (ce qui implique des considérations linguistiques et juridiques), et le contexte sociolinguistique (autorisation dans l'éducation et reconnaissance officielle récentes). Tout en tenant compte de ces spécificités, nous présentons dans la partie suivante la notion de représentativité et les moyens existants pour s'assurer de sa validité.

IV. Assurer la représentativité : quelles méthodes ?

Un des aspects qui nous intéresse actuellement est de savoir comment assurer en amont la représentativité d'un corpus. En effet, les études sur corpus menées jusqu'à présent, en France notamment, n'ont pas abordé cette question alors qu'elles tendent à proposer des théories générales sur la LSF. De plus la communauté sourde signante a des spécificités (son histoire, la situation de diglossie français/LSF, etc.) qui potentiellement influent sur la méthodologie à mettre en place pour garantir cette représentativité.

Dans un premier temps nous définissons la notion de « représentativité » dans le domaine des corpus de langues vocales (sous forme écrite ou orale) puis des corpus de langues signées. Ensuite nous présentons pour les mêmes domaines les méthodologies mises en place afin de garantir en amont et/ou vérifier en aval qu'il y a bien représentativité. Enfin, nous présentons notre méthodologie qui nous permet d'assurer une représentativité à deux niveaux.

a) Définition de la représentativité

Pour Greimas (1966), un corpus représentatif contient une et une seule instance d'un type d'élément issu d'un univers donné. Selon Bommier-Pincemin (1999), dans le domaine des corpus de textes, la représentativité consiste à recueillir le maximum d'éléments différents, sans précision sur le nombre d'instances de chaque élément. De manière implicite, en considérant qu'un corpus de référence est un corpus représentatif issu d'un corpus plus vaste, Sinclair (1996) abonde dans le même sens.

Il n'y a pas de définition particulière de la représentativité qui serait adaptée à la recherche sur les LS. Étant donné qu'une bonne partie des notions liées aux corpus de LS est empruntée aux LV, nous pouvons considérer que la définition proposée précédemment s'applique aux corpus de LS. Néanmoins, la notion de représentativité est fortement liée à l'objet considéré. Ainsi, nous constatons qu'il y a des spécificités de la LSF dues à des aspects sociaux : il s'agit d'une micro communauté (il n'existe pas de chiffres exacts dénombrant les personnes sourdes signantes en France, même si le chiffre de 300000 personnes signantes –sourdes ou entendantes– est le plus répandu (Gillot, 1998)) dont les membres sont éparpillés sur tout le territoire. Leur langue a été interdite dans l'enseignement pendant un siècle, l'éducation des

jeunes est encore balbutiante et cette langue n'a pas encore été complètement décrite linguistiquement (le modèle de Cuxac sur lequel nous nous appuyons est en cours d'évolution, même si cela reste le modèle le plus abouti en France).

La notion de représentativité pour les corpus de LV est acceptable telle quelle pour les corpus de LS. Les critères à respecter pour assurer la représentativité dans les corpus de LS seront issus aussi bien de la linguistique que de la sociologie, et ce même si le but de l'étude menée a une portée uniquement linguistique.

b) Représentativité dans les corpus oraux et les corpus de textes

Nous avons relevé dans la littérature quelques exemples de recommandations ou de manières de faire afin d'assurer la représentativité d'un corpus. Voyons tout d'abord ce qu'il en est pour les LV.

Habert (2000) explique qu'un choix peut-être fait en amont pour assurer la représentativité du corpus et que ce choix est fonction de la vision qu'a le chercheur de la représentativité. Il existe, selon l'auteur, trois choix possible : sélectionner avec soin un échantillon de chaque catégorie créée suivant une thématique donnée, constituer un corpus dans des conditions où seraient présentes toutes les catégories, regrouper les éléments du corpus selon des similarités linguistiques. Quel que soit le choix effectué, il permettrait de garantir une certaine représentativité du corpus. Le même auteur explique comment il est possible d'« améliorer la représentativité d'un corpus ».

