Etudier le Langage à l'Ere Numérique

Carmelo Alessandro Basile, Coline Caillol, Cameron Morin, Moisés Velasquez Perez, Chenyang Zhao, Ana Inés Ansaldo, Marie Bouchet, Edith Durand, Karl Seifen, Victoria Valentin

Étudier le Langage à l'Ère Numérique

Actes des 25èmes Rencontres des Jeunes Chercheurs (RJC) en Sciences du Langage

ED622 Sciences du Langage – Sorbonne-Nouvelle et Paris-Cité

Edité par Carmelo Alessandro Basile, Coline Caillol, Cameron Morin, Moisés Velasquez Perez & Chenyang Zhao

Sommaire

Repenser l'analyse de corpus numériques issus de sites web dynamiques

Marie BOUCHET

CLILLAC ARP, Université Paris Cité

marie.bouchet@u-paris.fr

RESUME

Les corpus numériques sont de plus en plus abordés en analyse de discours, ainsi qu'en langues de spécialité. L'objectif de cet article est d'élaborer une méthode d'analyse adaptée aux corpus numériques issus de sites web dynamiques. Pour cela, nous avons identifié les lacunes des corpus actuels à partir d'un corpus textuel traditionnel dans le domaine de l'accès aux droits. Nous avons ensuite identifié les caractéristiques principales d'un corpus numérique issu de sites web dynamiques. Enfin, nous nous sommes appuyés sur les recherches faites sur les réseaux sociaux et les sites web en analyse de discours afin de concevoir et tester notre modèle d'analyse de corpus adaptée aux sites web dynamiques. Ce modèle se décline en trois étapes : 1) Élaboration de sous-corpus à partir de parcours utilisateurs. 2) Annotation des corpus avec les fonctions des différentes parties de chaque page. 3) Exploration du corpus en prenant en compte sa dimension interactive. Nous avons ainsi réalisé l'annotation d'un corpus pilote issu du site officiel de l'administration française service-public.fr, qui a permis d'obtenir des résultats statistiques préliminaires. D'après ces résultats, notre méthode semble pertinente et doit être approfondie. Les premières analyses ont montré qu'une étude par parcours utilisateur propose une complémentarité à l'étude par corpus classique, car elle permet de prendre en compte l'ensemble de l'environnement discursif.

Mots-clés *: discours numérique, analyse du discours, parcours utilisateur, linguistique de corpus, accès aux droits.*

ABSTRACT

Digital corpora are increasingly influential in discourse analysis and languages for specific purposes. The aim of this article is to identify the main characteristics of a digital corpus extracted from dynamic websites in order to implement an efficient and accurate analysis model. We first identified shortcomings in conventional corpora analysis using a traditional textual corpus in the field of social rights. We then defined the main characteristics of a digital

corpus extracted from dynamic websites. Finally, we drew on research done on social media and websites corpora in discourse analysis to design and test our corpus analysis model. This model was developed in three steps: 1) Identify user paths to create sub-corpora. 2) Annotate the corpus with the functions of the different parts of each page. 3) Explore the corpus by taking into account its interactive dimension. We thus carried out the annotation of a pilot corpus from the official French government website service-public.fr. This pilot corpus produced preliminary statistical results. According to these results, our model appears to be relevant and will be extended to a larger corpus. This pilot study, relying on user paths, was successful in providing a comprehensive analysis of a discourse and its environment when coupled with the initial textual corpus study.
Key words*: digital discourse – discourse analysis – corpus linguistics – user path – social rights*

1. INTRODUCTION

En France, lors des dix dernières années, l'administration a subi une transformation numérique. L'ensemble des services d'accès aux droits sont désormais accessibles en ligne. Nous observons une dématérialisation des rapports entre les usagers et l'administration. Un grand nombre de démarches administratives se fait désormais en ligne : inscription à pôle emploi, demande d'acte de naissance ou de casier judiciaire, ou encore vote pour les Français de l'étranger. Cette transformation est accompagnée d'une modification des rituels d'interactions : les sites web sont désormais le lieu principal d'interactions avec l'administration, pour chercher des informations ou poser des questions. Ces sites sont dynamiques et contiennent des éléments interactifs, qui peuvent poser des difficultés d'analyse et d'intégration au corpus.

Nous nous intéressons dans cette étude aux nouvelles dynamiques discursives introduites lors de la transformation numérique du discours de l'accès aux droits. L'étude du discours de l'accès aux droits consiste ainsi à aborder et explorer non pas un support textuel, mais un support numérique : le site web dynamique. Les sites web dynamiques représentent de nouveaux terrains linguistiques qui ont fait l'objet de peu d'études en linguistique de corpus (Cacchiani, 2018). Nous nous demandons comment adapter les méthodes d'analyse de corpus aux corpus issus de sites web dynamiques.

Nous avons choisi de mettre au point une nouvelle méthode d'analyse à partir des caractéristiques principales d'un corpus numérique dynamique. Les trois caractéristiques principales de ce type de corpus sont : la délinéarisation du discours, la fonctionnalité de la page

web et la construction dynamique du discours. Ces caractéristiques ont permis d'élaborer un modèle d'analyse, ainsi qu'un corpus pilote.

Dans cet article, nous exposons dans un premier temps les recherches concernant les corpus numériques, ainsi que les données et le matériel utilisés dans notre étude. Nous détaillons ensuite la mise en place d'un modèle d'analyse à partir d'un corpus pilote. Enfin, nous présentons les premiers axes d'analyse et les perspectives de notre modèle.

2. ÉTAT DE L'ART

2.1. La linguistique de corpus

D'après Sinclair, repris par (Habert, 2000), un corpus est « une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques et extra-linguistiques explicites pour servir d'échantillon d'emplois déterminés d'une langue ». Cette définition met en lumière l'importance de la sélection et de l'organisation des textes en corpus réalisée par le chercheur en fonction des objectifs d'étude prédéfinis. Puisqu'il s'agit de données collectées pour servir d'échantillon, il est nécessaire que les textes choisis soient représentatifs du discours étudié. Pour cela, il convient d'établir des critères de sélection clairs et de définir les limites de notre analyse (Biber, 1993). De même, la taille du corpus dépend des critères d'analyse que nous définirons dans la partie 3.1.

De nos jours, les corpus sont stockés et utilisés sous forme numérique. Ce sont donc des « données langagières », mais également des textes lisibles par une machine[1] (McEnery & Wilson, 2001). Ils sont préparés et encodés afin d'être lus par des logiciels.

Le numérique a une place importante dans le développement de la linguistique de corpus, dont le développement coïncide avec l'ère du web 1.0 entre 1989 et 2005. L'ouvrage canonique de Sinclair *Corpus, Concordance, Collocation*, publié en 1991, met en avant le rôle important qu'a joué l'informatique dans le développement de la linguistique de corpus et de ses principes. Depuis 1991, les possibilités offertes par le numérique n'ont fait que s'accroître, et la linguistique de corpus a bénéficié de nombreuses innovations. Depuis le British National Corpus (BNC), le corpus de référence créé par l'Université d'Oxford, collecté dans les années 1980 à 1990 et composé de 100 millions de mots, les corpus ont évolué de façon fulgurante, à l'instar de l'English Web Corpus (enTenTen), collecté automatiquement et qui

[1] Citation originale : « machine-readable text » (McEnery & Wilson 2001 : 32).

rassemble 36 milliards de mots. Pourtant, à son époque, le BNC était une innovation en termes de nombre de mots (Rundell, 2008). Cette différence d'échelle montre les progrès importants accomplis depuis le début de la linguistique de corpus.

De nos jours, les corpus sont un outil incontournable dans plusieurs domaines, tant en sciences humaines qu'en sciences dites dures, notamment car le numérique permet de collecter des corpus de grande taille rapidement et de façon automatique. Les données récoltées peuvent ensuite être préparées et compilées en fichier afin de réaliser des analyses. Le numérique permet de les explorer de façon systématique, ce qui est également une grande avancée, car des logiciels permettent de faire des analyses complexes en quelques secondes.

