Problèmes et méthodes pour l'analyse d'énoncés en LSF

Balvet Antonio (1), Sallandre Marie Anne (2)

(1) UMR 8528 Silex – Université Lille 3
Université Lille 3, BP 60149 - 59653 Villeneuve d'Ascq Cedex
antonio.balvet@univ-lille3.fr

(2) UMR 7023 SFL – Université Paris 8
15 rue Catulienne, 93200 Saint-Denis
sallandre@yahoo.com

Mots-clés : LSF, transcription, traitement automatique des langues, linguistique de corpus
Keywords: LSF, transcription, analysis, natural language processing, corpus linguistics

Résumé Après avoir rappelé les principales caractéristiques de la construction des énoncés en LSF, nous abordons les questions liées à leur traitement automatique. Nous tentons ainsi de préciser dans quelle mesure les outils et méthodes aujourd'hui disponibles pour la modélisation et l'analyse automatique des langues naturelles apparaissent compatibles avec le cadre théorique et méthodologique ici adopté, centré sur les Structures de Grande Iconicité.

Abstract In this paper, we address the issue of the formal treatment of French Sign Language, in a functionalist and enunciativist framework. We advocate for a copus-driven approach and we emphasize the need for alternative frameworks for NLP, alongside classical generativist symbolic-computational ones.

1 Introduction

La LSF pose de nombreux défis tant pour sa description que pour sa modélisation, y compris par des moyens automatiques. En effet, cette langue construit l'espace comme un système linguistique, ce qui implique de passer par des transcriptions réalisées par des linguistes disposant d'une bonne connaissance de la langue. Du point de vue descriptif, l'approche de type énonciative et cognitive, retenue par (Cuxac, 2000) et (Sallandre, 2003), qui constituera la base des travaux ici exposés, semble difficilement conciliable avec les approches classiques en TALN, centrées sur l'élaboration de systèmes de règles formelles. En effet, le fonctionnement synthétique[1] et la nécessité d'intégrer le contexte élargi[2], mis en évidence par

[1] Intégration de plusieurs paramètres : signes manuels, mimiques faciales, posture du corps.

[2] Co-texte, mais également situation d'énonciation, attentes des interlocuteurs.

ces travaux, apparaissent comme une source de difficultés majeures pour toute tentative d'analyse automatique des énoncés en LSF par le biais de règles symboliques. Dans cet exposé, nous examinons, tout d'abord, les conditions d'un traitement formel des énoncés en LSF. Ensuite, nous présentons quelques outils et résultats d'exploration de corpus de transcription, afin d'en faire émerger des régularités au niveau syntagmatique.

2 Exploration de corpus de transcriptions en LSF pour le TALS

2.1 Quelques propriétés saillantes de la construction des énoncés en LSF

Le modèle proposé par (Cuxac, 2000) se situe dans une perspective sémiogénétique et considère l'iconicité (référentielle) non comme seul outil de description de la langue mais comme principe organisateur. Une bifurcation fonctionnelle a ainsi été postulée et détermine deux pôles entre lesquels le va-et-vient est constant. Elle se compose d'une part des structures de grande iconicité (SGI), qui *donne à voir tout en disant*, d'autre part des signes standard (SS), sans visée illustrative, qui *disent* seulement. Les SGI se structurent à partir des trois principaux transferts : transferts de taille et de forme, transferts de situation et transferts de personne. Ces SGI, contrairement aux signes standard[3], ont peu fait l'objet de descriptions formelles et encore moins de modélisations.

Le modèle de Cuxac a été confronté à un corpus de référence : LS-COLIN[4], présenté dans (Sallandre, 2003), constitué de trois discours de genres variés, sur une durée totale d'une heure et cinq minutes. L'analyse de ce corpus montre que plus des 2/3 des unités sont effectués avec une visée illustrative dans les deux narrations (3/4 pour le premier récit, 2/3 pour le deuxième) et environ 1/3 dans le genre explicatif. En raison de son importance, de la diversité des genres qui le composent et du nombre de locuteurs enregistrés, ce corpus constituera la base des observations et réflexions consignées dans le présent exposé, centré par conséquent sur les SGI plus que sur les SS.