En ce qui concerne l'évaluation de la représentativité, pour Abouda (2006), ce « qui permet de juger de la représentativité de corpus (…) », c'est « l'explicitation des bornes du corpus (conditions de productions, de réception, contexte des usages, informations sociologiques sur les producteurs, genre, etc.) ». On retrouve dans les deux citations une constante qui est l'importance de la clarification de la méthodologie de recueil des données : cela a une place très importante sur la représentativité ou non des données du corpus, et sur l'évaluation de cette représentativité.

c) Représentativité dans les corpus de langue des signes

La question de la représentativité dans les corpus de LS nous apparait importante dans la mesure où il nous semble difficile d'effectuer une analyse et de tirer des conclusions si nous ne sommes pas en mesure d'assurer que les données sur lesquelles nous nous appuyons sont représentatives du phénomène que nous voulons observer.

Prenons le cas du corpus LS-COLIN : il a été conçu dans le but de servir de base à des analyses linguistiques sur la présence et l'utilisation de certaines structures de la LS. Tant au niveau du choix des locuteurs qu'au niveau des genres de discours demandés, l'accent a été mis sur la diversité : pourtant certains genres discursifs (le dialogue par exemple) et certains locuteurs (très jeunes ou très vieux) n'ont pas été pris en compte dans ce corpus. Bien que la représentativité de ce corpus soit « partielle » (si on occulte les aspects non pris en compte lors de la constitution du corpus, les données effectivement recueillies sont représentatives d'une partie de la langue), les théories extraites de ce corpus restent valables mais demandent à être vérifiées sur un corpus qui, lui, serait « totalement » représentatif. Le problème que nous souhaitons ici mettre en avant est que la représentativité nous assure une validité de la théorie qui en découle, mais que cette représentativité n'a pas été « de fait » dans tous les corpus de LS qui ont servis à bâtir des théories sur la langue.

Dans le domaine des LS, la représentativité a été initialement assurée par le recours à l'exhaustivité dans le corpus. C'est toujours le cas pour certains corpus de linguistiques et pour tous les corpus dont un des objectifs est patrimonial. Cette exhaustivité à elle seule garantit-elle la représentativité ? Nous n'en sommes pas convaincus, d'autant plus que ce n'est pas le cas dans le domaine des LV. Cependant, historiquement, il s'agissait de savoir ce qu'était la langue pour ensuite mettre au point des critères de représentativité. Ces critères ne sont pas aujourd'hui complètement formalisés, néanmoins des spécificités de la LS il ressort que la représentativité d'un corpus de LS sera généralement fonction de considérations linguistiques, sociologiques, voire historiques.

Par exemple, un des critères sociolinguistique de la représentativité dans un corpus de LS est le type du locuteur : est-ce un signeur natif ou non ? La plupart des études privilégient le signeur dit « natif ». Mais que signifie « natif » ? Nous n'aborderons pas ce point en détail ici, puisqu'il tire sa réponse de données sociolinguistiques, ethnologiques, et linguistiques, entre autres qui ne sont à ce jour pas clairement définis.

Nous pouvons noter qu'il n'existe pas en LS de méthodologie qui puisse garantir la représentativité au sens où chaque élément doit être présent une et une seule fois dans le corpus. Cependant, nous poursuivons l'objectif double de création de corpus en vue d'une utilisation dans un système informatique, et d'analyse d'un phénomène linguistique (la coarticulation). Pour ce faire, nous devons concevoir un corpus représentatif pour valider notre recherche linguistique. Nous l'avons conçu non seulement en respectant les principes de clôture et de redondance, mais aussi en respectant les spécificités sociologiques liés à cette langue.

V. Étude de cas : le corpus SNCF

Notre méthodologie résulte de plusieurs années de pratique, de notre partenariat régulier avec des linguistes spécialistes de la LS et de notre connaissance de la communauté sourde et de la LSF. Nous avons réalisé plusieurs corpus dans le cadre de projet nationaux et internationaux, avec des visées de traitement informatique (pour l'analyse ou la synthèse d'images) et d'analyse linguistique (pour l'étude de la langue) : LS-COLIN, TALS05, SNCF, (précédemment cités), entre autres. Nous présentons ici la méthodologie mise en œuvre dans le cadre de la constitution du corpus SNCF, qui à notre sens est représentatif.