Un travail de préparation est nécessaire pour utiliser un corpus. Tout d'abord, le corpus est récolté, puis nettoyé. Un grand nombre d'informations et de textes jugés comme « parasites » peuvent être retirés. La mise en forme est également effacée. L'encodage et le format de fichier original sont changés. Il est parfois annoté (manuellement ou automatiquement), avant d'être exploité par un logiciel.

Nous proposons donc une définition pratique du corpus « classique » actuel. Un corpus est une collection de données langagières nettoyées, encodées et formatées pour être exploitées par un logiciel d'exploration de corpus.

Les textes utilisés peuvent provenir de supports imprimés, qu'il est nécessaire de numériser, ou bien de supports numériques. Dans les deux cas, le support utilisé est le plus souvent statique et comprend peu d'éléments interactifs. Cependant, certaines recherches s'intéressent aux textes numériques dits « natifs », qui sont des textes « produits originellement au sein d'espaces connectés » (Mayeur & Paveau, 2020 : 1), et qui ont des propriétés spécifiques, dont, depuis le web 2.0, leur caractère dynamique.

2.2 Les discours numériques

L'analyse de discours et les sciences du langage en général se sont récemment intéressées aux corpus numériques natifs en tant qu'objet d'étude. Ces discours, produits dans un univers numérique et consultables dans un environnement numérique, possèdent certaines caractéristiques communes.

Le discours étudié est considéré en analyse des discours numériques comme des productions numériques natives, un concept institué par Marie-Anne Paveau, afin de décrire ces nouveaux types d'écrits :

> « On appelle natives les productions élaborées en ligne, dans les espaces d'écriture et avec les outils proposés par internet, et non portées après

> numérisation d'espaces scripturaux et éditoriaux prénumériques aux espaces numériques connectés. » (Paveau, 2017 : 27)

Cette définition, ainsi que les travaux de la chercheuse ont mis en lumière l'importance d'étudier les discours numériques selon de nouveaux paramètres. Les discours étudiés ont été pour l'instant largement issus des réseaux sociaux (Emerit, 2016; Ghliss, 2019; Paveau, 2013). En effet, les études sur les discours des réseaux sociaux ont été les premières à se développer et à gagner de l'importance, car les réseaux sociaux interrogent de façon radicale nos habitudes discursives. Notre objet d'étude est le site web dont le développement a eu lieu plus progressivement. À titre d'exemple, le site service-public.fr a été mis en place en 2000, puis n'a cessé d'évoluer au cours des 20 dernières années. La discursivité d'un site web peut sembler moins radicalement transformée par le numérique, car la dimension interactive est moins présente. Sur les sites choisis pour cette étude, l'utilisateur ne peut par exemple pas interagir avec les autres utilisateurs. Cependant, les textes présents sont bien pensés et conçus sur des outils numériques, au même titre que les discours des réseaux sociaux. L'utilisateur doit également interagir avec la page web pour naviguer. Ces discours répondent ainsi à de nouvelles normes : celles des sites web.

2.2.1 Le site web

Un site web est un ensemble de pages web liées entre elles, comprenant une page d'accueil, et hébergé au même endroit (Web Characterization Terminology & Definitions Sheet, 1999). Il est généralement édité par une seule entité ou organisation. Cette dernière peut proposer un guide de style linguistique (destiné aux éditeurs et éditrices du site) et une identité visuelle. Une page web est un ensemble d'informations, composée d'une ou plusieurs ressources Web, destinée à être consultée comme une entité, et désignée par un identifiant unique (Web Characterization Terminology & Definitions Sheet, 1999).

Dans les sites sélectionnés pour notre étude, chaque page est mise en forme en HTML (Hypertext Markup Language), un langage de balisage permettant au webmaster de hiérarchiser et de mettre en forme les contenus. Les contenus discursifs doivent donc être pensés pour correspondre aux normes du site. Dans notre étude, nous nous intéressons à un type de site particulier, qui a vu le jour avec le web 2.0 : les sites dynamiques.

Un site web est dynamique lorsque les pages qu'il contient sont générées à la demande par le serveur, selon les requêtes de l'utilisateur. Il est donc nécessaire pour l'utilisateur d'interagir avec la page pour accéder à certains contenus. Le contenu des sites dynamiques est

donc instable. À l'inverse, un site statique est visible tel qu'il a été conçu et n'utilise pas de base de données. Il peut cependant avoir certaines animations (musiques, vidéos, menus…). Notre objet d'étude est donc le site web dynamique : un ensemble de pages avec lequel l'utilisateur doit interagir afin d'accéder à l'ensemble des ressources.

2.2.2 La fragmentation du discours

L'interaction nécessaire entre l'utilisateur et la page crée un nouveau type d'environnement discursif. Les informations ne sont pas accessibles de façon linéaire, et l'utilisateur doit cliquer et se repérer sur la page ou dans l'arborescence du site. Les textes issus de sites web dynamiques présentent notamment des traits de délinéarisation. La délinéarisation est « une élaboration du fil du texte dans laquelle les matières technologiques et langagières sont co-constitutives, et modifient la combinatoire phrastique en créant un discours composite à dimension relationnelle » (Paveau, 2015 : 14). Ces caractéristiques de sites web ont été également remarquées par Dominique Maingueneau qui les décrit en termes de « modules hétérogènes » et de « mosaïques » dans son analyse du web. Pour lui « l'écran ne propose qu'une vue partielle d'une totalité qui ne se donne jamais intégralement : il y a divergence entre les scansions d'Internet et la pagination de l'imprimé. » (Maingueneau, 2016 : 4).

Les deux chercheurs font donc les mêmes constats concernant le fil du discours numérique : il n'est plus linéaire, mais bien modifié par son environnement : il est « composite » ou ce sont des « scansions ». Ces deux termes décrivent la présence de plusieurs éléments séparés qu'il est nécessaire d'assembler afin de former un tout. Ils semblent particulièrement adaptés, car la page est effectivement composée de différents modules, délimités par des balises, qui sont rassemblés afin de former une entité. Le discours est conçu suivant les éléments technologiques du site : hyperliens, menus déroulants, mais aussi selon un système de pages imbriquées. Ces éléments sont particulièrement omniprésents, car ils permettent de naviguer dans l'arborescence du site.

Néanmoins, cette délinéarisation permet également à l'utilisateur de construire son propre fil du discours. En faisant des choix sur la page, l'utilisateur construit un fil discursif allant de page en page selon ses besoins et ses choix.

2.2.3 Les parcours utilisateur

Lorsque l'utilisateur se déplace de page en page, il crée un parcours utilisateur. « Un parcours sur le Web s'apparente à un parcours de lecture et d'action dans le cadre d'une navigation hypertextuelle » (Beauvisage, 2004). Le parcours utilisateur fait donc référence aux différentes

pages visitées par l'utilisateur dans l'ordre choisi par ce dernier. Ce concept est particulièrement utilisé en marketing et en ergonomie afin d'orienter l'utilisateur d'un site marchand vers l'achat.
Nous voulons dans cette étude nous réapproprier ce concept afin d'envisager le parcours utilisateur en tant qu'outil d'analyse du discours numérique natif. En effet, certains analystes du discours numérique font référence à la façon dont le discours prend forme avec la navigation. « L'espace virtuel est un espace qui se structure au fur et à mesure qu'on y navigue. » (Vitali-Rosati, 2014 : 4.2.3). Le parcours utilisateur sera donc une unité d'analyse de notre étude.

