Une grammaire formelle pour les langues des signes basée sur AZee : une proposition établie sur une étude de corpus

Camille Challant, Michael Filhol
Université Paris-Saclay, CNRS, LISN, 91400 Orsay, France
{camille.challant, michael.filhol}@lisn.upsaclay.fr

RÉSUMÉ

Cet article propose de premières réflexions quant à l'élaboration d'une grammaire formelle pour les langues des signes, basée sur l'approche AZee. Nous avons mené une étude statistique sur un corpus d'expressions AZee, qui décrivent des discours en langue des signes française. Cela nous permet d'entrevoir des contraintes sur ces expressions, qui reflètent plus généralement les contraintes de la langue des signes française. Nous présentons quelques contraintes et positionnons théoriquement notre ébauche de grammaire au sein des différentes grammaires formelles existantes.

ABSTRACT

A formal grammar for sign languages based on AZee : a proposal established on a corpus study

This article provides some initial thoughts about the development of a formal grammar for sign languages, based on the AZee approach. We have conducted a statistical study on a corpus of AZee expressions, which describe French sign language discourses. This allows us to glimpse some constraints on these expressions, which reflect more generally the constraints of French sign language. We present some constraints and theoretically position this draft grammar among the various existing formal grammars.

MOTS-CLÉS : AZee, langue des signes française, modélisation, grammaire formelle.

KEYWORDS: AZee, French sign language, modeling, formal grammar.

1 Introduction

Les langues des signes (LS) sont des langues visuo-gestuelles, qui se distinguent des langues audio-vocales sur de nombreux points. Les mains et les doigts mais aussi le buste, le regard, les sourcils sont autant d'articulateurs qui entrent en jeu dans les LS, et s'animent dans l'espace de signation pour produire des énoncés. Ces différents articulateurs rendent possible la multilinéarité, c'est-à-dire le fait de pouvoir réaliser plusieurs choses simultanément : les LS ne sont donc pas nécessairement des séquences de signes placés les uns à la suite des autres. Ces caractéristiques (spatialisation, multilinéarité) deviennent de véritables défis lorsque l'on s'intéresse à la modélisation des LS, qui est nécessaire en traitement automatique des langues pour des tâches de synthèse, de reconnaissance ou encore de traduction automatique. Le modèle formel sur lequel nous travaillons, nommé AZee, permet de représenter des discours en LS en tenant compte des spécificités qui viennent d'être évoquées.

Nous proposons, dans cet article, de premières réflexions quant à l'élaboration d'une grammaire

formelle pour la langue des signes française (LSF), établie à partir de l'étude d'un corpus décrit à l'aide d'AZee. Nous entendons par grammaire formelle un système de règles permettant de décrire la langue et de juger si un énoncé répond aux contraintes de celle-ci, système ne laissant aucune place à l'interprétation humaine et pouvant être utilisé par des programmes informatiques.

Nous commençons par présenter le modèle AZee et notre question de recherche, avant d'exposer notre méthode ainsi que nos premiers résultats. Nous comparons dans une dernière section notre potentielle grammaire AZee avec les différentes grammaires formelles existantes.

2 AZee

AZee (Filhol *et al.*, 2014) est une approche de description des LS fondée sur la notion essentielle de *règle de production*, qui associe à un sens identifié un ensemble de formes observables à produire. Cela permet de ne faire aucun présupposé concernant l'existence de niveaux linguistiques, de catégories grammaticales ou d'un ordre séquentiel.

Par exemple, en LSF, la règle de production `président` associe le sens 'président/présidence' à la forme illustrée en figure 1a. De même, la règle `info-about`, à deux arguments (*topic* et *info*), associe le sens '*info*, à propos de *topic*' à la synchronisation de formes présentée en figure 1b : les deux arguments sont placés en séquence et séparés par une durée contrôlée, chacun est maintenu (*hold*) sur une durée plus ou moins longue et un clignement des yeux (*el :cl*) se synchronise avec la fin du maintien d'*info*. Les différentes règles de production peuvent se combiner entre elles récursivement pour construire des *expressions AZee de discours*, qui reflètent le sens que l'on interprète à partir des formes qu'elles produisent. Un exemple d'une telle expression est donné en figure 1c : elle représente une production en LSF dont les formes correspondent à celles de la figure 1b, avec les règles `président` et `célèbre` respectivement en *topic* et *info*, et signifie « [le/un] président est célèbre ».

