Lidil

Revue de linguistique et de didactique des langues

60 | 2019

Langues des signes et genres discursifs

Étude exploratoire de la fréquence des catégories linguistiques dans quatre genres discursifs en LSF

An Exploratory Study of the Frequency of Linguistic Categories in Four Discourse Genres in LSF

Marie-Anne Sallandre, Antonio Balvet, Geoffrey Besnard et Brigitte Garcia

Édition électronique

URL : <http://journals.openedition.org/lidil/7136>

DOI : 10.4000/lidil.7136

ISSN : 1960-6052

Éditeur

UGA Éditions/Université Grenoble Alpes

Édition imprimée

ISBN : 978-2-37747-164-5

ISSN : 1146-6480

Référence électronique

Marie-Anne Sallandre, Antonio Balvet, Geoffrey Besnard et Brigitte Garcia, « Étude exploratoire de la fréquence des catégories linguistiques dans quatre genres discursifs en LSF », *Lidil* [En ligne], 60 | 2019, mis en ligne le 01 novembre 2019, consulté le 14 avril 2020. URL : <http://journals.openedition.org/lidil/7136> ; DOI : <https://doi.org/10.4000/lidil.7136>

Ce document a été généré automatiquement le 14 avril 2020.

© Lidil

Étude exploratoire de la fréquence des catégories linguistiques dans quatre genres discursifs en LSF

An Exploratory Study of the Frequency of Linguistic Categories in Four Discourse Genres in LSF

Marie-Anne Sallandre, Antonio Balvet, Geoffrey Besnard et Brigitte Garcia

1. Introduction

- 1 Cet article est la première étude portant sur la quasi-intégralité du corpus LS-COLIN (Cuxac et coll., 2002). Ce corpus avait été élaboré initialement avec le double objectif de créer un corpus de référence de la langue des signes française (LSF) et de comparer plusieurs genres discursifs chez des locuteurs sourds venant de régions différentes. Le premier objectif a été atteint, puisque de nombreuses personnes, chercheurs, étudiants, enseignants de LSF, utilisent ce corpus pour des applications diverses, mais, à notre connaissance, il ne semble pas que ce corpus ait été utilisé pour réaliser une étude exhaustive sur les genres discursifs¹. Cet article répond donc au souhait de départ des concepteurs du corpus. Par ailleurs, le corpus LS-COLIN étant de type monologique, il nous a semblé important de compléter ces données par des données de type dialogique ; des échantillons de dialogues entre adultes sourds provenant du corpus Creagest (Garcia, L'Huillier & Vincent, 2015) ont ainsi été sélectionnés pour le présent article.
- 2 Nous évoquerons d'abord les concepts-clés et les enjeux d'une étude linguistique portant sur les genres discursifs en langue des signes (LS). Puis la méthodologie du recueil et de l'annotation des données est présentée. Ensuite, les résultats sont abordés d'un point de vue quantitatif, par la synthèse des effectifs des catégories linguistiques par genre discursif. Enfin, une analyse statistique exploratoire est proposée.

2. Genres discursifs et linguistique de corpus

2.1. Hétérogénéité et genres discursifs

- 3 Rastier et Pincemin (1999) indiquent que le problème de la typologie des textes est central pour toute linguistique de corpus. C'est dans cette optique que la question des genres discursifs prend tout son sens. Ils précisent que « dans l'hypothèse que les genres sont les normes déterminantes de la sémiotique textuelle, la production et l'interprétation des textes en dépend étroitement » (résumé en ligne, p. 1). De plus, selon ces auteurs :

Un genre se définit en effet par (i) la cohésion d'un faisceau de critères, tant au plan du signifié qu'à celui du signifiant, et par (ii) son incidence sur la textualité, sur ces deux plans également [...]. Aussi, il détermine ce mode de corrélation entre plan du signifiant et plan du signifié que l'on peut nommer *sémiotique textuelle*. (p. 96)

- 4 Par ailleurs, Adam (2001) distingue cinq grands genres discursifs (récit, description, argumentation, explication et dialogue). L'auteur indique d'emblée qu'au sein de chaque texte, néanmoins, une hétérogénéité des séquences est présente : par exemple, au sein d'une narration, il est fréquent d'avoir une séquence purement descriptive.
- 5 Notre corpus, qui est constitué de quatre genres en LSF — narratif, argumentatif, explicatif, dialogal — n'échappe pas à l'hétérogénéité relevée par Adam (2001, p. 34). Par exemple, plusieurs de nos locuteurs, qui se remémorent le contexte dans lequel ils ont été informés des événements tragiques du 11 septembre 2001, insèrent des séquences narratives dans leur discours, comme le suggère Risler (2016, p. 47), alors que celui-ci est globalement argumentatif. Ainsi, il faut considérer que chaque discours de notre corpus s'inscrit dans un genre principal, mais que l'hétérogénéité est la norme, comme cela a été relevé dans la littérature.

2.2. Options théoriques et catégories linguistiques considérées

- 6 L'approche théorique adoptée dans cet article, appelée « approche sémiotique », a été développée pour la LSF et pour d'autres langues des signes (Cuxac, 2000 ; Garcia & Sallandre, 2014). Élaborée dans une perspective énonciative et fonctionnelle, cette approche défend l'idée que, à la différence des langues vocales, les langues des signes, structurées par l'iconicité, disposent de deux manières de *dire* selon deux visées sémiotiques, illustrative et non illustrative. Cuxac (2000) a montré que ces constructions iconiques à visée illustrative² relèvent de trois grands types de structures, dites « structures de transfert » : le transfert de taille et de forme (TTF), qui décrit dans l'espace la forme et/ou la taille d'une entité³ ; le transfert situationnel (TS), qui montre le déplacement d'un actant (main dominante) par rapport à un locatif (main dominée), et le transfert personnel (TP), par lequel le locuteur incarne l'entité dont il parle et dont il donne à voir les actions effectuées ou subies. À ces trois grands types s'ajoutent des sous-catégories de TP, dont deux sont prises en compte dans nos analyses statistiques : le double transfert (DT), qui est l'association simultanée d'un TP et d'un TS, et le transfert personnel en discours rapporté (TP dr), qui est utilisé quand un locuteur en prise de rôle fait dialoguer les personnages de son récit. Ces structures de transfert, qui mobilisent l'ensemble des composants manuels et non manuels, sont caractérisées par une rupture du regard du locuteur avec celui de l'interlocuteur. Elles génèrent, en nombre illimité, des unités dites « unités de transfert » (UT). Les UT

s'entrelacent avec l'autre grand type d'unités, relevant, elles, de la visée non illustrative, que sont les unités lexicales (UL), ainsi qu'avec les pointages (Point) et la dactylogogie (Dactylo). Les UL sont des unités stabilisées d'une LS nationale, et elles produisent leur sens comme le font les mots des langues vocales, par convention. Si une UL est employée dans le discours dans sa forme conventionnelle, par exemple pour introduire un nouveau référent, alors elle sera annotée dans la catégorie « UL ».

