Transcriptions reliées au traitement informatisé d'un corpus en langue des signes

Dans nos recherches sur la structure de la LSQ, la source la plus importante de données est un corpus d'enregistrements sur vidéo de conversations entre signeurs sourds. Le corpus LSQ88 contient 16 enregistrements de deux heures, pour un total de 34 heures d'enregistrements. Il y a dans le corpus 11 signeurs répartis en groupes de trois, sauf un enregistrement où il n'y a que deux signeurs et un autre où il y en a quatre. Sur les 11 signeurs, onze participent à une session seulement tandis que la onzième participe à toutes les sessions à titre d'animatrice. Lorsqu'on travaille avec de telles données primaires, on rencontre des problèmes qui ne se présentent pas de la même façon lorsqu'on obtient des données par «élicitation».

Jusqu'à présent, nous repérons nos données primaires surtout en visionnant partiellement les vidéos. Il ne s'agit pas, pourtant, d'un moyen efficace de profiter maximalement de la richesse du corpus qui est à notre disposition. Chaque fois que l'on entreprend une recherche sur un aspect de la structure de la langue, il faut recommencer à nouveau le visionnement des vidéos afin de repérer des exemples du phénomène qui fait l'objet de l'étude en cours. Ensuite, les portions pertinentes du corpus sont transcrites selon les besoins de l'étude. Le résultat cumulatif de cette approche est une prolifération de transcriptions partielles sur différents aspects de la structure qui sont rarement pertinentes pour d'autres études : il arrive même qu'on soit obligé de retranscrire la même portion du corpus afin de pouvoir utiliser les données pour une autre étude. Outre cela, la majorité des transcriptions partielles ont été faites sur papier. Pour celui ou celle qui veut consulter des transcriptions déjà complétées pour repérer des exemples d'un phénomène donné, la tâche devient longue et ardue.

Ces problèmes ne peuvent qu'être multipliés par le fait que nous sommes actuellement en train de constituer d'autres corpus vidéos, dont un corpus d'enfants apprenants de la LSQ et un corpus de signeurs âgés. Pour localiser de manière maximalement efficace les données désirées dans les divers corpus enregistrés, il faut un moyen de repérer automatiquement toutes les occurrences d'un comportement défini par l'individu qui entreprend la recherche, avec un moyen fiable de localiser les données brutes dans le corpus enregistré. Pour accomplir cette tâche, il faut minimalement une transcription suivie de l'ensemble du corpus enregistré.

Afin d'alléger la localisation et le traitement des donnés provenant de nos corpus enregistrés, nous nous sommes proposés d'implanter un outil informatisé de gestion de données. Dans l'étape actuelle, nous essayons de déterminer la meilleure forme que peut

prendre la transcription des données à l'ordinateur, tout en considérant les besoins des utilisateurs et les possibilités qui sont offertes par les logiciels disponibles.

L'expérience acquise par les chercheurs dans la gestion de grands corpus de textes est considérable, surtout dans les études littéraires. Dans la conception de notre outil informatisé de gestion de corpus, nous nous inspirons en grande partie des démarches décrites par Poplack (1989)[1] afin de rendre possible l'analyse linguistique et quantitative de l'ensemble du «mégacorpus» du français parlé d'Ottawa-Hull.

Le corpus Ottawa-Hull comprend des enregistrements sur cassette audio d'entrevues sociolinguistiques avec 120 locuteurs, pour un total de quelques 270 heures et plus de 3,5 millions de mots. L'énormité de ce corpus a rendu capital l'implantation d'un moyen de gérer les données afin de les rendre utilisables par les chercheur(e)s. La solution choisie par Poplack était de transcrire l'ensemble du corpus sur ordinateur et, à un stade ultérieur, de faire une concordance de l'ensemble du corpus au moyen du logiciel Oxford Concordance Program[2]. Au moyen de ce dernier, il est possible de consulter des listes de fréquences, des index et des concordances pour chaque mot prononcé par un locuteur du corpus et ce, pour l'ensemble du corpus ou par locuteur. Après la constitution de la concordance par l'ordinateur VAX de l'Université d'Ottawa, une impression sur papier de l'ensemble du corpus, des index, des listes de fréquence et de la concordance a constitué l'outil de consultation pour effectuer les recherches linguistiques. Il s'agit, en effet, d'une façon efficace de repérer les données requises et de les localiser par index temporel sur les enregistrements audios.