(a) forme pour "président/présidence" en LSF

(b) synchronisation de formes pour la règle `info-about`

```
:info-about
    'topic
    :président
    'info
    :célèbre
```

(c) Expression AZee de discours signifiant "le/un président est célèbre"

FIGURE 1 – Présentation d'AZee

Il est possible, de cette manière, de décrire de vraies productions en LSF avec AZee, ce que nous avons fait avec les 120 discours composant le corpus des "40 brèves" (40 entrées journalistiques en français écrit, chacune traduite en LSF par 3 traducteurs sourds) (Filhol & Tannier, 2014). Un corpus de 120 expressions AZee de discours représentant une heure de LSF au total est ainsi disponible, et

comporte 11 470 applications de règles de production (Challant & Filhol, 2022). Nous avons décidé de travailler sur ce corpus.

3 Question de recherche et méthode

Si une expression AZee modélise en effet une production langagière en LSF, alors toutes les contraintes qui pèsent sur la langue devraient être reflétées dans les expressions AZee de discours. Nous faisons donc l'hypothèse qu'il existe des contraintes qui régissent ces expressions. Si cette hypothèse est vérifiée, nous pourrons considérer ces contraintes comme des contraintes grammaticales de la LSF. Nous reviendrons sur cette considération plus en détails dans la section 5.

Plusieurs questions se posent alors pour identifier des contraintes sur les expressions AZee de discours : les différentes règles de production apparaissent-elles dans les mêmes contextes ? Leurs arguments sont-ils contraints dans leur complexité et si oui, comment ? Certaines règles de production sont-elles plus fréquentes que d'autres ? Afin d'être en mesure de répondre à ces questions, nous avons imaginé plusieurs tests, applicables par exemple, pour toute expression ou sous-expression E, à :

- $rootname(E)$: règle de production appliquée à la racine de E, identifiée par son nom ;
- $prodcount(E)$: nombre d'applications de règles de production dans E ;
- $contains\text{-}rule(E, R)$: vrai si et seulement si il existe une sous-expression E' dans E telle que $rootname(E') = R$ avec R le nom d'une règle de production ;
- $E.arg$: sous-expression utilisée comme argument arg de E (E est l'application d'une règle définissant arg comme nom d'un de ses arguments).

Ces tests sont combinables à l'aide d'opérateurs booléens afin de créer des requêtes plus complexes sur les expressions.

Nous avons ensuite appliqué ces tests sur le corpus d'expressions AZee présenté dans la section précédente. Nous présentons les premiers résultats que nous avons obtenus suite à cette étude dans la section suivante.

4 Premiers résultats

Dans un premier temps, nous pouvons constater que les règles de production à la racine des 120 expressions ne sont que des `context` (95 occurrences) et des `info-about` (25 occurrences). Il semblerait donc que la racine d'une expression de discours de genre journalistique soit plutôt contrainte. Nous pouvons formuler une hypothèse de contrainte sur toute expression de discours E :

$$rootname(E) \in \{\texttt{context}, \texttt{info-about}\}$$

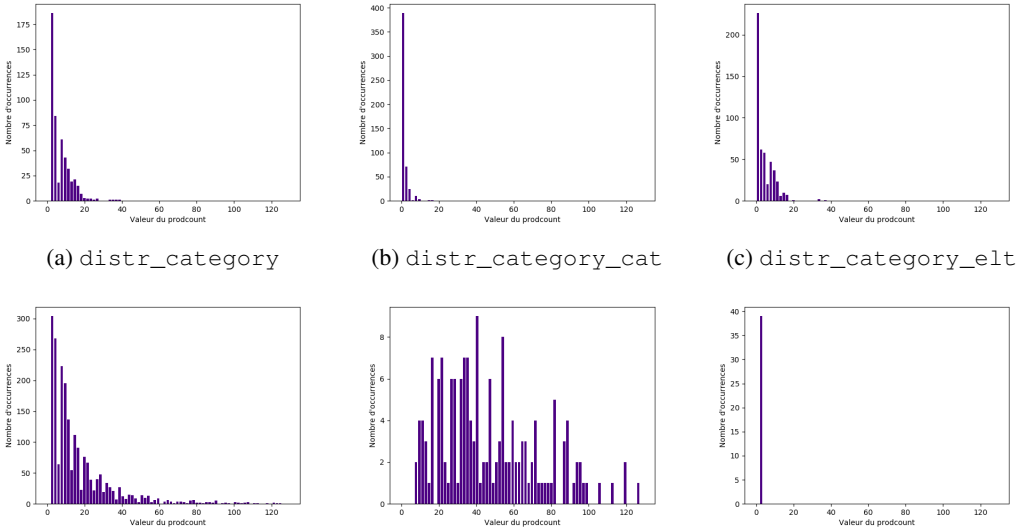
Dans un deuxième temps, nous nous sommes intéressés au poids des constituants, en mesurant leur $prodcount$, qui compte le nombre d'applications de règles de production contenues dans les expressions, ce qui reflète leur complexité.