3 Quels fondements méthodologiques et théoriques pour le TALS ?

Les questions habituellement traitées en TALN sont centrées sur les combinaisons possibles et impossibles de suites de symboles (mots, constituants), déterminées par l'application systématique de règles logiques. Or, le cadre ici choisi pour traiter des énoncés en LSF met l'accent sur des paramètres énonciatifs, ainsi que sur l'importance du contexte, tant dans l'exercice de la compétence linguistique des locuteurs que dans la pratique de transcription. Une accommodation est donc nécessaire entre les présupposés théoriques sous-jacents au TALN d'inspiration générativiste et le modèle de Cuxac.

---

[3] Voir (Lejeune, 2004) pour une analyse d'énoncés utilisant la Grammaire Applicative et Cognitive de Desclés.

[4] Action Cognitique 2000, LACO 39 (Université Paris 8, LIMSI-CNRS, IRIT), 39 discours, 13 locuteurs adultes.

3.1 Quelques difficultés pratiques et théoriques pour le TALS

Le degré de grammaticalité des phrases constitue habituellement le fondement de toute description formelle en TALN. Or, en LSF, la notion syntaxique de phrase, ainsi que celle de mot (typographique), paraissent peu pertinentes. Reste la question des conditions de bonne formation des énoncés, pour une langue dont la syntaxe semble guidée essentiellement par des principes d'optimisation de contraintes cognitives[5]. Il convient donc de déterminer à quelles conditions un énoncé sera refusé par un locuteur natif de la LSF. La méthode traditionnelle d'investigation en la matière, basée essentiellement sur l'introspection et l'intuition linguistique du chercheur, apparaît ici insuffisante, étant donné l'état actuel des connaissances. Un début de solution apparaît passer l'étude de corpus d'énoncés attestés, tel que LS-COLIN.

Par ailleurs, en TALN, l'un des prérequis de toute modélisation est la caractérisation des énoncés, dans les termes des grammaires formelles définies par N. Chomsky, caractérisées par deux ensembles d'éléments (auxquels s'ajoutent les règles de réécriture) : le Vocabulaire Non Terminal (i.e. catégories syntaxiques) constitué d'un nombre fini d'éléments, et le Vocabulaire Terminal (i.e. unités lexicales) constitué d'un nombre potentiellement infini d'éléments. Une des premières questions à résoudre pour la LSF serait donc celle du VNT. En effet, en LSF, l'identification des unités n'est souvent possible qu'en contexte, et en adoptant un point de vue fonctionnel. De ce fait, il semble difficile de proposer un système de conditions nécessaires et suffisantes pour identifier les éléments du VNT, en-dehors des catégories de grande iconicité dont une taxinomie est fournie dans (Sallandre, 2003). Ainsi, dans l'approche adoptée ici, il est possible que les éléments candidats pour un VNT de la LSF ne constituent pas une liste fermée. De plus, dans le cadre adopté ici, les problèmes centraux du sens, du contexte et de la multilinéarité, restent encore insuffisamment formalisés en TALN.

Ceci nous incite à aborder le problème du TALS d'abord par le biais des contraintes de combinaison des unités appartenant aux deux grands pôles de la bifurcation posée par Cuxac : SS et SGI. La caractérisation formelle des énoncés en LSF, par exemple sous la forme de grammaires hors contexte probabilistes, nous paraît pouvoir passer par une accumulation de données quantitatives et systématiques sur les possibilités de choix d'unités appartenant à chacun des pôles SS ou SGI. Les observations ici faites militent donc pour une exploration outillée de corpus attestés, comme préalable à toute tentative de TALS.

4 Méthodes et outils pour l'exploration de corpus en LSF

Nous présentons ici quelques résultats préliminaires de l'exploration de corpus de transcription d'énoncés en LSF tirés de LS-COLIN, en vue d'en faire émerger des régularités au niveau syntagmatique. Nous déterminons quels paramètres semblent les plus adaptés dans la perspective ici adoptée, puis nous discutons quelques résultats d'exploration de corpus de transcription, à l'aide des outils évoqués (i.e. n-grammes, concordances).

---

[5] Ex. : ordre privilégié Contenant>Contenu. Pour plus de détails, voir également (Lejeune, 2004).