Le corpus SNCF a été constitué dans l'objectif de servir à une diffusion d'information dans les gares ferroviaires. Le dispositif, un écran permettant de visualiser un signeur virtuel, est couplé au système de synthèse vocale existant : les informations sont délivrées après concaténation entre elles de petites unités de la langue. Il s'agissait de réaliser un corpus vidéo dans un double objectif : recueillir toutes les unités dont nous allions avoir besoin pour former les énoncés en LS, et permettre l'analyse du phénomène de coarticulation[3] en LS.

Nous avons souhaité que ce corpus soit représentatif de la langue utilisée au sein de la communauté sourde signante, aussi bien au niveau du lexique que de la grammaire. Ceci est rendu possible par le fait que les domaines lexical et grammatical sont clos : la liste des messages possibles implique un nombre fini et connu d'énoncés, et leurs possibilités de combinaison grammaticale sont connues. D'une part, nous avons essayé de proposer une LSF issue de la communauté elle-même : la représentativité ne vient donc pas de la diversité des éléments du corpus, mais d'un travail en amont de sélection de ces éléments. Pour cela, des locuteurs sourds de LSF sont intégrés tout au long du processus d'élaboration. D'autre part,

[3] La coarticulation est l'ensemble des modifications que subit un signe lorsqu'il est exprimé en contexte (Segouat 2009).

dans le cadre de notre étude de la coarticulation, nous avons recueilli diverses combinaisons de signes en contexte, avec plusieurs locuteurs, afin de garantir une diversité optimale qui permettra à notre modèle final d'avoir une certaine validité.

Afin de répondre au mieux à cette problématique de la représentativité, la méthodologie employée pour la constitution de ce corpus comporte plusieurs étapes.

Nous avons besoin d'avoir un modèle vidéo de chacun des signes. Ce modèle vidéo est créé en filmant un locuteur Sourd (étape 3), dont les productions signées sont soit exprimées directement, soit préparées en amont par un informateur Sourd qui maîtrise les concepts qui vont être signés (étape 2). Si besoin, l'informateur Sourd rencontre un informateur du domaine qui lui permet d'appréhender au plus juste les notions (étape 2). Auparavant, il faut donc choisir un locuteur sourd, un informateur Sourd et un informateur du domaine (étape 1)[4]. Ces différentes étapes sont présentées dans la figure suivante.