Avant même que la transcription du corpus soit prête à subir le traitement informatisé pour créer la concordance, il a fallu faire plusieurs étapes de révision du corpus transcrit sous plusieurs révisions successives afin d'éliminer le nombre maximal d'erreurs et de transformer certains aspects de la transcription selon un protocole explicite afin de distinguer différentes formes linguistiques qui seraient confondues si l'on gardait les normes orthographiques du français. L'ensemble du processus de transcription, de vérification et de correction — accompli par deux assistants travaillant à temps complet — s'est étalé sur une période de trois ans.

---

[1] Poplack, Shana. 1989. The care and handling of a megacorpus: The Ottawa-Hull French Project. In Fasold, Ralph W. et Deborah Schiffrin (dirs pub.), *Language change and variation. Current issues in linguistic theory* n° 52. Amsterdam : John Benjamins.

[2] Hockey, S. et I. Marriott. 1980. Oxford Concordance Program. Version 1.0. Oxford, Angleterr : Oxford University Computing Service.

Or, l'expérience acquise dans le contexte du projet Ottawa-Hull peut nous renseigner sur l'utilité de certaines démarches que nous pourrons considérer dans le contexte du corpus LSQ88 et les corpus ultérieurs. L'outil que nous proposons diffèrera, pourtant, de celui construit par Poplack, compte tenu des progrès effectués en technologie informatique pendant la décennie écoulée depuis l'implantation de la concordance du français d'Ottawa-Hull. En effet, nous nous proposons d'accomplir l'ensemble de la gestion de nos corpus sur micro-ordinateurs. Nous nous proposons donc d'emmagasiner l'ensemble des données sur micro-ordinateur afin de permettre aux chercheure(s) d'effectuer des recherches automatisées, donc efficaces et économiques en termes du temps et de l'effort requis. L'impression sur papier ne sera donc que l'étape finale de chaque recherche et ne comprendra que les informations qui sont pertinentes pour les fins de chaque projet de recherche.

Néanmoins, les particularités structurales des langues signées nous posent certains problèmes cruciaux outre ceux rencontrés dans le traitement des corpus textuels de langues orales. Tout d'abord, une transcription unilinéaire est insuffisante pour représenter fidèlement la structure d'une langue signée. Peu importe le niveau de structure qui constitue l'objet de notre intérêt, l'occurrence simultanée de plusieurs comportements structurellement autonomes est peut-être la caractéristique la plus frappante d'une langue signée.

Cette complexité simultanée s'exprime comme plusieurs couches de structure superposées, chacune distincte dans sa nature des autres. On observe une interaction constante entre éléments autonomes sur les différentes couches ce qui nécessite absolument que les comportements présents sur chaque couche soient encodés indépendamment : autrement, la nature des données serait faussée. Afin d'éclairer l'importance de ces divers niveaux de structure, nous décrivons ici la nature et l'importance de chaque type de comportements simultanément encodés.

L'utilisation morphosyntaxique de l'espace.

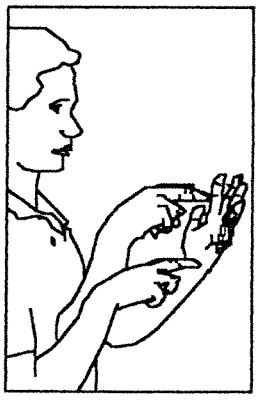
Un aspect de la structure des langues signées qui est fondamental à l'expression des relations grammaticales entre les arguments du verbe concerne l'utilisation de l'espace. Dans le discours, on associe souvent la référence des actants à des loci spatiaux définis soit en exécutant le signe au locus choisi, soit en le pointant ou en signant un verbe à partir de ou vers ce locus. Une fois le locus établi, on peut exprimer son référent de façon pronominal ou anaphorique soit en le pointant ou en exécutant un verbe directionnel vers ou à partir de ce point-ci.

L'utilisation indépendante des deux mains.

Le fait qu'il y ait deux mains qui sont capables, en principe, de signer indépendamment répercute de façon importante sur la structure grammaticale d'une langue signée et ce, aux niveaux de la phonologie, de la morphologie, de la syntaxe et du discours. L'utilisation autonome des deux mains entame souvent des interactions complexes avec l'utilisation de l'espace, la structure morphologique des signes et la structure syntaxique et disursive de l'énoncé. Il serait tout simplement imnpossible d'étudier ces phénomènes sans rendre compte, dans la transcription, de l'aspect bimanuel de la structure.