2.2.4 Affordances

La possibilité de naviguer sur la page et de construire son fil discursif à travers la navigation est due à une caractéristique intégrante du numérique : l'affordance.
Initialement, le terme affordance est issu de la psychologie et utilisé pour la première fois par Gibson (1977) pour faire référence aux possibilités d'action offertes à un animal par son environnement en fonction de ses capacités. En français, on parle parfois de potentialité. L'affordance décrit la « coévolution entre les acteurs et leur environnement » (Ostern & Rosemann, 2021 : 1). C'est un concept qui est donc applicable aux technologies et notamment au numérique, et qui a été repris par différents domaines en lien avec le numérique, en particulier celui de l'expérience utilisateur.

Pour notre objet d'étude, l'affordance fait référence à des possibilités ou des contraintes offertes par une page web, mais aussi à la capacité d'un dispositif, la page, à suggérer son utilisation (Burger, 2018). Autrement dit, l'utilisateur d'un site web construit son parcours utilisateur en s'appuyant sur ses objectifs, ses capacités, mais aussi selon les potentialités incluses par les concepteurs du site.

2.3 Nouvelles méthodes pour l'analyse des discours numériques natifs

Afin d'étudier les textes numériques, certaines chercheuses ont mis au point de nouvelles méthodes d'analyse qui prennent en compte leurs caractéristiques.

Pour parler des réseaux sociaux et notamment de Facebook, Laetitia Emerit (2016) choisit d'introduire la notion de « lieu de corpus ». Il s'agit d'un « lieu à partir duquel il est possible de créer des corpus numériques et jusqu'auquel il est nécessaire de remonter pour interpréter ces corpus » (Emerit, 2016). Cette notion est introduite par la chercheuse afin de proposer un complément au corpus qu'elle caractérise comme « stabilisé », donc extrait de sa page web native et décontextualisé. Le corpus et le lieu de corpus sont donc complémentaires dans l'analyse du discours numérique natif. Ces recherches mettent en avant les limites des corpus tels qu'ils sont utilisés en science du langage et tentent de les dépasser. Les données

doivent être extraites de leur environnement et stabilisées pour être étudiées, tandis que les textes numériques natifs ne cessent d'être actualisés. Il est donc important de s'intéresser à l'environnement du discours.

Mel Stanfill (2015) propose dans son étude d'analyser l'environnement discursif du site web. Elle élabore une méthode : *discursive interface analysis*, qui consiste à étudier les fonctionnalités de la page, ses menus, son organisation, ses caractéristiques, afin d'en dégager des normes. En s'appuyant sur les théories de Foucault (1972) concernant le discours, Mel Stanfill étudie la manière dont l'interface est construite. Un site est construit selon certains objectifs. Par exemple le site officiel de l'administration française est construit dans le but d'informer les usagers. À partir de ces objectifs, l'interface donne alors à l'usager des possibilités de navigation et d'action[2]. Ces possibilités sont étudiées dans la recherche avec le concept d'affordance.

Les discours numériques sont formatés par les affordances de leur dispositif digital (Burger, 2018). Nous pouvons en conclure que l'interface conditionne le discours étudié. Il est donc essentiel de l'intégrer à une étude du discours.

Les deux méthodes présentées nous permettent d'envisager les particularités et les difficultés que présente l'étude d'un site web. Nous avons également présenté les concepts d'affordance et de parcours utilisateur, qui seront les fondamentaux du modèle développé dans les parties suivantes. La méthode du lieu de corpus s'applique particulièrement aux réseaux sociaux, ce qui n'est pas notre objet d'étude, mais elle propose une appréciation utile des discours numériques que nous allons explorer, celle de la nécessité de prendre en compte l'environnement discursif initial. La *discursive interface analysis*, même si elle prend en compte le discours, est une méthode d'analyse ergonomique et sociologique du site web. Il convient donc de mettre en place une méthode d'analyse d'un corpus issu d'un site web dynamique, en s'inspirant des méthodes adaptées à notre objet d'étude.

3. MATERIEL ET DONNEES

3.1. Le corpus

Le corpus initialement constitué pour cette étude est composé de sites web dynamiques publiés par des organisations officielles d'accès aux droits et de service public. Il est bilingue, et donc divisé en deux sous-corpus : un français et un britannique. Il comprend des sites

[2] Citation originale : "a site's design makes a normative claim about its purpose and appropriate use that both demonstrates an understanding of users and builds a set of possibilities into the object" (Stanfill, 2015 : 1060).

gouvernementaux comme le site officiel de l'administration française service-public.fr, mais aussi des sites issus d'organisations privées chargées de service public comme le site de la Caisse d'allocations familiales. Afin de sélectionner les sites étudiés, nous les avons divisés en trois thèmes : Aides sociales, Immigration et Citoyenneté.

Le corpus Aides sociales est composé de quatre sites pour le français : le site de la Caisse d'Allocations Familiales (caf.fr), le site du Crous de Paris (crous-paris.fr), la partie correspondant aux aides sociales du site web officiel de la ville de Paris (paris.fr) et enfin la partie correspondant aux aides sociales du site officiel de l'administration française (service-public.fr).

Le corpus Aides sociales est composé de six sites pour le corpus anglais britannique : la partie aides sociales des sites des conseils municipaux de Bornemouth, Bristol et Falkirk. La partie aides sociales du site officiel du gouvernement du Royaume-Uni (gov.uk) et du Citizen Advice Bureau (citizensadvice.org.uk). Le site de l'agence pour l'emploi (jobcentreguide.co.uk) et understandinguniversalcredit.gov.uk.

Le corpus Immigration est composé de deux sites pour le français : la partie immigration des sites service-public.fr et du Service d'information à destination des talents étrangers et de leur famille (welcometofrance.fr).

Le corpus Immigration est composé d'un site pour l'anglais britannique : la partie immigration de gov.uk. En effet, il n'existe pas d'autres sites officiels d'information concernant les droits des étrangers.

Le corpus Citoyenneté est composé de quatre sites pour le français : la partie citoyenneté de service-public.fr, et des sites officiels des villes de Paris, Marseille et Rennes.

Le corpus Citoyenneté est composé de quatre sites pour l'anglais britannique : la partie citoyenneté des sites des Councils (municipalités) de Bristol, Falkirk et York, et gov.uk.

Lorsque seulement une partie du site a été sélectionnée, les pages incluses ont été sélectionnées manuellement à partir de mots-clés extraits d'un corpus pilote.

Le tableau ci-dessous présente un récapitulatif de notre corpus en termes de taille.

	Français (FR)	**Anglais (R-U)**
Aides sociales	722 000 mots — 4 sites	211 000 mots — 6 sites
Citoyenneté	711 000 mots — 4 sites	303 000 mots — 4 sites

Immigration	1 233 000 mots — 2 sites	124 500 mots — 1 site

Tableau 1 Récapitulatif du corpus

Le Tableau 1 montre que le corpus français comporte plus de mots que le corpus britannique. Cela est dû à la législation, mais aussi au nombre de sites et de dispositifs disponibles en France. Les sites choisis pour ce corpus sont représentatifs des sites d'accès aux droits disponibles en France et au Royaume-Uni. On peut donc remarquer que la France multiplie les sites et les dispositifs tandis que le Royaume-Uni présente une tendance inverse.
Il s'agit d'un format de corpus classique utilisé en science du langage. Les pages HTML sont nettoyées et transformées en un unique fichier par site au format texte brut.

Pour être exploité, ce corpus doit correspondre aux critères présentés dans la partie 2.1 : il doit notamment être lisible par une machine. Il est donc nécessaire de préparer les fichiers qui vont être intégrés aux corpus. Pour cela, nous avons établi des critères de sélection afin de nettoyer les corpus. Comme nous nous intéressons au discours de l'accès aux droits, nous avons décidé de garder uniquement le contenu principal de la page. Pour cela, nous avons utilisé le code HTML de la page. Ce critère de sélection nous a permis de faire une sélection objective des textes retenus dans le corpus.