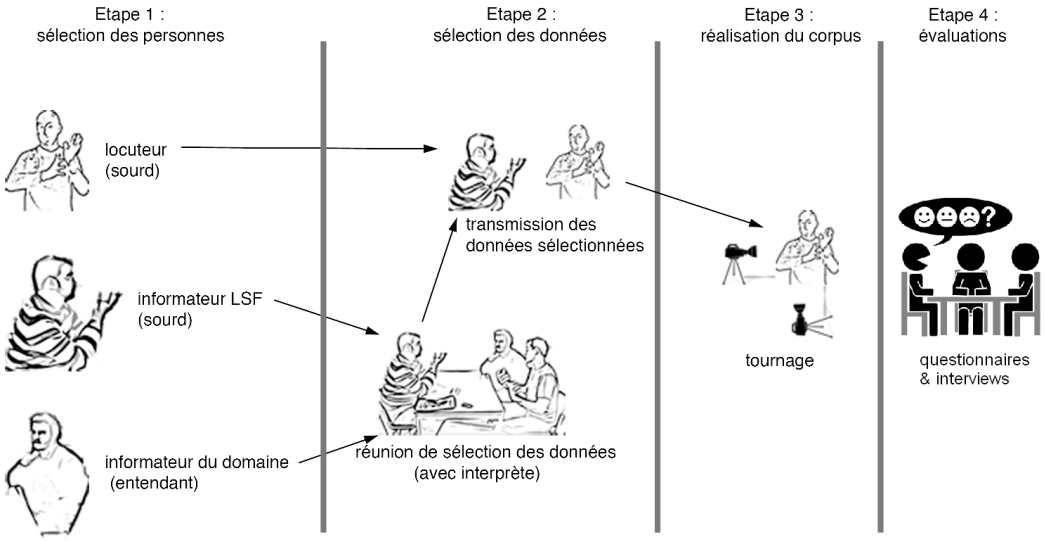

Figure 2 : schéma descriptif de notre méthodologie

La première étape de notre méthodologie consiste à *sélectionner les personnes impliquées* dans la conception du corpus. Dans un premier temps, nous définissons qui sera notre « locuteur ». Par ces termes nous désignons la personne qui sera employée pour exprimer en LSF tous les signes dont nous avons besoin. Cette personne doit être un sourd, maîtrisant parfaitement la LSF, ayant une grande facilité ainsi qu'une grande clarté d'expression. Dans le cas du projet SNCF, nos locuteurs sont des personnes ayant suivi une formation universitaire de traducteur, dont le travail quotidien est la traduction vers la LSF de dépêches d'informations écrites en français. Leurs prestations sont numérisées et diffusées sur le Web, ce sont donc des personnes qui ont l'habitude d'être face à la caméra. Elles satisfont ainsi pleinement aux critères fixés pour le choix du locuteur. En parallèle, nous contactons des « informateur LSF »: des personnes sourdes spécialistes de leur langue mais aussi d'un des domaines pour lequel nous souhaitons recueillir des signes. L'informateur LSF peut ne pas s'estimer assez compétent quant à la sémantique de la terminologie d'un domaine pour lequel nous souhaitons disposer de signes. Ses compétences en LSF et ses connaissances dans des domaines proches lui permettent

[4] Suivant les travaux de Mondada (2005) sur l'analyse des corpus conversationnels, nous avons décidé d'adopter ses dénominations pour les différents intervenants. L'auteur définit le rôle d'« informateur » comme renvoyant « à une conception de l'enquête comme ayant pour objectif de recueillir des informations et donc à identifier les meilleures sources pour cela » et le locuteur comme « une catégorie qui (…) privilégie la parole ».

d'être le mieux placé pour appréhender au mieux les notions qui lui échappent : il va rencontrer un « informateur » du domaine qui lui expliquera ces notions. Nous avons choisi des informateurs de LSF reconnus par la communauté scientifique de leur domaine et par la communauté sourde. Pour cela, dans le même temps, nous prenons contact avec des « informateurs » entendants de chaque domaine pour lequel nous voulons recueillir du lexique. Dans le cas du système d'information en gare, nous sommes en partenariat avec l'entreprise WebSourd[5] pour un projet commandé par la SNCF. Il s'agit de réaliser des messages d'information tels que ceux diffusés vocalement. La SNCF fournit à l'entreprise les phrases dont elle désire la traduction et des informateurs de la SNCF sont à la disposition des informateurs de LSF, si besoin est.

La prise de contact avec les locuteurs, les informateurs de LSF et du domaine étant effectuée, nous passons à la seconde étape qui est la *conception du corpus vidéo*. Il peut être fait appel à plusieurs informateurs, de LSF, d'un domaine, voire de plusieurs domaines (pour des raisons de clarté de discours nous nous exprimerons au singulier). Ils se réunissent afin de permettre à l'informateur LSF de réfléchir à la meilleure façon de traduire les différentes notions : l'entretien a lieu sous la forme d'une discussion, durant laquelle l'informateur LSF propose des interprétations en LSF des termes en français écrit, ces propositions étant validées ou nuancées par l'informateur du domaine. Enfin, l'informateur LSF rencontre le locuteur afin de lui transmettre tous les signes et explications qui vont constituer le corpus vidéo.

Lors de la troisième étape nous procédons à la *réalisation du corpus* vidéo.

La quatrième et dernière étape concerne l'*évaluation du corpus* et sa validation. Cette évaluation se situe sur deux plans : un dit « académique », un autre plus « communautaire ». Dans le premier cas il s'agit de faire évaluer les différentes productions par des chercheurs en linguistique, des professionnels sourds de différents domaines proches de ceux dont il est question dans le corpus, etc. Dans l'autre cas l'évaluation est faite par le public sourd : dans le cas du système d'information en LSF dans les gares, ce seront des usagers des transports ferroviaires. Les retours collectés concernent bien sûr des aspects techniques et technologiques, mais aussi des aspects linguistiques et sociologiques. Cette évaluation nous permet de nous assurer de l'acceptation par l'ensemble des locuteurs de la LS des données (les signes) sélectionnées au départ.