- 7 Enfin, grâce à leurs composants manuels (configuration, orientation, emplacement et mouvement) et non manuels (regard, mimique faciale, posture corporelle et mouvement labial) qui sont les mêmes quelle que soit l'unité — UT, UL et pointages — sont compositionnelles. Par exemple, il arrive qu'une UL apparaisse dans le cadre d'une structure de transfert. C'est le cas notamment des semi-TP, dans lesquels l'action effectuée par l'entité incarnée est signée au moyen d'une UL⁴ ; dans ce cas, qui représente en moyenne 5 % du total des unités, les unités ont été annotées par l'item « semi-TP », puis regroupées dans la super-catégorie « TP » dans cet article. C'est le cas également avec les TP dr, qui utilisent des UL, dans l'immense majorité des cas, pour faire dialoguer les personnages d'un récit. Là, en revanche, une catégorie autonome a été dégagée pour nos analyses, celle de « TP dr » afin de mettre en valeur l'importance quantitative — jusqu'à 28 % du total des unités d'un discours — de ces types de structures.

2.3. Genres discursifs en langues des signes

- 8 Les recherches portant spécifiquement sur les genres discursifs en langues des signes sont peu nombreuses. Nous en avons sélectionné deux pour leur relative proximité avec notre propre étude. La première, celle de Russo (2004), compare trois types de textes en LS italienne : des poésies, des narrations et des conférences, qui met en relation avec des types et des degrés d'iconicité variables. Russo considère en effet que l'iconicité affecte potentiellement tous les niveaux de la langue, du niveau sublexical au niveau discursif. À tous les niveaux, il est possible selon lui de distinguer deux types d'iconicité, l'iconicité *productive* (ou *dynamique*) et l'iconicité *figée* (ou *dormante*). L'iconicité dynamique est ce que Cuxac (2000) appelle « structures de grande iconicité » et qui relève des UT, tandis que l'iconicité dormante fait référence à ce que Cuxac (2004) appelle « iconicité dégénérée » et qui affecte les UL, pour un ou plusieurs de leurs composants sublexicaux. Sur les 823 signes composant le corpus de Russo (2004), la distribution est la suivante : d'une part, l'iconicité dynamique est particulièrement massive dans le genre poétique (53,4 %) et le genre narratif (43 %), comme on pouvait s'y attendre, mais beaucoup moins fréquente (13,5 %) dans les conférences, qui relèvent du genre explicatif. D'autre part, l'iconicité dormante est particulièrement massive, également, dans la plupart des signes qui compose son corpus, 77 % des signes présentent en effet une forme d'iconicité figée pour au moins l'un de leurs composants sublexicaux dans les poèmes, 64 % dans le genre narratif et 47 % dans le genre explicatif. Par ailleurs, au sein même des deux grands types d'iconicité, l'auteur distingue des sous-types que nous ne développons pas ici mais qui présentent un intérêt certain pour la description fine des genres et des types de textes en LIS.
- 9 La deuxième étude (Meurant & Sinte, 2016) porte sur la LS de Belgique francophone et analyse les phénomènes de reformulation dans trois genres discursifs, narration,

explication et conversation. Cette étude présente divers points de proximité avec la nôtre, ce qui rend les comparaisons plus aisées : d'une part, la terminologie utilisée est en partie issue de l'approche sémiologique (par exemple, les TP) et, d'autre part, les annotations ont été faites avec le logiciel ELAN à partir d'une grille assez proche de la nôtre. Contrairement à ce que les auteurs avaient anticipé, les formes lexicales (LE), équivalentes de nos UL, sont largement plus présentes dans leur corpus que toutes les formes partiellement lexicalisées (PL) cumulées, qui recoupent en grande partie nos UT. La prédominance du lexique est la plus nette dans les conversations (LE : 83 %, PL : 14 %), elle est également importante dans les explications (LE : 61 %, PL : 31 %), mais elle est moins marquée dans les narrations (LE : 58 %, PL : 42 %). Selon les auteurs, cette particularité de la narration doit être mise en relation avec le constat de Sallandre (2003) que les narrations contiennent une plus forte proportion de structures partiellement lexicalisées que les autres genres.

3. Méthodologie

3.1. Les corpus utilisés

- ¹⁰ Dans cet article, les données de trois genres discursifs provenant du corpus LS-COLIN (Cuxac et coll., 2002) sont les suivantes : deux narrations à partir de stimuli en images (*Cheval*, *Oiseaux*), une explication-prescription (une recette de cuisine, abrégée *Cuisine*) et une argumentation (*Onze-Septembre*) ; ce dernier consistait, pour chaque locuteur, à expliquer ce qu'il pensait des retombées géopolitiques mondiales après les attentats du 11 septembre 2001 à New York⁵. À ces données de type monologique, s'ajoutent des données de type dialogique issues de six extraits de dialogues entre adultes sourds (corpus Creagest : Garcia, L'Huillier & Vincent, 2015, abrégé *Dialogue*). Dans ces extraits, chaque locuteur explique à un enquêteur sourd son métier, ses loisirs, etc. La durée totale des discours étudiés est ainsi d'une heure et vingt-cinq minutes. Le tableau 1 présente une synthèse des métadonnées.

Tableau 1. – Récapitulatif des métadonnées⁶.