La complexité morphologique s'exprime souvent, dans les signes à deux mains, par l'encodage sur chaque main d'un morphème distinct, comme dans les signes RENCONTRER ou LIRE (1). RENCONTRER est formé avec le même classificateur (Cl. 1 «personne») sur les deux mains, représentant deux individus qui se rencontrent. Il peut arriver que les deux classificateurs soient situés chacun au début du signe à un locus spatial distnct, agissant ainsi comme entités syntaxiquement autonomes.

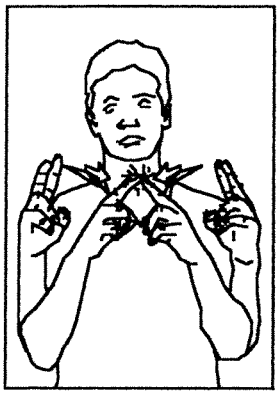
(1)



LIRE

Mais les signes à deux mains ne sont pas tous morphologiquement complexes : une grande proportion des signes sont produits sur les deux, par exemple CONNAÎTRE (2) et PAS-ASSEZ. Néanmoins, ces signes peuvent aussi participer à certains phénomènes importants aux niveaux syntaxique et discursif : il faut rendre compte, dans la transcription, du fait qu'ils sont produits par les deux mains.

(2)



CONNAÎTRE

Au niveau syntaxique, il faudra tenir compte de la tenue d'une main d'un signe pendant la production de signes sur l'autre et ce, en fonction des différences déjà mentionnées au niveau de la structure morphologique. Plus précisément, il faudra tenir compte des phénomènes suivants dans la transcription du corpus : les encodages parallèles de signes; les inversions de dominance; les interactions entre ces deux phénomènes et la structure spatiale de l'énoncé ou avec la structure morphologique des signes, selon qu'ils sont uni- ou bimanuels.

Oralisation

La présence de l'oralisation de mots provenant du français (et parfois d'autres langues, surtout de l'anglais) est important pour l'étude des effets du contact avec les langues orales environnantes, e.g. pour déterminer la nature de l'emprunt lexical, du mélange de codes ou d'autres résultats du contact qui pourrait être spécifiques aux langues signées.

Les expressions faciales ou comportements du corps

Un autre niveau de structure qui est d'une importance primordiale concerne les expressions faciales et certains autres comportements non manuels. L'expression faciale ainsi que d'autres comportements non manuels est souvent, un peu comme l'intonation dans certaines langues orales, une clé fondamentale pour l'interprétation des énoncés. Non seulement ont-ils des fonctions syntaxiques, mais certains comportements non manuels sont significatifs au niveau morphologique, surtout en ce qui a trait à l'expression de nuances aspectuelles sur les verbes et adverbiuales sur les adjectifs. Il existe également certains comportements non manuels qui peuvent être considérés comme unités lexicaux.

Comment gloser?

dans les gloses elles-mêmes, ainsi que d'ajouter des informations supplémentaires sur des lignes ajoutées (ou dans des colonnes ajoutées en format transposé).

Nous reproduisons en annexe un exemple d'une transcription Excel (importée dans MS Word sous forme de tableau) afin d'illustrer les conventions que nous avons adoptées.

Premièrement, il est nécessaire de séparer sur des lignes différentes les comportements distincts afin d'en indiquer clairement la durée relative : par rapport à la notation des comportements non manuels distincts, c'est déjà pratique courante chez les chercheurs sur les langues signées. Nous transcrivons aussi les signes produits par les deux mains sur des lignes séparées afin de capter les phénomènes d'encodage parallèle de signes et d'inversions de dominance. Ainsi, dans la première tranche de transcription dans l'annexe, on observe des signes distinctes sur les deux mains.

Il est également nécessaire de noter les loci spatiaux des deux mains indépendamment parce qu'ils peuvent pointer, ou être exécutés, à des loci spatiaux distincts. Un exemple de ce phénomène se trouve dans la première colonne de la tranche f) de la transcription : la main droite pointe une région de l'espace devant le signeur situé légèrement vers le bas, à droite sur l'écran vidéo, et à une distance moyenne à peu près équivalente à la longueur de l'avant-bras tendu.

Afin de faciliter les analyses des phénomènes impliquant des interactions entre la structure spatiale de l'énoncé ou la structure morphologique des signes et l'utilisation des deux mains, nous indiquons, par le biais d'indices, la classe structurale de chaque signe en relation avec la structure bimanuelle. Les indices utilisées sont les suivantes : *a* indique la main active; *b* indique la main servant de base non morphémique; *c* indique la main servant comme classificateur; et 1 et 2 sont utilisés pour indiquer les mains dominante et non dominante, respectivement.