Nous avons, pour chaque règle de production R connue, établi :

- la distribution $distr\_R$ des $prodcount(E)$ sur l'ensemble des expressions ou sous-expressions E du corpus telles que $rootname(E) = R$

— pour chaque argument arg défini par R, la distribution $distr\_R\_arg$ des $prodcount(E.arg)$ sur l'ensemble des expressions E telles que $rootname(E) = R$

Nous avons obtenu des distributions, dont nous montrons quelques exemples en figure 2. Celles-ci sont très différentes les unes des autres : certaines règles semblent présenter une limite maximale (comme ici `category` ou `info-about`), d'autres une limite minimale (`prise-de-parole`) et d'autres encore ont une valeur de *prodcount* qui est fixe (`tens-unit`). Nous avons décidé de nous concentrer sur les règles de production que nous venons de citer, car elles présentent des contrastes intéressants. Nous les présentons succinctement dans le tableau 1.



(a) `distr_category` (b) `distr_category_cat` (c) `distr_category_elt`

(d) `distr_info-about` (e) `distr_prise-de-parole` (f) `distr_tens-units`

FIGURE 2 – Distribution du *prodcount* de différentes règles de production

Nom de la règle	`category`	`prise-de-parole`	`tens-units`
Arguments	*cat, elt*	*sig*	*tens, units*
Nombre d'occurrences dans le corpus	500	165	44
Sens	*elt*, interprété comme une instance de *cat*	pause rhétorique avant *sig*	nombre formé par deux chiffres, *tens* et *units*

TABLE 1 – Présentation de trois règles de production

On remarque, sur la figure 2a, qu'il n'y a pas de prodcount supérieur à 20, ce qui montre une limite maximale pour les expressions dont `category` est à la racine. On remarque la même limite sur la figure 2c, c'est-à-dire pour les expressions utilisées comme argument *elt* de `category`. En revanche, cette limite est de 10 sur la figure 2b ce qui signifie que l'argument *cat* semble davantage limité en complexité que *elt*. Un contraste similaire est observé entre les arguments *topic* et *info* de la règle `info-about`, la limite se dégageant pour *info* (80) étant le double de celle pour *topic* (40).

Pour `prise-de-parole`, au contraire, aucune limite maximale ne se dégage (figure 2e), mais toutes les valeurs sont supérieures à 7. Ainsi, la limite semble être minimale : on observe un comportement inverse au précédent.

Enfin, un troisième comportement se dégage avec `tens-units`, comme l'illustre la figure 2f : sur les 44 occurrences du corpus, toutes les expressions ayant pour racine cette règle ont un $prodcount$ de 3. Plus précisément même, par argument, si $rootname(E) = \texttt{tens-units}$ alors :

$$prodcount(E.tens) = 1, prodcount(E.units) = 1$$

Dans un troisième temps, nous nous sommes penchés sur les règles pouvant être contenues dans la descendance d'une règle de production. Nous avons considéré l'ensemble $\mathcal{R}$ des règles ayant au moins un argument obligatoire et présentant plus de 20 occurrences dans le corpus, soit 19 règles de production.

Nous avons, pour chaque couple R_1, R_2 de règles de $\mathcal{R}$, compté le nombre de fois où R_2 apparaît dans R_1, ainsi que dans chacun de ses arguments pris séparément, c'est-à-dire :

- le nombre d'expressions E telles que $rootname(E) = R_1$ ET $contains\text{-}rule(E, R_2)$
- pour chaque argument arg défini par R_1, le nombre d'expressions E telles que $rootname(E) = R_1$ ET $contains\text{-}rule(E.arg, R_2)$

Dans le cas où $R_1 = \texttt{category}$, nous avons remarqué que 9 règles de production n'étaient jamais contenues dans l'argument *cat* alors qu'elles ne sont que 3 à n'être jamais contenues dans *elt*, et deux qui ne sont ni contenues dans *elt*, ni dans *cat* : `prise-de-parole` et `context`. En revanche, il n'y a pas de telle contrainte concernant `prise-de-parole` et `info-about` : toutes les règles de production de $\mathcal{R}$ apparaissent dans les expressions dont elles sont les racines. Pour finir, aucune règle de production de $\mathcal{R}$ n'est contenue dans les expressions ayant pour racine `tens-units`. Toutes les règles de production ne sont donc pas contraintes de la même façon sur leur descendance.