Genre discursif	Type	Titre	Nombre de locuteurs	Durée totale (h:min:sec)
narratif	monologique	<i>Cheval</i>	13	00:19:27
narratif	monologique	<i>Oiseaux</i>	13	00:15:01
explicatif/ prescriptif	monologique	<i>Cuisine</i>	13	00:31:05
argumentatif/mixte	monologique	<i>Onze-Septembre</i>	6	00:11:57
mixte	dialogique	<i>Dialogue</i>	6	00:07:29
Total			(51 discours)	01:24:59

3.2. Profil des locuteurs

- 11 Bien que ne constituant pas des variables pour notre étude, les informations concernant l'âge, le sexe, la profession, etc., des dix-neuf locuteurs de notre corpus sont importantes pour comprendre le contexte de ces données. Elles sont donc fournies à titre indicatif.
- 12 Les treize locuteurs du corpus LS-COLIN sont sourds de naissance et utilisent la LSF comme langue principale dans leur vie quotidienne. Une majorité d'entre eux est issue de parents entendants (11 personnes sur 13) mais a bénéficié d'un bain linguistique en LSF relativement précoce, soit parce qu'au moins un membre de la fratrie est sourd (5 personnes sur 11), soit par les écoles dans lesquelles ils ont été scolarisés⁷. Les huit hommes et cinq femmes qui ont été enregistrés avaient entre 23 et 60 ans au moment du tournage. Celui-ci a eu lieu à Paris mais les locuteurs résidaient dans diverses villes de France. La majorité d'entre eux exerce le métier d'enseignant de LSF au sein d'associations ou d'établissements scolaires⁸.
- 13 Les extraits du corpus Creagest ont été sélectionnés parmi une vaste base de données comprenant les discours de 55 adultes sourds dans quatre régions françaises. La sélection a été faite sur la base de la variété thématique des contenus signés d'une part et du type de métadonnées d'autre part⁹. Ainsi, sur les six locuteurs présents dans les extraits, il y a trois femmes et trois hommes, âgés de 31 à 48 ans. Il y a deux locuteurs par région (Angers, Lyon et Paris). Ils ont des professions variées : informaticien, enseignante, responsable d'association, etc. Trois locuteurs sur six sont issus de familles sourdes, mais tous ont appris la LSF de manière relativement précoce, comme les locuteurs du corpus LS-COLIN. La durée moyenne des extraits est d'une minute et quinze secondes (voir le tableau 1), ce qui représente un très court échantillon des dialogues qui ont une durée moyenne d'une heure trente chacun¹⁰. Cependant, la variété des thèmes abordés et le type dialogique de ces données nous ont semblé intéressants à prendre en compte.

3.3. Annotations

- 14 D'après Adam (2001), le texte est le plus souvent hétérogène par nature, alors que c'est au niveau de la séquence que le chercheur peut identifier des segments homogènes. Bien qu'étant d'accord avec le niveau de la séquence, nous n'avons pas annoté notre corpus en retenant ce niveau textuel intermédiaire. Nous avons en effet annoté, d'une part, à un niveau macro, par type de texte, c'est-à-dire qu'un discours produit par un locuteur représente un fichier d'annotation. Ainsi nous avons 51 fichiers d'annotation au total. Et, d'autre part, au niveau micro, nous avons annoté en « unité de sens », c'est-à-dire en unité minimale de réalisation (UL, UT, etc.).
- 15 Les annotations des corpus LS-COLIN et Creagest ont été réalisées avec le logiciel ELAN (Crasborn & Sloetjes, 2008) par deux annotateurs¹¹ qui sont aussi auteurs de l'article. Il s'agit donc d'annotations de première main qui ont été effectuées avec un objectif de recherche précis, celui d'identifier, pour chaque unité minimale de réalisation, la catégorie linguistique qui lui est assignée (UL, TP, etc.). La grille d'annotation, ou *template*, est constituée de plusieurs lignes, dont voici les principales : « stimulus », « unité de sens », « catégorie linguistique », puis une ligne pour chacun des composants

manuels et non manuels. Certaines lignes sont liées à un vocabulaire contrôlé, d'autres non¹². Pour les dialogues, les lignes de la grille d'annotation ont simplement été dédoublées, pour chacun des interlocuteurs¹³.

4. Analyse

4.1. Exploration en corpus

- 16 Nous présentons ici une exploration de différents genres discursifs en LSF. Bien que la LSF et les productions considérées présentent des spécificités (modalité visio-gestuelle, influence limitée du code écrit, oralité), nous nous inscrivons dans une démarche proche de celle exposée dans Biber (1991) : en posant que les productions étudiées se répartissent dans des genres discursifs distincts, nous cherchons à identifier les catégories d'unités linguistiques les plus saillantes pour chaque genre. Ce faisant, nous nous attendons à observer des effets de continuum entre fréquence des unités considérées et genre discursif, pour les quatre genres principaux analysés ici. En d'autres termes, nous nous attendons à observer des associations ou corrélations significatives entre paramètres (les unités considérées), ainsi que des correspondances entre ces paramètres — éventuellement corrélés entre eux — et les différents genres supposés. La présente étude consiste donc en l'application d'une méthode statistique exploratoire pour l'étude des genres en LSF. Nous examinons la répartition des différentes catégories par genre discursif. Puis nous exploitons différents outils statistiques reposant sur des mesures de variation (*i.e.* variance, corrélation, association entre paramètres), en préambule à une analyse en composantes principales (ACP), afin d'identifier d'éventuelles correspondances entre unités linguistiques et genres discursifs.
- 17 Dans cette partie, nous cherchons donc avant tout à caractériser les différents genres supposés en termes d'unités linguistiques de la LSF. Soulignons que, contrairement à Biber et d'autres auteurs, la présente étude ne repose pas sur l'analyse de la distribution de formes linguistiques (unités lexicales isolées, patrons récurrents), ni même de catégories syntaxiques, mais plutôt de catégories linguistiques définies dans le cadre de l'approche sémiologique (UL, UT, pointages et dactylologie).

4.2. Statistiques descriptives : répartitions des catégories linguistiques par genre

- 18 Le tableau 2 synthétise la répartition des différentes unités, en effectifs, issues des catégories principales, et par type de productions, tandis que le tableau 3 en donne la répartition en pourcentages.

Tableau 2. – Répartition des différentes catégories linguistiques par production (effectifs).