Lors de la seconde étape de notre méthodologie, nous impliquons des informateurs de LSF et des informateurs du domaine. Le rôle des informateurs de domaine est de s'assurer que les informateurs de LSF ont toutes les connaissances nécessaires pour déterminer quels signes et énoncés ils vont devoir sélectionner. Les informateurs de LSF établissent alors une liste des signes et énoncés potentiels et, suivant des critères qui leur sont propres, sélectionnent un élément unique parmi chaque liste. Nous pouvons dire que l'ensemble des listes établies par les informateurs de LSF est un ensemble représentatif de la LS dans le domaine considéré. Nous pouvons également considérer que les données du corpus, puisque choisies parmi un ensemble représentatif afin d'être compris par le plus grand nombre, constituent une variété standard de la langue. Ces affirmations demandent à être confirmées par les résultats des évaluations que nous allons mener sur les corpus réalisés avec notre méthodologie.

La représentativité du phénomène de la coarticulation est de fait : le domaine lexical est clos, et la combinaison des signes entre eux est exhaustive. Le premier point est confirmé par les informateurs du domaine, qui assurent que les concepts du domaine sont tous présents dans la sélection du corpus. Le second point est justifié par le fait que ce sont des experts qui ont travaillés à la mise en signe des énoncés en français. Il y a donc autant de combinaisons des

[5] http://www.websourd.org

énoncés en LSF possibles que celles qui sont produites dans le système vocal. L'ensemble des combinaisons d'énoncés possibles est donc exhaustif.

Conclusion

Les recherches dans le domaine des LS, que ce soit en informatique ou en linguistique, s'appuient de plus en plus sur des corpus. La notion de corpus est à ce jour implicite, de même que les méthodologies qui permettent de créer ces corpus. Or, la définition du concept de corpus de LS et l'explicitation des principes méthodologiques de constitution de tels corpus ont un impact certain quant aux traitements et aux analyses qui peuvent être effectués. Nos recherches abordant la question de la représentativité, nous avons pu noter que cette notion, bien que manipulée par d'autres chercheurs, n'était pas définie de manière claire mais là encore implicitement.

Cet article constitue une première réflexion sur les définitions des concepts liés aux corpus de LS tels que la notion même de corpus, les principes méthodologiques de constitution des corpus, et la représentativité. En ce qui concerne les principes méthodologiques, nous avons noté des points communs mais aussi des différences, qui ne permettent pas de conclure sur une éventuelle possibilité d'utiliser telle quelle une méthodologie issue de corpus oral ou écrit, mais qui enjoint à approfondir la réflexion.

Une des pistes que nous envisageons sera de proposer une définition de ce qu'est un corpus de LS, qui fasse un lien avec la notion de scripturalité et d'oralité. En effet, étant donné que les corpus sont définis différemment dans le domaine « oral » et dans le domaine « écrit », mais que ces notions ne sont pas très bien établies (elles sont toujours en discussion dans le domaine de la linguistique de corpus, entre autres), il nous semble intéressant de considérer des notions comme la scripturalité ou l'oralité, dont nous avons encore à appréhender la définition exacte. Ensuite il s'agira de décider de principes méthodologiques à respecter dans le cadre de la constitution de corpus de LS, en lien avec la définition de corpus précédemment établie. En ce qui concerne plus particulièrement notre méthodologie, qui propose une approche de la représentativité et de sa validation, nous allons débuter son évaluation. Les résultats nous permettront de valider notre processus de constitution et d'utilisation de corpus.

Les réflexions proposées dans cet article sont applicables aux autres champs disciplinaires s'intéressant aux LS et qui voudraient constituer un corpus. Dans le domaine des LS, puisque différentes théories linguistiques s'affrontent, nous pouvons émettre l'idée que des corpus dits représentatifs constitués à partir de chacune des théories pourraient permettre de déterminer celle qui représente le mieux les LS. Enfin, ces questions pourraient fédérer une ligne directrice commune à plusieurs chercheurs, ce qui permettrait une plus grande interconnexion des méthodologies de constitution, des traitements, des méthodologies d'évaluation, et des résultats.

Remerciements

Nous tenons à remercier tout particulièrement les relecteurs de cet article, messieurs Jean Chuquet et Christophe Benzitoun, ainsi que toute l'équipe d'organisation du colloque du CerLiCo, et en particulier madame Hélène Chuquet. Merci également aux nombreuses personnes qui nous ont fait des retours très intéressants lors de notre communication au colloque. Nous n'oublions pas les nombreuses personnes sourdes avec qui nous collaborons, avec qui les échanges sont très riches et dont les retours sont toujours très précieux.