Les textes sélectionnés ont ensuite été extraits et nettoyés automatiquement par le logiciel KOALA (3,2). Chaque donnée sélectionnée pour un site a été rassemblée en un fichier texte brut, prêt à être exploité par un logiciel d'exploration de corpus.

3.2. Outils

Le logiciel KOALA a été créé pour notre étude par Amine Benamara (LISN, Université Paris Saclay). Il a servi à extraire la balise HTML sélectionnée, mais aussi à nettoyer le texte afin d'enlever les balises HTML et toute trace de code. Le logiciel est en cours d'actualisation afin de proposer une visualisation dynamique des pages HTML et .txt.

Afin d'explorer notre corpus dit classique, nous avons utilisé le logiciel Sketch Engine (Kilgarriff et al., 2014) d'exploration de corpus qui permet des annotations automatiques en catégories prédéfinies (partie de discours, collocations, etc.). À titre d'exemple, la recherche de collocation par catégorie syntaxique nous a été particulièrement utile.

L'annotation, ainsi que les résultats préliminaires obtenus sur le corpus pilote, ont été réalisés sur UAM Corpus Tool (O'Donnell, 2009). Ce logiciel a permis de construire un schéma

d'annotation, mais aussi d'annoter manuellement (ou automatiquement) l'ensemble du corpus. Des résultats statistiques sont ensuite disponibles à partir de l'annotation.

3.3. Étude préalable

À partir d'une étude réalisée avec le corpus présenté dans la partie précédente (3.1), nous avons identifié des lacunes des méthodes classiques. L'identification de ces lacunes nous a orientés vers la conception d'un nouveau modèle d'analyse. Nous avons donc choisi d'identifier les caractéristiques de nos textes, et du discours numérique en général, afin de voir en quoi celles-ci pouvaient être différentes d'un discours non numérique.

Nous avons effectué une phase d'observation, puis d'étude des textes. Tout d'abord, les textes ont été analysés grâce à des logiciels de corpus classiques afin de noter certaines caractéristiques linguistiques. Ensuite, nous avons navigué sur les sites et étudié leur code HTML afin d'identifier leur structure.

Ces observations et analyses nous ont menés à identifier trois caractéristiques principales sur lesquelles nous avons basé notre méthode d'analyse des corpus numériques.

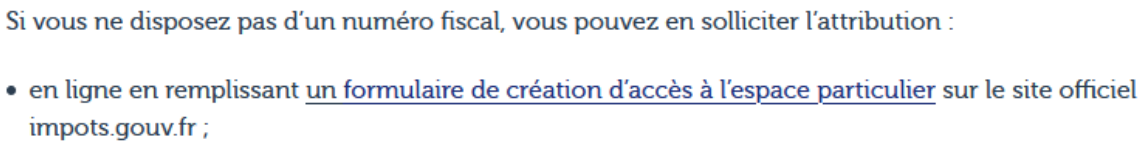
3.3.1. Limites des outils d'études classiques

Avec le logiciel Sketch Engine, nous avons étudié la collocation [pouvoir solliciter]. Celle-ci nous a permis de nous rendre compte des lacunes d'une étude textuelle classique. Par exemple, notre corpus possède un grand nombre de phrases répétées (Figure 1).

Figure 1 collocation de [pouvoir solliciter] sur Sketch Engine

Sketch Engine identifie des collocations grâce à une mesure statistique appelée logDice qui s'appuie sur la fréquence d'utilisation des collocats et des collocations (Rychlý, 2008). Nous avons identifié 72 occurrences de cette phrase exactement identiques. Et il ne s'agit pas de la seule phrase répétée : sur les 582 occurrences de la collocation [pouvoir solliciter], aucune phrase n'est unique. Il est évident que la présence de nombreuses phrases répétées peut fausser la mesure statistique qui identifie la cooccurrence des deux verbes comment étant significative à partir d'un corpus composé d'un grand nombre de phrases répétées.

Il est important de noter que la répétition de ces phrases est en partie due au domaine choisi : l'accès aux droits. En effet, le discours administratif est très normé, et les démarches sont souvent les mêmes pour différents profils. Par exemple, la transmission d'un avis d'imposition, dont il est question dans notre exemple, est commune à un très grand nombre de démarches. Il semble donc naturel que la même phrase soit répétée plusieurs fois. Cependant, cette redondance inhérente au discours administratif est accentuée par le site web, qui est construit en arborescence. Chaque page est indépendante et requiert donc une réitération des informations. Dans la phrase étudiée sur la Figure 1, la fonction de cette répétition est de rediriger l'utilisateur vers un service en ligne.


Si vous ne disposez pas d'un numéro fiscal, vous pouvez en solliciter l'attribution :

- en ligne en remplissant un formulaire de création d'accès à l'espace particulier sur le site officiel impots.gouv.fr ;

Figure 2 Contexte de la phrase étudiée

Cette observation est vérifiable en se rendant sur le site, et en regardant la phrase dans son contexte numérique, comme présenté dans la Figure 2 ci-dessus. Le syntagme nominal « formulaire de création d'accès à l'espace particulier » est en réalité un hyperlien. La fonction de cette répétition est donc de rediriger l'utilisateur. Un discours numérique natif étant délinéarisé, cette redondance peut être vue comme un moyen de la réduire. Il s'agit d'un effet de reprise communément observable sur les sites web (Pecman, 2018).

3.3.2. Particularités d'une page web

La méthode choisie (sélection des contenus, nettoyage du corpus et analyse avec un logiciel d'exploration) semble donc présenter des problèmes pour analyser le discours de l'accès aux droits. Comme nous l'avons présenté précédemment (2), le discours n'est pas constitué uniquement d'un texte. En effet, si le discours est un continuum entre le verbal et non-verbal, le corpus collecté ne saurait être une représentation fidèle du discours de l'accès aux droits.

Afin de proposer une analyse complète du discours, nous allons intégrer cet environnement discursif à notre étude. Pour cela, il est nécessaire de revenir au site web tel qu'il est affiché en ligne, ou ce que la chercheuse Emerit (2016) appelle « lieu de corpus » dans son étude des réseaux sociaux. L'analyse du site web dans son intégralité et la prise en compte de ses fonctionnalités pourra servir à mettre en lumière les particularités d'une page web. Tout comme pour le corpus, nous sommes dans l'obligation d'en présenter une version statique, ou stabilisée, par image.

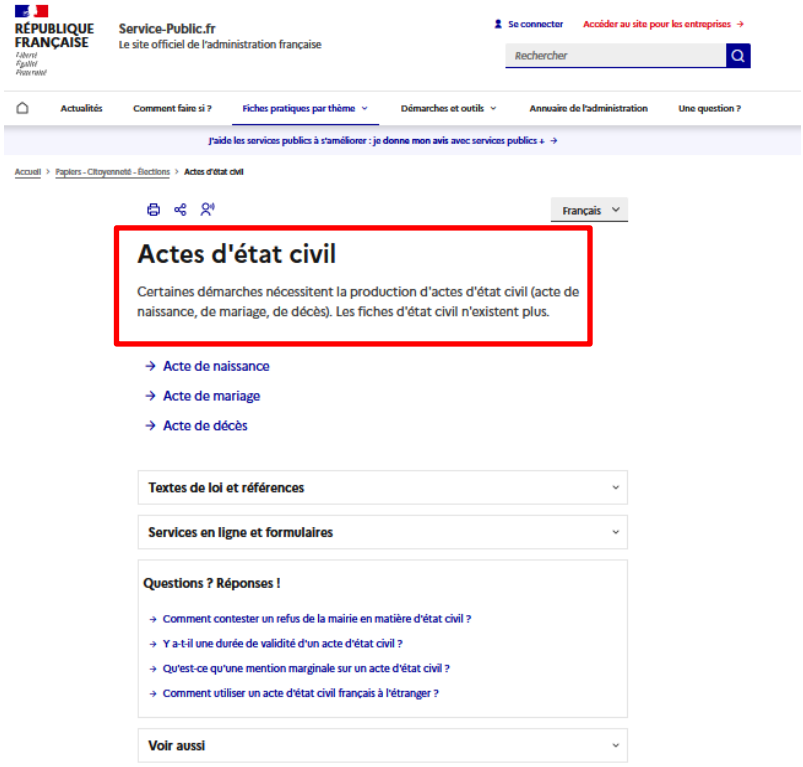


Figure 3 Page www.service-public.fr/particuliers/vosdroits/N359 au 5 septembre 2022