	<i>Dialogue</i>	<i>Onze-Septembre</i>	<i>Cheval</i>	<i>Oiseaux</i>	<i>Cuisine</i>	Total
UL	592	747	288	330	1389	3346
Point	90	108	41	46	109	394

Dactylo	9	2	3	0	23	37
TP	13	126	447	310	392	1288
TP dr	37	29	385	139	99	689
DT	0	24	58	51	163	296
TS	10	72	62	111	52	307
TTF	14	63	100	97	234	508
Total	765	1171	1384	1084	2461	6865

- 19 Les deux catégories linguistiques présentant des effectifs très élevés sont, avant tout UL, puis TP. Notons que, à l'inverse, certaines catégories présentent des effectifs très bas, voire nuls : en particulier, la catégorie Dactylogologie est globalement peu représentée, voire totalement absente dans *Oiseaux*, et les Doubles Transferts (DT), bien que plus fréquents, sont malgré tout absents de *Dialogue*. De leur côté, les pointages sont bien représentés dans les trois genres principaux, bien qu'avec des distributions assez contrastées en apparence.

Tableau 3. – Répartition des catégories pour chaque production (pourcentages).

	<i>Dialogue</i>	<i>Onze-Septembre</i>	<i>Cheval</i>	<i>Oiseaux</i>	<i>Cuisine</i>
UL	77,39 %	63,79 %	20,81 %	30,44 %	56,44 %
Point	11,76 %	9,22 %	2,96 %	4,24 %	4,43 %
Dactylo	1,18 %	0,17 %	0,22 %	0,00 %	0,93 %
TP	1,70 %	10,76 %	32,30 %	28,60 %	15,93 %
TP dr	4,84 %	2,48 %	27,82 %	12,82 %	4,02 %
DT	0,00 %	2,05 %	4,19 %	4,70 %	6,62 %
TS	1,31 %	6,15 %	4,48 %	10,24 %	2,11 %
TTF	1,83 %	5,38 %	7,23 %	8,95 %	9,51 %
Total	100 %	100 %	100 %	100 %	100 %

- 20 Les pourcentages permettent de mieux apprécier l'importance relative de chaque catégorie pour les différents genres. Les pourcentages empilés sont repris dans la figure 1.

Figure 1. – Catégories linguistiques par production (pourcentages empilés).

- 21 Bien que les répartitions soient très contrastées, il apparaît que les principales catégories sont représentées, quel que soit le genre. La distinction entre genre narratif (*Cheval*, *Oiseaux*) et non narratif (*Dialogue*, *Onze-Septembre* et *Cuisine*) apparaît clairement liée à la prépondérance, d'un côté des unités avec visée illustrative (UT), de l'autre des unités sans visée illustrative (UL, Point, Dactylo). Par ailleurs, pour un même genre, on peut constater une certaine variabilité : les profils de distribution des différentes catégories, bien que similaires entre *Cheval* et *Oiseaux*, ne sont pour autant pas identiques. *Oiseaux* présente une proportion plus importante d'UL et de TS que *Cheval*, notamment. Ces différences s'expliquent certainement pour deux raisons, liées au stimulus : la première est qu'il y a plus de personnages dans *Oiseaux* que dans *Cheval*, et les locuteurs doivent nommer ces personnages, en général par des UL ; la deuxième tient à la présence de l'entité « arbre » dans *Oiseaux* qui est une contrainte spatiale importante, obligeant la plupart des locuteurs à garder, par le bras de leur main dominante, ce référent spatial, tandis qu'ils signent le déplacement des personnages (oiseaux, chat, chien) par rapport à celui-ci ; le résultat est la présence accrue de TS dans cette histoire.
- 22 Enfin, comme on peut le voir, les proportions respectives des UL par rapport aux différents transferts s'inversent en fonction des genres : la proportion des UL suit la progression suivante *Dialogue* > *Onze-Septembre* > *Cuisine* > *Oiseaux* > *Cheval*. La figure 1 montre clairement que les types de TP (TP, TP dr et DT) progressent lorsque la proportion des UL régresse¹⁴. On aurait donc, du point de vue du genre discursif la progression suivante : dialogique > argumentatif > explicatif > narratif, une chaîne qu'il nous faut explorer statistiquement.

4.3. Exploration statistique par catégories linguistiques et genres discursifs

- 23 Après avoir présenté les effectifs bruts et les proportions de chaque catégorie dans les différentes productions étudiées, nous présentons quelques outils statistiques afin d'explorer plus précisément l'association (ou dépendance) entre catégories linguistiques et genres discursifs. Dans l'ensemble des manipulations présentées ici, nous nous appuyons sur les principes d'analyse des données linguistiques aidées du logiciel R détaillés dans Levshina (2015).

4.3.1. Tests d'indépendance : χ^2

- 24 Le tableau 2 ci-dessus regroupe les différents facteurs Catégories x Genres. Il constitue un tableau de contingence, sur lequel il est possible d'appliquer un test d'indépendance, le test du χ^2 . Ce test repose essentiellement sur la différence de répartition entre effectifs observés (noté o), dans un tableau de contingence (ici de 5 lignes par 8 colonnes), et effectifs attendus (*expected*, noté e), par le simple jeu du hasard. Le χ^2 de Pearson s'obtient par la formule générale suivante :

Formule du test de χ^2

$$\sum \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

- 25 Dans l'application du test du χ^2 , nous cherchons à déterminer dans quelle mesure les variables Catégories linguistiques et Genres discursifs sont dépendantes l'une de l'autre. Pour ce faire, nous estimons, à l'aide du test d'indépendance du χ^2 la certitude de l'hypothèse inverse, à savoir que les variables Catégories et Genres sont indépendantes, c'est-à-dire que les distributions observées des catégories dans les différents genres s'expliquent par le simple fait du hasard.
- 26 Nous calculons le score de χ^2 grâce à la fonction R 'chisq', qui livre le résultat du test de Pearson : X-squared = 1887.9, df = 28, p-value < 2.2e-16. Le résultat du test semble indiquer une association significative entre les variables considérées. En effet, le score enregistré, rapporté au nombre de degrés de liberté (ici 28), indique une p-value de $2,2 \times 10^{-16}$, largement inférieure au seuil de rejet de 0,05 (ainsi que celui plus strict de 0,01). Autrement dit, la probabilité de se tromper en affirmant que les paramètres Catégories et Genres sont liés est très faible.
- 27 Signalons toutefois que, en raison de la distribution des annotations dans le corpus, le résultat de ce test est à prendre avec précaution. En effet, comme indiqué plus haut, le test repose sur l'estimation d'une distribution théorique pour chaque catégorie linguistique, basée sur les effectifs observés. Or, l'effectif théorique obtenu pour la catégorie Dactylo pour *Dialogue*, est inférieur à 5. En principe, toute valeur inférieure à 5 doit faire préférer l'adoption de tests alternatifs au χ^2 , tel que le test exact de Fisher. Nous gardons toutefois la valeur incriminée, au motif qu'il s'agit avant tout ici d'une démarche exploratoire. Toutefois, en toute rigueur, la catégorie Dactylo devrait soit être éliminée de l'étude, soit être fusionnée avec une autre catégorie. En effet, outre la fréquence théorique très basse soulignée, c'est la classe Dactylo tout entière qui est très peu fréquente, quelle que soit la production considérée.
- 28 Une fois le test du χ^2 réalisé, nous pouvons déterminer les résidus de Pearson, notés r , par la fonction suivante :