Nous utilisons les trois points (...) pour indiquer la tenue d'un signe ou partie d'un signe, ou d'un autre comportement précédent pendant l'enclenchement d'autres signes ou comportements. Ceci permet de rendre compte de la portée temporelle des signes et des autres comportements et éviter que chaque occurrence séquentielle de la même glose soit interprétée comme une occurrence indépendante du même signe ou comportement. Cette convientionn tient la place de la ligne horizontale couramment utilisée dans la littérature sur les langues signées.

Nous transcrivons sur plusieurs lignes, de façon détaillée, les expressions faciales (haut et bas du visage; bouche); nous laissons aussi ouvert, si nécessaire, la possibilité de noter le

Étant donné le but que nous nous sommes fixé, à savoir d'utiliser le corpus informatisé comme outil de recherche aux différents niveaux de la structure, il faut structurer la transcription des données d'une façon maximalement transparente, tout en représentant sans ambiguïté les aspects de la structure qui pourraient avoir une influence sur un phénomène donné. Nous rencontrons un problème important au niveau des gloses utilisées. Lacerte (ce volume) décrit la problématique des gloses au niveau sémantique; outre cette difficulté, nous devons tenir compte du fait qu'une glose en français peut souvent obscurcir la structure morphologique d'un signe, surtout lorsque celle-ci se réalise sur les deux mains, par exemple. Le signe normalement glosé RENCONTRER illustre bien ce problème. Il est, en effet, composé de deux morphèmes classificateurs qui peuvent très bien agir indépendamment ailleurs dans une phrase.

Transcription dans Microsoft Excel

Après avoir considéré plusieurs moyens possibles d'entrer les données transcrites dans l'ordinateur, nous avons décidé d'utiliser le chiffreur Microsoft Excel. Ce logiciel permet d'entrer des données sur plusieurs lignes superposées, autant que nécessaire pour inclure toutes les informations qui pourraient être utiles pour l'analyse linguistique. L'entrée de données sur une table de calcul permet, en particulier, d'ajouter des informations manquantes à un stade ultérieur de révision sans perturber l'alignement des éléments sur différentes lignes de la transcription. De plus, il est possible d'entrer les données soit en format horizontal/en lignes parallèles (ce qui en facilite la lecture), soit en format vertical/en colonnes (ce qui est nécessaire pour convertir le corpus en base de données afin d'effectuer des recherches ainsi que pour stocker les transcriptions continues de plus de 256 éléments successifs sur une ligne). Les fonctions de copiage et collage du logiciel permettent de transposer facilement les données entre les deux formats.

Une fois un corpus entré sur une table de calcul en format vertical, il peut être converti en base de données sur laquelle il est possible de faire des recherches automatisées de plusieurs sortes, des tris et certains calculs statistiques. Ces bases de données peuvent également être transportées à un stade ultérieur dans un fichier créé par un logiciel de traitement de bases de données tel que Microsoft FoxPro qui permet des traitements plus sophistiqués tels que des enquêtes relationnelles portant sur plusieurs bases de données en même temps.

La flexibilité d'un logiciel chiffreur permet de transcrire les données d'une façon qui serait moins facile dans un simple logiciel de traitement de texte. Ainsi, il est possible, au fur et à mesure, de séparer certaines informations sur des lignes distinctes plutôt que de les inclure

sens global d'une expression faciale sur une autre ligne. Nous considérons la possibilité d'adopter cette convention pour encoder simultanément le sens lexical d'un signe ainsi que sa structure morphologique.

Finalement, nous plaçons la ligne «regard» en bas de la transcription, près des comportements manuels pour faciliter la comparaison de la direction du regard et la localisation spatiale des signes.