La Figure 3 est une capture d'écran d'une page du site service-public.fr, qui fait partie de notre corpus. Le carré rouge sur l'image encadre le texte retenu pour le corpus présenté dans la section 3.1. Il est évident qu'une majeure partie du texte présent sur la page n'est pas incluse dans le corpus. Pourtant, ce texte est présent sur la page. L'utilisateur interagit avec l'ensemble de la page, et pas seulement avec le contenu. De plus, l'ensemble de la page n'est pas visible sur cette capture d'écran. Certains éléments sont interactifs et l'usager doit cliquer afin de les faire apparaître. D'autres éléments ne peuvent pas être affichés simultanément.

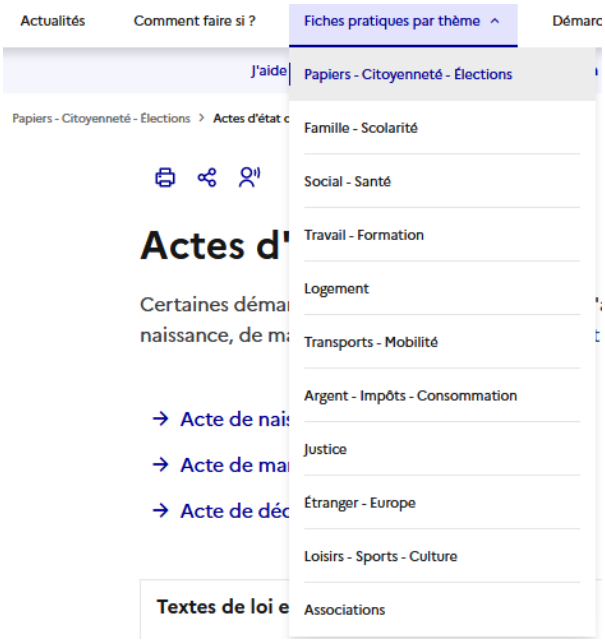


Figure 4 Menu déroulant de la page http://www.service-public.fr/particuliers/vosdroits/N359 au 5 septembre 2022

La Figure 4 présente une capture d'écran d'un menu déroulant provenant de la page présentée en Figure 3. Suivant les choix de l'utilisateur, le contenu et le texte affiché à l'écran sont transformés. Le contenu principal peut même être recouvert par un autre module : ici la barre de menu. Il ne semble donc pas représentatif de ce discours de ne considérer que le contenu principal de la page quand l'utilisateur peut interagir avec l'ensemble de la page web.

Sur la même page, on peut également trouver deux rubriques déroulantes (Figure 5 ci-dessous).

Figure 5 Menu déroulant de la page www.service-public.fr/particuliers/vosdroits/N359r au 5 septembre 2022

Sur la Figure 5 et la page web correspondante, nous observons deux rubriques complémentaires qui apparaissent en cliquant sur leur titre, respectivement « Textes de loi de références » et « Services en ligne et formulaires ». Le clic fait apparaître une série de liens s'apparentant à un menu.

L'analyse de cette page met donc en avant deux particularités : l'ensemble de la page web peut être considérée comme du contenu et chaque utilisateur construit son discours en naviguant.

4. MISE EN PLACE D'UN MODELE D'ANALYSE

Nous allons maintenant présenter le modèle d'analyse mis au point pour étudier les corpus issus de sites web dynamiques. Nous ferons une description générale du modèle avant de présenter ces trois caractéristiques principales.

4.1. Description générale du modèle

Le modèle pilote comprend trois étapes correspondant aux trois caractéristiques mises en évidence dans la partie précédente : la délinéarisation du discours, la fonctionnalité de la page web et la construction dynamique du discours.

Tout d'abord, afin de prendre en compte la délinéarisation du discours et la fonction de navigation sur les sites web, les corpus seront formés à partir de parcours utilisateur. Ces parcours permettront d'identifier le fil du discours tel qu'il est perçu par l'utilisateur. Nous prendrons donc en compte la navigation page par page de l'utilisateur, au lieu d'étudier un ensemble de texte rassemblé en un fichier.

Ensuite, afin de prendre en compte la cohérence de la page, un schéma d'annotation a été mis en place à partir des fonctions du texte. Une page web est composée de différents modules qui répondent chacun à un objectif et forment une mosaïque (Maingueneau, 2016). Afin de prendre en compte la spécificité de chaque module de la page, il est nécessaire d'indiquer les différentes parties de chaque page, qui remplissent des fonctions différentes. Pour cela, nous nous sommes inspirés des travaux de Swales (2004) sur les *moves*, que nous expliquons en détail dans la partie 4.3.

Enfin, afin d'intégrer l'affordance de l'environnement discursif, le corpus est composé en prenant en compte la dimension interactive : les pages de chaque corpus sont sélectionnées selon la possibilité de cliquer sur un lien pour y accéder. Cela permet, en complément de la création d'un parcours utilisateur, de ne pas transgresser les normes ou les objectifs du site. Les

liens de chaque page sont également annotés afin d'étudier le corpus en prenant en compte sa dimension interactive et ses potentialités.

4.2. Sélection du parcours utilisateur

Une partie importante du texte présent sur la page a été écarté en sélectionnant le contenu principal lors de la création du corpus présenté à la partie 3.1. Cependant, l'usager interagit avec l'ensemble du texte du site, et pas seulement avec le contenu principal. Ainsi, afin de réintégrer le discours dans son intégralité, nous avons décidé de recréer des parcours utilisateurs, qui sont des parcours de navigation sur le site.

Pour l'étude pilote, les parcours utilisateur sélectionnés seront ceux intégrés par le site. Chaque page web a une affordance : des potentialités co-conçues avec l'utilisateur et la technologie. Pour construire ce premier parcours utilisateur, nous avons utilisé la terminologie du domaine ainsi que les potentialités incluses dans le site web.

Nous avons choisi un terme, correspondant à un document administratif précis : l'Acte de naissance. Nous avons ensuite navigué sur les pages comme un utilisateur cherchant la page correspondante. Notre parcours a commencé sur la page « Accueil Particulier », puis nous avons suivi les indications en cliquant sur la catégorie « Papiers – Citoyenneté » dont la légende indiquait « Etat-civil ». Nous avons ensuite sélectionné la page « Actes d'état civil », et enfin la page « Acte de naissance ». Afin de compléter ce parcours, et de correspondre aux critères définis : une navigation autour de l'Acte de naissance, nous avons également visité les pages « Qu'est-ce qu'une mention marginale sur un acte d'état civil ? » et « Demande d'une copie d'un extrait conservé au répertoire civil ».[3]

[3] Le corpus parcours utilisateur a été composé en avril 2022. Le site a depuis été modifié.