4.1 Analyse de concordances tirées des corpus de transcription

Les observations relatées ici reposent sur l'analyse de concordances réalisées sur le champ « étiquetage linéaire » des transcriptions tirées de (Sallandre, 2003). En effet, ce champ contient l'ensemble des unités identifiées, qu'elles soient du type standard ou non, il constitue donc le lieu d'observation privilégié pour les alternances entre unités standard et non standard. Les concordances permettent l'analyse des contraintes distributionnelles des unités linguistiques, elles peuvent être réalisées grâce à une base de 5-grammes construits à partir de la fusion des étiquettes linéaires de toutes les transcriptions, production par production, de LS-COLIN. La base de données ainsi constituée permet, par exemple, d'aborder la description synthétique des contraintes s'exerçant sur les unités et segments pertinents ci-dessous.

4.1.1 Début et fin de récit (Histoire du Cheval)

Pour l'*Histoire du Cheval*, seuls 2 locuteurs sur 13 utilisent des unités de la catégorie transferts (TTF) dès le début de leur production. Pour ce récit, 11 locuteurs sur 13 suivent un schéma narratif d'introduction du thème par des SS. Ce comportement va dans le même sens que l'intuition des transcripteurs, qui ont isolé le schéma : [introduction de thème en SS] suivi de [SGI]. L'examen des contextes immédiats de la marque de fin de récit montre que seuls 3 locuteurs sur 13 terminent leur récit par des unités appartenant aux SGI.

4.1.2 Contextes gauche et droit de Signe Standard (Histoire du Cheval)

Dans la plupart des cas, les SS sont enchaînés en séquences de plusieurs unités, puis, dans l'ordre décroissant, les SGI (TP, TTF, etc.), puis les Pointages. Un SS semble être avant tout un introducteur d'un autre SS, puis d'une SGI (i.e. TP, TTF, TS, et d'autres sous-types). Par comparaison entre les contextes gauche et droit, on peut souligner que, pour le récit concerné, les SS sont plus souvent introduits par un pointage qu'ils n'introduisent un pointage.

5 Conclusion

Comme nous l'avons évoqué dans le présent exposé, le TALS se heurte à des difficultés, tant pour la description que pour la modélisation des énoncés. Ces difficultés sont essentiellement liées à la dynamique énonciative, aux problèmes d'interprétation en contexte et à l'intégration synchronisée de plusieurs niveaux linguistiques à l'œuvre en LSF. Une des avancées que nous attendons de l'interaction entre linguistique formelle, linguistique de corpus et LSF est, précisément, de poser la nécessité d'un réel traitement du contexte, de la sémantique et de la dynamique conversationnelle tant pour les LS que pour les autres LN, qui ne peut, à notre sens, que passer par un travail sur corpus attestés et par l'exploration de nouveaux cadres formels non nécessairement logiques, telles que les Grammaires de Construction Radicales de (Croft, 2001) ou la Théorie de l'Optimalité (Prince & Smolensky, 1993). Nous proposons donc, comme étape préalable à un TALS la mise en œuvre et le développement d'outils adaptés pour le travail d'exploration de corpus de transcription en LSF, en ayant en tête l'élargissement à d'autres langues, voire d'autres domaines d'exploration. Une partie de ces outils est en cours d'élaboration et de validation, ils ont vocation à être intégrés à la plate-

forme CoPT[6] (Corpus Processing Tools) dont la constitution a été, en partie, dictée par les explorations sur corpus de transcription en LSF ici relatées.

Références

CROFT W. (2001), *Radical Construction Grammar*, Oxford University Press.

CUXAC C. (1996), *Fonctions et structures de l'iconicité des langues des signes*, thèse de Doctorat d'Etat, Université Paris V.

CUXAC C. (2000), La Langue des Signes Française; les Voies de l'Iconicité, *Faits de Langues* n°15-16, Paris: Ophrys.

LEJEUNE F. (2004), *Analyse sémantico-cognitive d'énoncés en Langue des Signes Française pour une génération automatique de séquences gestuelles*, thèse de doctorat, Université Paris XI.

PRINCE A., SMOLENSKY P. (1993), *Optimality Theory, Constraint interaction in generative grammar*, Technical Report, ROA.

SALLANDRE M.-A. (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 VIII.

[6] Voir le descriptif de la plate-forme, ainsi que quelques modules à l'adresse : http://copt.sourceforge.net.