Annexe: Transcription partielle d'un signeur du corpus LSQ88

a)

temps							
tête							
corps							
haut du visage							
bas du visage							
expression globale			surpris et effrayé	...surpris et effrayé			
bouche							
main droite (gauche de l'écran)	POINTÉ		PRENDRE-PAR-LA-JOUE	C'EST-ÇA	PRENDRE-PAR-LA-JOUE		...PRENDRE-PAR-LA-JOUE
coordonnées spatiales	index			g			
référence du locus/vecteur							
main gauche (droite de l'écran)	UN		~POINTÉ	...~POINTÉ	...~POINTÉ		POINTÉ
coordonnées spatiales			1ère	1ère	1ère		nc1
référence du locus/vecteur							
regard	mains	bat	hc1	interloc. (ML?)	bg1	bat	nc1

b)

temps							
tête							hoch. horiz.
corps							
haut du visage							
bas du visage							
expression globale							
bouche							
main droite (gauche de l'écran)	...PRENDRE-PAR-LA-JOUE		(inc.)1	"QUOI"1		SIGNER1	
coordonnées spatiales							
référence du locus/vecteur							
main gauche (droite de l'écran)	SE-SOUVENIR	C'EST-ÇA	(inc.)2	"QUOI"2		SIGNER2	
coordonnées spatiales	...nc1						
référence du locus/vecteur							
regard	...nc1	interloc. (ML?)	interloc. (ML?)	interloc. (ML?)	bat	nc2	

c)

temps										
tête	hoch. horiz. g-c	neutre vers bas								
corps										
haut du visage				sourcils relevés						
bas du visage										
expression globale										
bouche										
main droite (gauche de l'écran)	PAS-ENCORE1	C'EST-ÇA		PROCHAINE-CHOSE1		POINTÉ	UN	MOT		POINTÉ
coordonnées spatiales						majeur				1ère
référence du locus/vecteur										
main gauche (droite de l'écran)	PAS-ENCORE2	...PAS-ENCORE2		PROCHAINE-CHOSE2		DEUX				
coordonnées spatiales										
référence du locus/vecteur										
regard	interloc. (ML?)		bat		bat	bc2	...bc2	...bc2	bat	...bc2

d)

temps										
tête					hoch. vert. ++	balancement g-à-d				
corps										
haut du visage										
bas du visage				joue relevée						
expression globale										
bouche				mm	ouverte		six		ans	six ans
main droite (gauche de l'écran)	TAQUINERa	POINTÉ		SOURD	ENFANT	(AGE(esc.?))	SIX		ALENTOUR	SIX
coordonnées spatiales	1-à-2	bc2								
référence du locus/vecteur										
main gauche (droite de l'écran)	TAQUINERb			~POINTÉ						
coordonnées spatiales	1-à-2									
référence du locus/vecteur										
regard	...bc2	...bc2	bat	...bc2	...bc2	...hc2	...hc2	bat	...hc2	...hc2

e)

temps								
tête								
corps						c-g	g-c	
haut du visage	sourcils relevés							
bas du visage			joue relevée					
expression globale								
bouche			mm		moqué			
main droite (gauche de l'écran)		POINTÉ	TAQUINERa	POINTÉ	MOQUER1	PETIT-PEU	MOQUER1	OUBLIER
coordonnées spatiales		1ère	1-à-2	1ère	1-bc2		1-bc2	
référence du locus/vecteur								
main gauche (droite de l'écran)			TAQUINERb		MOQUER2	...MOQUER2	MOQUER2	POINTÉ
coordonnées spatiales			1-à-2		1-bc2		1-bc2	bc2
référence du locus/vecteur								
regard	bat	bc2	...bc2	...bc2	...bc2	interloc. (ML)	...interloc. (ML)	

f)

temps									
tête									
corps									
haut du visage									
bas du visage									
expression globale									
bouche	ou		...ou					boump (10 fois)	
main droite (gauche de l'écran)	POINTÉ		FACHÉ		POUSSER-FORT1		POINTÉ	CI-V:-TOMBER	
coordonnées spatiales	bd2				bd2		1ère	en bas de l'escalier spiral(10x)	
référence du locus/vecteur									
main gauche (droite de l'écran)	...POINTÉ		...POINTÉ		POUSSER-FORT2				
coordonnées spatiales	...bc2		...bc2		bd2				
référence du locus/vecteur									
regard	OO, bc2	bat	ng2 (ML)	bat	maindom.	bat	ng2 (ML)	main dominante, Bc2, ng2 (ML), Bc2	bat

g)

temps		
tête		
corps		c-d
haut du visage		
bas du visage		
expression globale	pleurer	...pleurer
bouche		
main droite (gauche de l'écran)	POINTÉ	PLEURER (emphatique)1
coordonnées spatiales	1ère	
référence du locus/vecteur		
main gauche (droite de l'écran)		PLEURER (emphatique)2
coordonnées spatiales		
référence du locus/vecteur		
regard	nd2 (JB)	nd2 (JB)