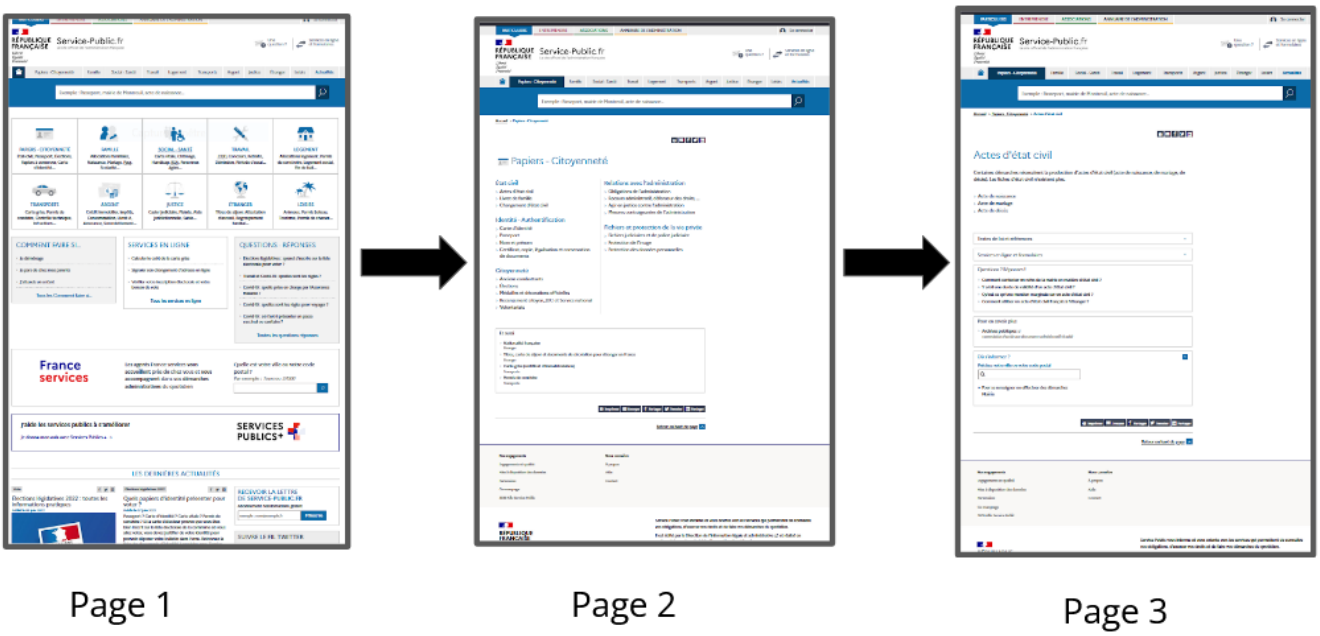


Figure 6 Illustration d'un parcours utilisateur au 2 avril 2022[4]

La Figure 6 ci-dessus est une représentation d'une partie du parcours utilisateur. Comme le montrent les images, l'ensemble de la page est pris en compte, des bannières aux mentions légales. Nous avons ainsi récupéré l'ensemble des pages HTML du parcours, que nous avons numérotées de 1 à 6. Les balises HTML ont ensuite été nettoyées afin de ne laisser qu'un fichier texte, sans mise en forme. Le corpus obtenu est de 25 000 mots.

Une fois le parcours utilisateur pilote collecté et mis en forme, nous avons pu procéder à la mise en place d'un schéma d'annotation.

4.3. Mise en place du schéma d'annotation

Nous nous sommes inspirés des travaux de Swales concernant les *moves* (2004) pour la création du schéma d'annotation. A l'origine, les *moves* sont utilisés pour analyser les genres académiques. Cette approche consiste à identifier les séquences ou étapes (appelées *moves*) d'un article de recherche. Les unités discursives qui forment le texte sont ainsi annotées selon leurs fonctions communicatives. Cette approche permet donc d'analyser la place de chaque unité discursive dans le fil du discours. Elle a ensuite servi à décrire la structure générique sous-jacente de plusieurs genres (Moreno et Swales, 2018). Elle pourrait donc servir à décrire la structure du discours de l'accès aux droits dans le cadre de sa caractérisation. Cependant, afin de créer notre schéma d'annotation, il convient d'identifier les fonctions spécifiques à notre discours. Selon Bathia (2014), les *moves* sont inhérents au texte : ce ne sont pas des

[4] L'interface du site a été modifiée en août 2022.

constructions des lecteurs. Pour identifier les différentes fonctions de notre texte, nous nous sommes donc intéressés au code HTML, qui met en forme la page. Il s'agit d'un critère externe qui permet d'identifier les fonctions des différentes parties du discours. Sur le site gov.uk, nous avons identifié les fonctions « browse » et « container », qui correspondent respectivement aux fonctions « navigation » et « contenu » de notre schéma. Sur le site service-public.fr, les balises ont des attributs nommés « role », qui indiquent le rôle du segment dans la page. On retrouve par exemple les rôles « navigation » et « main », qui correspondent aux fonctions du même nom dans le schéma. L'analyse des codes HTML des deux sites nous a permis de créer différentes catégories de textes présents sur la page web. Cette étude a été complétée par une analyse des textes en ligne en double vue HTML / page web.

Cette analyse nous a permis de mettre en place un schéma d'annotation pilote commun pour les sites gov.uk et service-public.fr, qui est représenté dans la figure ci-dessous.

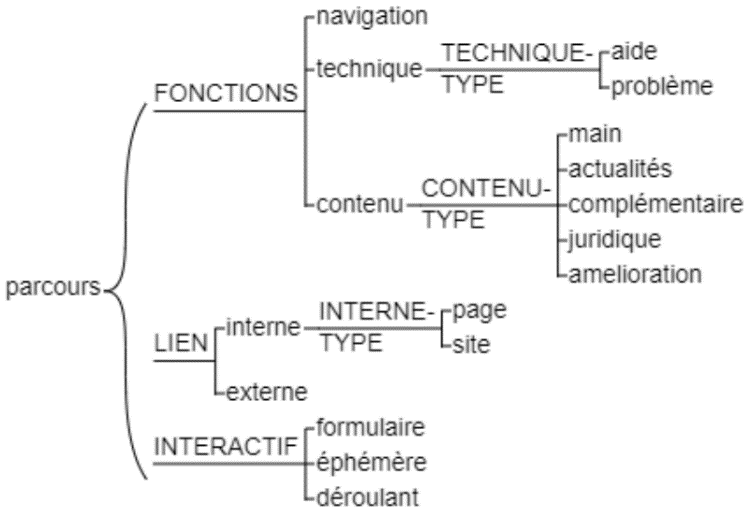


Figure 7 Schéma d'annotation construit sur UAM Corpus Tool

Le Figure 7 ci-dessus est une représentation du schéma d'annotation construit sur UAM Corpus Tool. Ce schéma a permis d'annoter le parcours utilisateur pilote. Tous les termes présents dans la figure (et mis en gras ci-dessous) correspondent à une couche d'annotation.

Le texte issu d'un site web correspond à trois fonctions principales. Tout d'abord, une grande partie sert à naviguer dans l'arborescence du site : c'est le cas des menus par exemple. La première fonction est donc **Navigation**. Ensuite, chaque page contient du **Contenu**. Ce contenu correspond à l'objectif de la page : la ressource contenue sur la page web que l'utilisateur vient chercher. Enfin, une partie du texte présent sur un site web concerne le site lui-même, c'est ce que nous avons appelé **Technique**. Il s'agit alors soit de l'annonce d'un **problème** technique

(généralement à l'aide d'une fenêtre pop-up), ou bien d'une **aide** à la navigation (informations complémentaires pour remplir des formulaires par exemple).