Résidus de Pearson

$$r = \frac{o - e}{\sqrt{e}}$$

- 29 Ici encore, o correspond aux effectifs observés, et e aux effectifs attendus (*expected*). Les résidus sont calculés pour chaque cellule¹⁵ dans le tableau 4. Ces scores permettent d'identifier quelles cellules du tableau de contingence participent le plus au score global du χ^2 . En ce sens, les résidus de Pearson permettent d'identifier quelles catégories sont les plus liées à un genre donné.

Tableau 4. – Résidus de Pearson (arrondis à la deuxième décimale).

	UL	Point	Dactylo	TP	TP dr	DT	TS	TTF
<i>Dialogue</i>	11,35	6,96	2,4	-10,9	-4,54	-5,74	-4,14	-5,66
<i>Onze-Septembre</i>	7,38	4,98	-1,72	-6,32	-8,17	-3,73	2,71	-2,54
<i>Cheval</i>	-14,88	-4,31	-1,63	11,63	20,88	-0,22	0,01	-0,24
<i>Oiseaux</i>	-8,63	-2,06	-2,42	7,48	2,9	0,62	8,98	1,87
<i>Cuisine</i>	5,47	-2,71	2,67	-3,25	-9,42	5,52	-5,53	3,85

- 30 L'identification des cellules aux résidus les plus importants est rendue plus facile grâce au diagramme en bulles de la figure 2.

Figure 2. – Diagramme en bulles des résidus de Pearson.

- 31 Dans ce diagramme, la taille des bulles est proportionnelle au résidu de Pearson pour la cellule considérée. La couleur et l'intensité de la couleur de la bulle indiquent respectivement l'attraction entre Genres discursifs et Catégories linguistiques (en bleu : associations positives), et la répulsion entre ces deux paramètres (en rouge : associations négatives). Ce diagramme indique une forte association positive entre le genre narratif, et en particulier l'histoire du *Cheval*, et la catégorie TP dr, ainsi que TP. Ce résultat va dans le sens d'un constat général, déjà formulé par Bouvet (1996) et conforté dans Sallandre (2003), selon lequel les TP en discours rapporté seraient plus spécifiques au genre narratif. Cela confirme également une forte structuration linguistique entre dialogue (ou « discours ») et récit (Benveniste, 1966), décrite par Millet (2019, p. 143-146) pour la LSF.
- 32 À l'inverse, *Cheval* est très négativement associé à la catégorie UL. *Dialogue*, quant à lui, est positivement associé aux UL, ainsi qu'aux pointages dans une moindre mesure. Ils sont par ailleurs négativement associés aux unités de transferts (TP, TP dr, DT, TS, TTF).

4.3.2. Analyse en composantes principales (ACP)

- 33 Afin de mieux saisir les dépendances entre les variables considérées, il est utile d'exploiter l'analyse en composantes principales (ACP), car elle permet de projeter les différentes dimensions de variations présentes dans le jeu de données vers un nombre réduit de dimensions (généralement deux ou trois), plus facilement interprétables. Avant toute chose, il convient de déterminer dans quelle mesure les données sont corrélées ou non. Le tableau 5 fournit les scores de corrélation entre les différents facteurs¹⁶. Dans ce tableau, un score égal à 0 signale une absence de corrélation, alors qu'un score proche de 1 (ou -1) signale une corrélation forte.

Tableau 5. – Scores de corrélations entre catégories linguistiques.

	UL	Point	Dactylo	TP	TP dr	DT	TS	TTF
UL	1	0,89	0,71	-1	-0,87	-0,58	-0,57	-0,66
Point	0,89	1	0,5	-0,92	-0,66	-0,88	-0,38	-0,89
Dactylo	0,71	0,5	1	-0,72	-0,46	-0,25	-0,89	-0,45
TP	-1	-0,92	-0,72	1	0,82	0,65	0,59	0,73
TP dr	-0,87	-0,66	-0,46	0,82	1	0,24	0,22	0,26
DT	-0,58	-0,88	-0,25	0,65	0,24	1	0,23	0,97
TS	-0,57	-0,38	-0,89	0,59	0,22	0,23	1	0,45
TTF	-0,66	-0,89	-0,45	0,73	0,26	0,97	0,45	1

- 34 Les scores de corrélations relevés nous indiquent que les données ne sont ni trop fortement ni trop faiblement corrélées, bien que des corrélations faibles (inférieures

à 0,3 en valeur absolue) soient présentes¹⁷. L'examen des valeurs propres (*eigenvalues*) nous indique que trois dimensions suffisent à expliquer 98,96 % de la variation observée dans les données.

- 35 À partir de ces résultats, il est possible d'étudier les deux premières dimensions identifiées par l'ACP, grâce au diagramme de la figure 3.

Figure 3. – Corrélations entre facteurs (Catégories linguistiques) sur les dimensions 1 et 2.