Enfin, dans un dernier temps, les règles `category`, `tens-units`, `info-about` peuvent être observées à la racine de n'importe quel argument d'une règle de production de l'ensemble $\mathcal{R}$. Au contraire, `prise-de-parole` ne se trouve à la racine que d'arguments de trois règles : `context`, `info-about` et `each-of`.

Pour conclure, à partir des observations réalisées sur notre corpus, nous avons pu identifier des contraintes sur les expressions AZee de discours. On remarque, par exemple, des règles de production qui peuvent accepter des arguments très complexes tandis que cela n'est jamais observé chez d'autres, certaines règles de production peuvent contenir n'importe quelle autre règle de production là où d'autres sont contraintes à ce niveau.

5 Discussion et positionnement théorique

Nous venons, dans la section précédente, d'identifier plusieurs contraintes auxquelles les différentes règles de production peuvent être soumises. Selon nous, et comme mentionné plus haut, une contrainte sur une expression AZee peut être perçue comme une contrainte grammaticale, qui gouverne la combinaison, la taille, la position ou encore la fréquence d'apparition des différentes unités de la langue. Ces contraintes formelles composent ensemble un système, qui peut être considéré comme une grammaire.

Ainsi construite, notre grammaire comporterait plusieurs caractéristiques qui la distingueraient des autres grammaires, à commencer par son absence de présupposés, sur plusieurs plans. Tout d'abord, la séquence n'est pas admise d'office : l'ordre linéaire dans lequel certains éléments apparaissent dans la forme produite par une expression est simplement le résultat de l'application de règles de production combinées. Les niveaux linguistiques ne sont pas non plus présupposés, pas plus que les catégories syntaxiques. Nous revenons ainsi à quelque chose de plus fondamental, sans pour autant nier l'existence de ces notions en LSF : il s'agit simplement là de remettre en question leur caractère fondamental. Si la séquence, les niveaux linguistiques ou les catégories syntaxiques sont des notions pertinentes pour décrire les LS, elles peuvent être définies à partir de critères formels plus fondamentaux, et sont donc en réalité émergentes. Par exemple, en nous appuyant sur les résultats présentés dans la section 4, nous pourrions définir des classes de règles qui répondent aux mêmes contraintes sur leur nombre d'arguments ou encore sur le poids de leurs arguments, Ensuite, notre grammaire a la caractéristique de porter de la sémantique à tous les niveaux de l'expression : chaque noeud est porteur de sens, quelle que soit sa position dans l'expression. Enfin, chaque expression AZee détermine des formes à produire, en prenant en compte les articulateurs manuels comme non manuels : aucune hiérarchie n'est présumée entre ces différents articulateurs.

Ces caractéristiques opposent notre approche aux grammaires génératives (Chomsky, 1965), centrées sur la syntaxe et basées sur un ordre séquentiel, qui ont sont encore très utilisées pour décrire les LS (Aristodemo & Hauser, 2021; Kimmelman & Pfau, 2021; Napoli & Sutton-Spence, 2014). Les grammaires génératives s'intéressent notamment à l'ordre des unités lexicales dans le discours, chacune étant pré-étiquetée avec une catégorie syntaxique. Les règles s'appliquant à ces catégories et aux nœuds dans les arbres syntaxiques (VP, NP, etc.) opèrent sans recours au niveau sémantique.

En revanche, plusieurs idées appartenant à d'autres approches formalistes retiennent notre attention, bien que la notion de catégorie syntaxique reste au cœur de toutes celles-ci. Par exemple, les grammaires cognitives (Martínez *et al.*, 2020; Langacker, 1987) accordent une grande place à la sémantique, ce qui est essentiel dans l'approche AZee. De plus, nous pouvons apercevoir quelques points communs entre notre grammaire et les grammaires de construction (Beuls & Van Eecke, 2023; van Trijp, 2015; Fillmore, 1988), qui prennent également pour base l'association forme-sens présente dans les langues. Ces grammaires mettent en avant un continuum lexique-syntaxe et des niveaux linguistiques qui ne sont pas clairement distingués, ce qui fait écho à notre approche. Enfin, les grammaires de propriétés (Blache, 2001) nous intéressent particulièrement : les propriétés sont des contraintes, et ce système de contraintes permet de caractériser les énoncés grâce à un gradient de grammaticalité (plutôt qu'un jugement binaire) dont la valeur peut être déterminée par la nombre de contraintes satisfaites, ce qui nous semble tout à fait approprié pour les langues orales que sont les LS.