Cinq types de contenu différents ont été identifiés : le contenu **principal** (appelé « **main** », en référence à la balise HTML), les **actualités**, le contenu **complémentaire**, le contenu **juridique**, et le contenu **amélioration**. Dans le cas de notre parcours, le contenu principal de la page « Acte de naissance » donne des informations sur cette pièce, comment et où l'obtenir, etc. Le contenu **complémentaire** concerne les informations adjacentes. Sur le site service-public.fr, les informations complémentaires sont par exemple les rubriques Questions — Réponses, qui peuvent apporter des précisions sur des termes de la page. Les **actualités** sont généralement des contenus éphémères liés au sujet principal de la page. Par exemple, depuis le début de la guerre en Ukraine, des bannières et des pop-ups apparaissent sur les pages concernant les réfugiés ou les démarches exceptionnelles pour obtenir une carte de séjour. Ces pop-ups seraient alors annotés comme actualité. Enfin, l'annotation contenu **juridique** correspond aux informations juridiques concernant le site (licence, condition d'utilisation, etc.).

Comme nous l'avons mis en avant à multiples reprises, le site est interactif et renferme de multiples possibilités inscrites dans son design, des affordances, qui permettent aux utilisateurs de naviguer et de construire le fil de leur discours. Afin de rendre compte de ces particularités dans notre analyse, nous avons intégré deux catégories d'annotations : les **Liens** et les éléments **Interactifs**.

Pour les liens, nous avons annoté s'ils étaient **internes** ou **externes**, c'est-à-dire s'ils renvoient sur le site où se trouve déjà l'utilisateur, ou sur un autre site. Pour les liens **internes**, nous avons annoté s'ils redirigeaient sur la même **page**, ou sur le **site** en général.

Nous avons identifié plusieurs types éléments **Interactifs**. Tout d'abord, les **formulaires**. Notre schéma n'a pas pour but d'annoter les formulaires administratifs, mais bien les formulaires numériques. Un formulaire est un espace (généralement un rectangle), où l'utilisateur peut transmettre des informations aux serveurs du site web. Nous avons ensuite annoté les éléments **éphémères**, qui sont les pop-ups ou informations d'actualités. Enfin, nous avons annoté les **menus déroulants**, c'est-à-dire les textes sur lesquels l'utilisateur doit cliquer afin de faire apparaître du texte.

Une fois le schéma mis au point, nous avons réalisé l'annotation de façon manuelle sur UAM Corpus Tool. L'annotation est réalisée en double écran avec fichier .txt dans l'interface du logiciel et la page web disponible en ligne et navigable.

5. RESULTATS ET DISCUSSION

Les résultats générés par le premier parcours utilisateurs proviennent d'un corpus pilote de 25 000 mots. Ce parcours a été composé autour d'un terme du domaine. Les données et résultats obtenus devront être étayés par d'autres parcours utilisateurs afin de proposer une étude plus complète de ce discours.

5.1. Résultats statistiques préliminaires

Des premières analyses ont été ensuite réalisées par UAM Corpus Tools. Ces résultats concernent pour l'instant les statistiques d'un parcours.

Nous allons présenter dans cette partie ces résultats préliminaires, qui nous permettront de mettre en lumière certaines caractéristiques du parcours, mais aussi d'évaluer sa pertinence.

Fonctions	**Occurrences totales**	**Pourcentage**
Navigation	259	25
Technique	185	18
Contenu	570	55
Total	1 014	98

Tableau 2 Pourcentage moyen de chaque fonction dans le parcours — UAM Corpus Tools

Le Tableau 2 ci-dessus concerne le pourcentage moyen de chaque fonction dans le parcours. Il s'agit donc d'une moyenne réalisée sur l'ensemble des pages. Le texte annoté comme étant du contenu ne concerne que 55 % de la page. La navigation concerne quant à elle 25 % du texte présent sur la page. La partie technique est présente sur 18 % du texte. Enfin, ce tableau indique qu'environ 2 % des segments n'ont pas été annotés. Le contenu de la page ne représente que la moitié du texte présent sur la page en moyenne. Il existe plusieurs types de contenu, il est donc important d'observer la répartition des différents types de contenu sur la page (Tableau 3 ci-dessous). Cette répartition nous donne des informations précieuses sur le discours de l'accès aux droits.

Fonctions	Occurrences totales	Pourcentage
Main	53	55
Actualité	10	10
Complémentaire	12	13
Juridique	4	4
Amélioration	17	18
Total	96	100

Tableau 3 Pourcentage moyen de chaque type de contenu — UAM Corpus Tools

Le Tableau 3 ci-dessus présente le pourcentage moyen de chaque type de contenu. Nous observons que le contenu principal (appelé Main dans notre schéma d'annotation, en référence à la balise HTML qui le définit) ne représente que 55 % du texte. Le contenu principal ne correspond donc qu'à 21 % de la page en moyenne, ce qui signifie qu'il y a 79 % de contenu annexe, dédié aux autres catégories, ou fonctions. Remis en contexte, ces résultats nous permettront ensuite d'étudier la répartition des fonctions sur la page, mais aussi de caractériser le discours de l'accès aux droits numérique dans son environnement discursif.

Le deuxième objectif de la méthode mise au point est d'intégrer la dimension interactive à notre analyse. Cela a été fait dans un premier temps dans la collecte du parcours utilisateur. Cependant, nous souhaitons également étudier l'interactivité d'une page web. Les liens ont donc été annotés pour l'ensemble du parcours. Le Tableau 4 ci-dessous en présente les résultats.

Liens	Occurrences totales	Pourcentage

Internes	317	30,4
Externes	70	6,7
Total	387	37,1

Tableau 4 Pourcentage de liens dans le parcours — UAM Corpus Tools

Dans le Tableau 4, nous pouvons observer qu'en tout 37 % du texte annoté est un hyperlien. Cela signifie qu'en moyenne 37 % de la page web est cliquable. Il s'agit d'un pourcentage conséquent, surtout lorsque le contenu principal ne représente que 21 % du texte présent sur la page. Ces statistiques indiquent également que la plupart des liens redirigent l'utilisateur sur une page interne au site, ou même à la page elle-même.

Les résultats statistiques préliminaires générés par le logiciel UAM Corpus Tool nous permettent d'obtenir des résultats préliminaires concernant notre parcours utilisateur pilote. Ces résultats permettent d'évaluer la pertinence de cette analyse, mais aussi d'envisager des pistes d'analyses futures afin de mettre en place cette méthode d'analyse.

5.2. Premières interprétations

Cette étude pilote, ainsi que ces résultats permettent de faire des premières interprétations. Tout d'abord, les positions de Marie-Anne Paveau et Dominique Maingueneau, qui concernent initialement différents types de sites ou de plateformes numériques, semblent se vérifier dans notre corpus également. Le discours est délinéarisé, car entouré de textes divers, organisés en différentes rubriques (ou modules) et répondant à différents objectifs (naviguer, informer, etc.). Le fil du discours, si l'on considère qu'il s'agit du contenu principal, est non-linéaire, car entouré de discours complémentaires, mais aussi déterminés par des clics.

Le concept d'affordance semble également être une piste à explorer afin de rendre compte de l'interactivité et de la personnalisation du discours. En effet, chaque utilisateur a une expérience différente du site, car il ne visite pas les mêmes pages. Les potentialités du site sont conditionnées par l'ergonomie du site, mais aussi par les capacités de l'utilisateur. Nous

pourrions donc envisager différents profils d'utilisateur afin de proposer différents types de parcours.

Grâce aux statistiques obtenues, nous pouvons observer que le contenu principal ne concerne pas la majorité du discours présent sur la page dans notre corpus pilote, comme l'illustre le graphique ci-dessous.