- 36 L'ACP livre des résultats qui peuvent être résumés comme suit : premièrement, la variation¹⁸ observée dans la distribution des catégories par production peut être ramenée à seulement deux dimensions. Deuxièmement, en prenant appui sur ces deux dimensions, il est possible de visualiser dans quelle mesure différents facteurs sont liés entre eux. Ainsi, sur l'axe horizontal (dimension 1), on observe une répartition qui peut être interprétée comme l'existence d'un continuum entre visée illustrative (à droite), et non illustrative (à gauche). Il faut souligner que cette distinction n'est pas explicitement codée dans les annotations ; elle est détectée au travers de la distribution de chaque catégorie dans les différentes productions. Les sous-catégories de transferts se trouvent ainsi regroupées, du simple fait de leur distribution au sein du corpus, dans la partie droite du diagramme. Il en va de même pour les catégories Dactylo, UL et Point, qui sont regroupées à gauche. Sur l'axe vertical (dimension 2), on peut observer une distinction, opérée là encore sur la simple observation des distributions des différentes catégories, entre TP, TP dr et TS d'une part (quadrant inférieur droit) et TTF + DT d'autre part (quadrant supérieur droit). De même, pour la visée non illustrative, Point et Dactylo sont placés respectivement dans les quadrants inférieur gauche et supérieur gauche, de part et d'autre de UL. Enfin, UL est positionnée à l'extrême opposé des TP, conformément aux attentes des premières observations des distributions mutuellement exclusives de ces deux catégories (tableau 2).
- 37 Grâce aux dimensions identifiées par l'ACP, il est possible de projeter le jeu de données sur les deux premières dimensions (celles qui capturent le plus de variations), dans un plan, comme le montre la figure 4.

Figure 4. – Répartition conjointe unités/genres dans le plan.

- 38 Rappelons que la répartition des unités linguistiques est réalisée de façon purement automatique par la bibliothèque FactoMiner du logiciel R. Elle s'appuie uniquement sur les distributions relevées en corpus. Or, la répartition proposée de façon automatique par l'algorithme ACP peut être interprétée comme une partition entre catégories à visée non illustrative, en l'occurrence ici dans le quadrant inférieur gauche du plan, par opposition aux catégories de la visée illustrative (quadrant supérieur droit). Les UT apparaissent disjointes des autres catégories, et parmi celles-ci, les TP dr semblent à part.
- 39 La projection dans le plan peut également s'interpréter sous l'angle des frontières entre genres discursifs. Comme on peut le constater, l'ACP semble sensible à la partition principale entre genre narratif et non narratif : *Oiseaux* et *Cheval* sont représentés en partie droite du plan, alors que les autres productions sont regroupées en partie gauche. Toutefois, du côté non narratif, on peut constater que *Cuisine* est positionné dans le quadrant supérieur gauche, alors que *Dialogue* et *Onze-Septembre* sont placés dans le quadrant inférieur gauche. En d'autres termes, il semble que les frontières entre genres explicatif (*Cuisine*) d'un côté, et argumentatif/mixte (*Onze-Septembre*) de l'autre soient objectivables à partir de la distribution observée des différentes catégories dans le corpus. *Dialogue*, bien que placé du côté des productions non narratives, semble cependant distinct de *Cuisine*, mais également de *Onze-Septembre*. Cette répartition rejoint la caractérisation des *Dialogue* comme genre mixte.

5. Discussion et perspectives

- 40 Cette étude portant sur les discours de dix-neuf locuteurs sourds a révélé des tendances dans les types de catégories utilisés au regard du genre discursif. Globalement, la catégorie UL est la catégorie la plus fréquente, quel que soit le genre discursif. La deuxième catégorie la plus fréquente est le TP, non seulement dans les narrations, où elle est très présente, mais aussi dans les autres genres, où sa fréquence n'est pas anecdotique. Enfin, les TP en discours rapporté sont également fréquents, et sont préférentiellement utilisés dans le genre narratif. Deux des trois catégories les plus fréquemment utilisées relèvent donc de la prise de rôle, c'est-à-dire de situations dans lesquelles le locuteur-énonciateur s'efface pour incarner l'entité qu'il représente.
- 41 Les statistiques exploratoires, présentées ici pour la première fois pour ce corpus de LSF, apportent une vision nouvelle des données. Par exemple, l'ACP montre que les paramètres de variation (catégories) et les différentes productions étant projetés ensemble sur le même plan, il est possible d'identifier quelles catégories linguistiques sont spécifiques d'un genre discursif donné. En l'occurrence, la corrélation entre la catégorie Pointage et le genre *Dialogue* semble la plus claire. Ensuite, la catégorie UL semble fortement associée à *Onze-Septembre*, genre plutôt argumentatif. Enfin, les catégories TS et TP semblent toutes deux fortement corrélées avec *Oiseaux*. En ce qui concerne les deux autres genres, *Cuisine* ne semble pas spécifiquement corrélé à l'une ou l'autre des catégories, mais semble plutôt exploiter l'ensemble des unités (Dactylo, UL, TS, TP, TTF) à l'exception des Pointages et des TP dr. De son côté, *Cheval* semble exploiter l'ensemble des transferts (TS, TP, DT et TP dr). Au sein du genre narratif, la différence principale entre *Oiseaux* et *Cheval* semble tenir à la force de la corrélation entre TS et TP : forte dans le cas de *Oiseaux*, répartie sur l'ensemble des catégories de transferts pour *Cheval*.
- 42 Par ailleurs, même s'il est difficile de faire des comparaisons strictes entre nos résultats et ceux de Russo (2004), nos critères de segmentation et nos catégories linguistiques étant différents, des similitudes apparaissent toutefois : l'iconicité dynamique de Russo est à mettre directement en relation avec l'iconicité d'image propre aux UT, or c'est la catégorie dominante selon Russo pour les genres poétique et narratif. Les résultats de Meurant et Sinte (2016), quant à eux, sont plus comparables avec les nôtres, car nos catégories linguistiques se recoupent partiellement. Le phénomène marquant est la forte similitude dans les proportions trouvées en LSF et en LSFB pour les catégories du lexique et des transferts, dans les différents genres étudiés. Quoi qu'il en soit, ces deux études, portant sur deux langues des signes autres que la LSF apportent des informations précieuses sur la répartition des types d'unités linguistiques selon les genres discursifs.
- 43 Les résultats présentés dans cette étude sont éclairants à plusieurs titres, bien que les outils statistiques employés aient essentiellement comme finalité de faire émerger des corrélations difficiles à repérer à l'œil nu au sein des dimensions de variation du corpus. Tout d'abord, les catégories principales, rattachées aux deux types de visée proposées dans l'approche sémiologique, apparaissent clairement distinguées. L'ACP identifie d'un côté les unités de la visée non illustrative : UL, Point, Dactylo ; de l'autre, les unités de la visée illustrative : TP, TTF, TS, DT, etc. Ce premier constat donne une indication sur la stabilité des critères d'annotation de ces catégories principales, et donc de la pertinence de la notion de visée pour des annotateurs humains. Ces premiers

résultats semblent donc pertinents pour une exploitation didactique (enseignement de la LSF) : la notion de visée (illustrative, non illustrative) semble à même d'offrir aux apprenants un critère organisateur pouvant faciliter la compréhension des structures de la langue.