La recherche ayant conduit à ces résultats a reçue un financement du 7ème programme cadre de la Commission Européenne (FP7/2007-2013) sous la convention de subvention n°231135.

Bibliographie

ABOUDA, Lofti, et BAUDE, Olivier, 2006, « Constituer et exploiter un grand corpus oral : choix et enjeux théoriques. Le cas des ESLO », *Corpus en Lettres et Sciences sociales : des documents numériques à l'interprétation, Actes du colloque international d'Albi, juillet 2006*.

BAUDE, Olivier, 2006, *Corpus oraux, Guide des bonnes pratiques 2006*, Paris et Orléans , CNRS éditions et PUO.

BENZÉCRI Jean-Paul et coll., 1981, « Pratique de l'analyse des données », *Linguistique et lexicologie*, Paris, Dunod.

BOMMIER-PINCEMIN, Bénédicte, 1999, *Diffusion ciblée automatique d'informations : conception et mise en œuvre d'une linguistique textuelle pour la caractérisation des destinataires et des documents*, Thèse de Doctorat, Université Paris IV (Sorbonne).

CRASBORN, Onno, 2009,

CUXAC, Christian, 2000, *La Langue des Signes Française (LSF) – Les voies de l'iconicité*, Paris, Faits de Langues vol 15-16, Ophrys.

CUXAC, Christian, 2001,

GILLOT, Dominique, 1998, « Le droit des sourds » .Rapport commandé par le Premier Ministre.

GREIMAS, Algirdas Julien, 1966, *Sémantique structurale*, Paris, Larousse.

GUITTENY, Pierre, LEGOUIS, Philippe, et VERLAINE, Laurent, 2004, *La langue des signes*, Centre d'Information sur la Surdité d'Aquitaine.

HABERT, Benoît, 2000 « Des corpus représentatifs : de quoi, pour quoi, comment ? », Mireille Bilger (dir.), *Linguistique sur corpus : Études et réflexions*. Perpignan : Presses Universitaires de Perpignan, Cahiers de l'Université de Perpignan N 30 – 2000, pp 11-58.

JOHNSTON, Trévor, 2009,

LABOV, William, 1973, *The boundaries of words and their meanings. New ways of analyzing variation in English*, Washington, Georgetown University Press, pp 340–373.

MAYAFFRE, Damon, 2005, *Rôle et place des corpus en linguistique : réflexions introductives*, Toulouse, édité par P. VERGELY, 5-18.

MONDADA, Lorenza, 2005, « Constitution de corpus de parole-en-interaction et respect de la vie privée des enquêtés : une démarche réflexive », *Rapport sur le projet « Pour une archive des langues parlées en interaction »*. Université Lyon 2 et CNRS.

O'DONNELL, Matt, 2008, présentation au Corpus Linguistics Summer Institute, Liverpool, du 30 juin au 3 juillet.

SALLANDRE, Marie-Anne, 2003, *Les unités du discours en Langue des Signes Française. Tentative de catégorisation dans le cadre d'une grammaire de l'iconicité*, Thèse de Doctorat, Université Paris 8 Saint-Denis.

SEGOUAT, Jérémie, et BRAFFORT, Annelies, 2008, « Proposition d'une méthodologie de réalisation d'un corpus de signes 3D isolés de LSF », *Atelier Traitement Automatique des Langues des Signes - TALN RECITAL 2008*, Avignon, France.

SINCLAIR, John, 1996, « Preliminary recommendations on corpus Typology », Technical Report, Eagles.

STOKOE, William, 1960, « Sign Language Structure: An Outline of the Visual Communication System of the American Deaf », *Studies in Linguistics*, New York.

THOUTENHOOFD, Ernst, et CRASBORN, Onno, 2007, « Sign language corpus creation and interface concerns », The Emergence of Corpus Sign Language Linguistics, 40èmes Rencontres Annuelles de l'Association de Linguistique Appliquée (BAAL), du 6 au 8 Septembre, Edimbourg.