6 Conclusion et perspectives

Pour conclure, nous avons présenté dans cet article les premières bases d'une grammaire formelle fondée sur AZee. Nous avons également positionné cette ébauche de grammaire au sein des grammaires formelles existantes. Une de nos perspectives à court terme est d'expérimenter l'extraction automatique de motifs réguliers de notre corpus, à l'instar de Herrera *et al.* (2022). Nous aimerions pouvoir adapter leurs méthodes et outils à nos données. De plus, nous souhaiterions augmenter notre corpus d'étude, tout en diversifiant le genre de discours décrits. Nous envisageons de décrire avec AZee le corpus Mocap1 (Benchiheub *et al.*, 2016), qui comporte un grand nombre de structures

dites iconiques, structures propres aux LS et reconnues comme très fréquentes. Nous pourrons ainsi l'explorer avec notre méthodologie et comparer les résultats obtenus avec ceux de notre corpus actuel. Enfin, une de nos perspectives est de générer avec un avatar (puisque cela est possible avec AZee !) des discours qui répondent – ou non – à nos contraintes, et de les présenter à des locuteurs natifs de la LSF afin de confirmer nos hypothèses.

Références

ARISTODEMO V. & HAUSER C. (2021). Similar but Different : Investigating Temporal Constructions in Sign Language. *Glossa : a Journal of General Linguistics 6(1) : 2*, **6**. DOI : 10.5334/gjgl.999.

BENCHIHEUB M.-E.-F., BERRET B. & BRAFFORT A. (2016). Collecting and Analysing a Motion-Capture Corpus of French Sign Language. In *7th International Conference on Language Resources and Evaluation - Workshop on the Representation and Processing of Sign Languages (LREC-WRPSL 2016)*, p. 7–12, May 23-28, Portoroz, Slovenia. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr, v1, https ://hdl.handle.net/11403/mocap1/v1.

BEULS K. & VAN EECKE P. (2023). Fluid Construction Grammar : State of the Art and Future Outlook. In *Proceedings of the First International Workshop on Construction Grammars and NLP (CxGs+NLP, GURT/SyntaxFest 2023)*, p. 41–50, Washington, D.C. : Association for Computational Linguistics.

BLACHE P. (2001). *Les grammaires de propriétés : des contraintes pour le traitement automatique des langues naturelles*. Collection Technologies et cultures. Hermès Science publications.

CHALLANT C. & FILHOL M. (2022). A First Corpus of AZee Discourse Expressions. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, p. 1560-1565, Marseille, France.

CHOMSKY N. (1965). *Aspects of the Theory of Syntax*. Cambridge : MIT Press.

FILHOL M., HADJADJ M. & CHOISIER A. (2014). Non-Manual Features : The Right to Indifference. In *International Conference on Language Resources and Evaluation(LREC)*, p. 49–54, Reykjavik, Iceland.

FILHOL M. & TANNIER X. (2014). Construction of a French–LSF Corpus. In *Building and Using Comparable Corpora, Language Resource and Evaluation Conference (LREC)*, p. 2–5, Reykjavik, Iceland.

FILLMORE C. J. (1988). The Mechanisms of "Construction Grammar". *Annual Meeting of the Berkeley Linguistics Society*, **14**(00), 35–55. DOI : 10.3765/bls.v14i0.1794.

HERRERA S., KAHANE S. & GUILLAUME B. (2022). Extraction de règles de grammaire à partir de treebanks : développement d'un outil et premiers résultats. *Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain (LIFT) et Traitement Automatique des Langues (TAL)*, (93–98).

KIMMELMAN V. & PFAU R. (2021). Information Structure – Theoretical Perspectives. *The Routledge handbook of theoretical and experimental sign language research.*, p. 591–613.

LANGACKER R. W. (1987). *Foundations of Cognitive Grammar : Volume I : Theoretical Prerequisites*. Stanford University Press.

MARTÍNEZ R., SIYAVOSHI S. & WILCOX S. (2020). Advances in the Study of Signed Languages within a Cognitive Perspective. *Hesperia : Anuario de Filología Hispánica*, **23**, 29–56. DOI : 10.35869/hafh.v23i0.1654.

NAPOLI D. J. & SUTTON-SPENCE R. (2014). Order of the Major Constituents in Sign Languages : Implications for All Language. *Frontiers in Psychology*, **5**, 376. DOI : 10.3389/fpsyg.2014.00376.

VAN TRIJP R. (2015). Towards Bidirectional Processing Models of Sign Language : A Constructional Approach in Fluid Construction Grammar. In *Proceedings of the EuroAsianPacific joint conference on cognitive science*, p. 668–673, Turin : Univeristy of Torino.