- 44 Par ailleurs, du point de vue du traitement automatique des LS (Hadjadj, Filhol & Braffort, 2018), les résultats présentés militent là encore pour une prise en compte des deux dimensions principales d'organisation du discours en LSF : visée et genre. Toutefois, soulignons ici que la notion de visée, en particulier, est accessible à des signeurs humains à même d'élaborer une représentation des buts communicatifs de l'allocutaire. Ce type de représentation fait appel à plusieurs fonctions cognitives de haut niveau (théorie de l'esprit, neurones miroirs, entre autres) pour l'heure inaccessibles à des machines. De notre point de vue, il est donc illusoire d'espérer voir émerger la dimension « visée », ou même « genre », au sein de dispositifs mécaniques, même nourris de données d'entraînement en grand volume. En effet, bien que des initiatives de collecte de signes isolés, de configurations manuelles ou de mimiques faciales faisant appel aux foules (crowdsourcing) aient permis de constituer des bases de données disponibles gratuitement, il n'en reste pas moins que ces initiatives ne prennent en général pas en compte le fonctionnement multilinéaire des LS. Dans le meilleur des cas, les bases de données ainsi constituées peuvent être réexploitées en tant que modules au sein d'un modèle plus large de traitement automatique des LS.

BIBLIOGRAPHIE

- ADAM, Jean-Michel. (2001). *Les textes : types et prototypes*. Paris : Nathan Université.
- BENVENISTE, Émile. (1966). *Problèmes de linguistique générale*. Paris : Gallimard.
- BIBER, Douglas. (1991). *Variation across Speech and Writing*. Cambridge : Cambridge University Press.
- BOUVET, Danielle. (1996). *Approche polyphonique d'un récit produit en langue des signes française*. Lyon : Presses universitaires de Lyon.
- CRASBORN, Onno & SLOETJES, Han. (2008). *Enhanced ELAN Functionality for Sign Language Corpora*. Présentation orale à la 6th International Conference on Language Resources and Evaluation (LREC 2008) / 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora.
- CUXAC, Christian. (2000). *La langue des signes française. Les voies de l'iconicité*. Paris : Ophrys.
- CUXAC, Christian et coll. (2002). *Corpus LS-COLIN sur plusieurs genres discursifs*. Programme Cognitique 2000, Langage et Cognition, Université Paris 8, IRIT, LIMSI-CNRS, Université Paris-Sorbonne. Disponible en ligne sur <<https://cocoon.huma-num.fr/exist/crdo/search2.xql>>.
- CUXAC, Christian. (2004). Phonétique de la LSF : une formalisation problématique. *Sillexicales*, 4, 93-113.

GARCIA, Brigitte, L'HUILLIER, Marie-Thérèse & SALLANDRE, Marie-Anne. (2013). CREAGEST : enjeux linguistiques, patrimoniaux et socio-éducatifs d'un grand corpus de langue des signes française. *La nouvelle revue de l'adaptation et de la scolarisation*, 64, 81-91.

GARCIA, Brigitte & SALLANDRE, Marie-Anne. (2014). Reference Resolution in French Sign Language (LSF). Dans P. Cabredo Hofherr & A. Zribi-Hertz (dir.), *Crosslinguistic Studies on Noun Phrase Structure and Reference* (p. 316-364). Leiden : Brill.

GARCIA, Brigitte, L'HUILLIER, Marie-Thérèse & VINCENT, Coralie. (2015). *Corpus Creagest, dialogues en LSF d'adultes sourds*. Projet ANR. <www.ortolang.fr/market/corpora/ortolang-000926>.

HADJADJ, Mohammed N., FILHOL, Michael & BRAFFORT, Annelies. (2018). Modeling French Sign Language: A Proposal for a Semantically Compositional System. Dans *Proceedings of the Language Resource and Evaluation Conference (LREC)*. Miyazaki, Japon. Disponible en ligne sur <www.aclweb.org/anthology/L18-1671.pdf>.

LEVSHINA, Natalia. (2015). *How to Do Linguistics with R: Data Exploration and Statistical Analysis*. Amsterdam : John Benjamins Publishing Company.

MEURANT, Laurence & SINTE, Aurélie. (2016). La reformulation en langue des signes de Belgique francophone (LSFB). Analyse dans un corpus de trois types de discours : narration, explication et conversation. *L'Information grammaticale*, 149, 32-44.

MILLET, Agnès. (2019). *Grammaire descriptive de la langue des signes française. Dynamiques iconiques et linguistique générale*. Grenoble : UGA Éditions.

RASTIER, François & PINCEMIN, Bénédicte. (1999). Des genres à l'intertexte. *Cahiers de praxématique*, 33, 83-111. Disponible en ligne sur <<http://praxematique.revues.org/1974>>.

RISLER, Annie. (2016). Parler de soi, parler des autres en langue des signes française. Autopointage et prise de rôle. *L'Information grammaticale*, 149, 45-54.

RUSSO, Tommaso. (2004). Iconicity and Productivity in Sign Language Discourse: An Analysis of Three LIS Discourse Registers. *Sign Language Studies*, 4, 164-197.

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). Disponible en ligne sur <<https://halshs.archives-ouvertes.fr/tel-00185376>>.

SALLANDRE, Marie-Anne. (2014). *Compositionnalité des unités sémantiques en langues des signes. Perspective typologique et développementale* (Thèse d'habilitation à diriger des recherches, Université Paris 8). Disponible en ligne sur <<https://hal.archives-ouvertes.fr/tel-01336182>>.

NOTES

1. Les études antérieures ont porté sur une partie du corpus LS-COLIN (Sallandre, 2003 et 2014, sur les genres narratif et explicatif; Risler, 2016, sur des données relevant du genre argumentatif/mixte).

2. Précisons que ce que l'approche sémiologique décrit de manière unifiée comme relevant de « structures de transfert » recouvre à la fois ce qui est couramment désigné ailleurs par le terme de « classifier constructions » et « role shift » ou « constructed actions ». Sur les correspondances terminologiques avec la littérature internationale, voir Garcia & Sallandre (2014, p. 332-333).

3. Cela renvoie en partie aux « spécificateurs de taille et de forme » chez Millet (2019).

4. Par exemple l'énoncé [TP^{cheval} APERCEVOIR VACHE] dans lequel l'UL APERCEVOIR est signée alors que le locuteur reste en TP du cheval. C'est donc bien le personnage qui fait l'action, pas le locuteur-énonciateur.
5. Le genre métalinguistique, également présent dans ce corpus, n'est pas étudié ici.
6. Les corpus classés par thème, *Onze-Septembre*, *Cheval*, *Oiseaux* et *Cuisine* ne sont pas toujours homogènes du point de vue du genre, comme évoqué précédemment, tandis que le genre « *Dialogue* » n'est, quant à lui, pas homogène du point de vue des thèmes abordés puisque chaque locuteur aborde un thème différent. Cependant, il semblait intéressant de comparer ces différentes productions de types monologique et dialogique.
7. Instituts spécialisés pour jeunes sourds ou classe bilingue français-LSF relevant de l'Éducation nationale.
8. Pour plus de détails sur la méthodologie utilisée dans le corpus LS-COLIN, voir Sallandre (2003, p. 92-100). Les vidéos sont téléchargeables sur le lien : <<https://cocoon.huma-num.fr/exist/crdo/search2.xql>>.
9. Les vidéos sont téléchargeables sur le lien : <www.ortolang.fr/market/corpora/ortolang-000926>.
10. Pour des détails sur la méthodologie utilisée dans le corpus Creagest, voir Garcia, L'Huillier & Vincent (2013).
11. Les deux annotateurs, l'un ayant la LSF pour langue première, l'autre pour langue seconde, ont effectué leurs annotations d'abord en parallèle puis ils les ont mises en parallèle afin de les homogénéiser. Il n'a pas été effectué d'étape d'évaluation de l'écart inter-annotateur de manière statistique, mais une étape de vérification a été réalisée systématiquement par l'annotatrice la plus expérimentée.
12. Le template ELAN est téléchargeable sur le lien : <www.ortolang.fr/workspaces/templates-lsf>.
13. Et, par souci de clarté, seules ont été retenues ici les annotations des tours de paroles du locuteur enquêté, qui représentent autour de 90 % du temps de parole, et pas celles de l'enquêteur.
14. Précisons que certaines UL sont présentes dans les TP (par exemple, dans les TP en discours rapporté, quand les propos entre les personnages du récit sont exprimés par des unités lexicales). Dans ce cas, ces UL sont comptabilisées dans les TP.
15. Grâce à la fonction 'chisq\$residuals' du logiciel R.
16. Fonction 'cor' du logiciel R.
17. L'ACP elle-même est réalisée par la fonction 'PCA' du logiciel R.
18. Dans une ACP, la variation est capturée par la variance, définie comme la moyenne des carrés des écarts à la moyenne d'une série statistique. La variance constitue l'un des indicateurs de variation fondamentaux.

RÉSUMÉS

Cet article présente une analyse de quatre genres discursifs en LSF : narratif, explicatif, argumentatif et dialogal. Les données sont issues de deux corpus, le corpus LS-COLIN (Cuxac et coll., 2002) et le corpus Creagest (Garcia, L'Huillier & Vincent, 2015) et ont été annotées avec les logiciels Excel et ELAN.

L'objectif est de voir s'il existe une systématique dans la fréquence et le type de catégories linguistiques au regard du genre discursif. Les catégories linguistiques utilisées sont celles proposées dans l'approche sémiologique (Cuxac, 2000) qui suggèrent l'existence de deux grands types d'unités en LSF, toutes deux morphologiquement compositionnelles : d'une part, les unités relevant d'une visée illustrative, et appelées « unités de transferts » et, d'autre part, les unités relevant d'une visée non illustrative que sont les unités lexicales, les pointages et la dactylographie. Des statistiques descriptives viennent compléter la vision globale des données fournies par les effectifs bruts et les proportions des catégories linguistiques pour chaque genre discursif. Par ailleurs, et c'est un fait intéressant, les statistiques retrouvent de manière automatique la répartition entre les deux grands types d'unités, alors que les données n'ont pas été annotées de manière explicite avec cette distinction.

This paper presents an analysis of four discourse genres in LSF: narrative, explanatory, argumentative and dialogical. The data are derived from two corpora, the LS-COLIN corpus (Cuxac et al., 2002) and the Creagest corpus (Garcia, L'Huillier & Vincent, 2015) and the annotations were made using Excel and ELAN software.

The aim is to see whether systematicity is visible in the frequency and type of linguistic categories with regard to discourse genre. The linguistic categories used are those proposed by the semiological approach (Cuxac, 2000), which suggests the existence of two main types of LSF units, both morphologically compositional: on the one hand, units with an illustrative intent, grouped under the name "transfer units" and on the other hand, units with a non-illustrative intent, i.e., lexical units, pointings and fingerspelling.

Descriptive statistics complement the overall view of the data provided by raw frequencies and the proportions of linguistic categories for each discourse genre. Moreover, and this is an interesting fact, the statistics automatically find the distribution between the two main types of units, while the data do not explicitly encode this distinction.

INDEX

Mots-clés : genre discursif, langue des signes française (LSF), corpus, unité lexicale, unité de transfert, statistiques

Keywords : discourse genre, French Sign Language (LSF), corpus, lexical unit, transfer unit, statistics

AUTEURS

MARIE-ANNE SALLANDRE

UMR 7023 SFL, Université Paris 8 – Université Paris Lumières et CNRS

ANTONIO BALVET

UMR Savoirs, Textes, Langage, Université de Lille, CNRS

GEOFFREY BESNARD

Université Paris 8

BRIGITTE GARCIA

UMR 7023 SFL, Université Paris 8 – Université Paris Lumières et CNRS