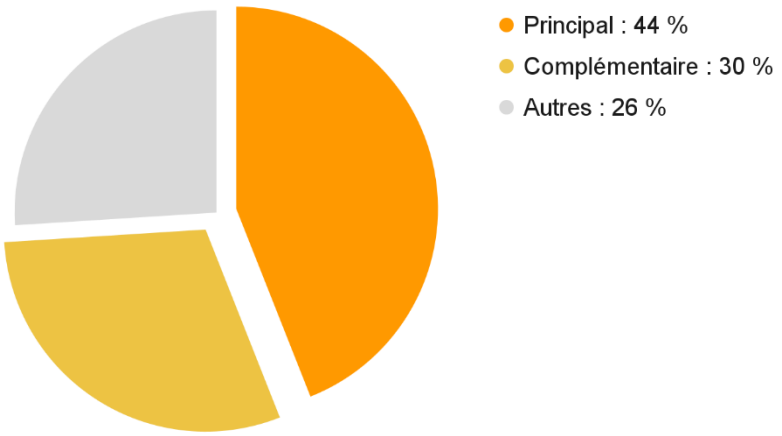


Figure 8 Représentation graphique des différents types de contenu

Sur le graphique ci-dessous réalisé à partir des statistiques obtenues sur le logiciel UAM Corpus Tool, le contenu principal ne constitue par la majorité du discours présent sur la page web. Cependant, la ressource recherchée par l'usager ne concerne souvent que le contenu principal. Le reste du discours peut donc être considéré comme du bruit pour l'usager à la recherche d'information. Les fonctions du discours sont diverses et complémentaires. La navigation permet de trouver le contenu recherché. Elle est donc partie intégrante du discours, et ne peut être ignorée. Il est important de ne pas étudier l'ensemble de la page web de façon linéaire, sans apporter de précision sur les objectifs et caractéristiques de chaque partie du discours.

5.3. Résultats terminologiques et phraséologiques préliminaires

Le schéma d'annotation permet d'isoler les segments ne concernant que la navigation, le contenu ou les liens par exemple. En étudiant ces différents types de segments, il est possible d'identifier des caractéristiques pour chaque type de segments. Ces analyses n'ont pas encore été généralisées sur l'ensemble du parcours, mais nous présentons ici quelques observations issues de ce modèle pilote.

Dans les segments correspondant au contenu, nous avons observé un grand nombre de phrases conditionnelles. Nous avons identifié le schéma « [*Si vous êtes] [situation administrative]* » à l'aide du corpus présenté dans la partie 3.1 sur le logiciel Sketch Engine. Dans notre corpus pilote, ce schéma apparaît uniquement dans les segments annotés comme

étant du contenu. Il conviendra d'identifier si d'autres types de segments indiquent une situation administrative, afin de voir si ce schéma se manifeste d'une façon différente suivant les fonctions de la page. Par exemple, il existe des simulateurs qui permettent aux usagers de définir leur profil administratif pour obtenir des informations personnalisées. Nous pouvons faire l'hypothèse que ce schéma se trouverait une forme modifiée dans ce simulateur.

Dans les segments correspondants à la navigation, nous avons identifié une grande densité lexicale. Le logiciel UAM Corpus Tool indique 99 % de lexème dans les segments étiquetés comme étant de la navigation. En effet, ces unités font majoritairement partie des menus. Les menus sont composés en majorité de termes décrivant le thème des informations se trouvant sur la page. Le segment suivant, qui est présent 7 fois dans les 5 pages de notre corpus, illustre ce phénomène : « • copie intégrale • extrait avec filiation • extrait sans filiation ». Il s'agit d'un sous-menu thématique, qui a pour but d'aider l'usager en l'orientant vers les pages qui pourraient l'intéresser. D'un point de vue terminologique, nous pouvons remarquer qu'il s'agit de 3 termes, qui sont des hyponymes d'Acte de naissance, qui est le thème de notre parcours utilisateur. La terminologie pourrait donc avoir une place importante dans l'arborescence du site.

Ces pistes d'analyses phraséologiques et terminologiques ont été réalisées à partir de l'étude pilote et devront être approfondies. Elles montrent que l'application de ce modèle à d'autres parcours utilisateur pourrait être pertinente.

5.4. Deux méthodes complémentaires

Le modèle présenté dans cet article met en lumière certaines parties du discours qui sont souvent effacées : menu, éléments interactifs, hyperliens, etc. Le modèle nous semble donc pertinent afin de caractériser un discours issu de sites web dynamiques, et sera étendu à d'autres parcours utilisateurs. Cependant, il ne saurait être suffisant. Afin d'apporter un éclairage complet sur le discours, nous ne pouvons faire l'économie d'un corpus classique. Ce dernier est utile pour étudier la terminologie et la phraséologie du domaine, mais il présente surtout l'information recherchée par l'utilisateur. Il est donc essentiel pour étudier les termes et leurs définitions, mais aussi la façon dont les informations sont classées et mises à disposition. La mise en place de cette annotation pilote ne vise pas à nier l'importance d'une analyse de corpus classique, qui reste essentielle. Au contraire, elle cherche à compléter ces analyses et à les mettre en contexte afin de proposer une caractérisation complète du discours étudié. Le modèle sera enrichi dans de futurs travaux par une étude des schémas phraséologiques selon leur fonction et leurs caractéristiques sur la page, notamment en prenant en compte les particularités du site web

dynamique : la délinéarisation du discours, la fonctionnalité de la page web et la construction dynamique du fil discursif.

BIBLIOGRAPHIE

Sources primaires

Beauvisage, Thomas. (2004). *Sémantique des parcours des utilisateurs sur le Web* [Université Paris X]. http://www.revue-texto.net/1996-2007/Inedits/Beauvisage/Beauvisage_Parcours.html

Bhatia, V. K. (2014). Analysing genre: Language use in professional settings. Routledge.

Biber, Douglas. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing*, *8*(4), 243–257. https://doi.org/10.1093/llc/8.4.243

Burger, Marcel. (2018). Entre affordances et multimodalité : de nouveaux enjeux pour l'analyse des discours du digital. *Cahiers Du Centre de Linguistique et Des Sciences Du Langage*, *55*(55), 3–24. https://doi.org/10.26034/la.cdclsl.2018.269

Cacchiani, Silvia. (2018). Webpage usability and utility content: Citizens' rights and the law on Gov.uk. *IPERSTORIA*, *12*(Fall-Winter 2018), 192–205. https://iris.unimore.it/handle/11380/1200602

Emerit, Laetitia. (2016). La notion de lieu de corpus : un nouvel outil pour l'étude des terrains numériques en linguistique. *Corela. Cognition, Représentation, Langage*, *14*(1). https://journals.openedition.org/corela/4594

Foucault, Michel. (1971). The Archaeology of Knowledge and the Discourse on Language. In *Pantheon, New York*. Pantheon Book.

Ghliss, Yosra. (2019). Les photo-discours WhatsApp : éléments d'analyse d'une affordance d'une application mobile. *Corela. Cognition, Représentation, Langage*. https://journals.openedition.org/corela/8480

Gibson, James J. (1977). *The Theory of Affordances*. Hilldale. https://doi.org/10.1075/lsse.2.03bli

Habert, Benoît. (2000). Des corpus représentatifs : de quoi, pour quoi, comment. *Cahiers de l'Université de Perpignan*.

Kilgarriff, Adam, Baisa, Vít, Bušta, Jan, Jakubícek, Miloš, Kovár, Vojtech, Michelfeit, Jan, Rychlý, Pavel, & Suchomel, Vít. (2014). The Sketch Engine: Ten Years on. *Lexicography*, *1*(1), 7–36. https://doi.org/10.1007/s40607-014-0009-9

Maingueneau, Dominique. (2016). L'ethos discursif et le défi du Web. *Itinéraires*, *2015–3*, 2015–2018. https://doi.org/10.4000/itineraires.3000

Mayeur, Ingrid, & Paveau, Marie-Anne. (2020). Présentation. Les devenirs du texte numérique natif. *Corela*, *HS-33*. https://doi.org/10.4000/corela.11749

McEnery, Tony, & Wilson, Andrew. (2001). *Corpus Linguistics* (Second Edi). Edinburgh University Press.

Moreno, A. I., & Swales, J. M. (2018). Strengthening move analysis methodology towards bridging the function-form gap. *English for Specific Purposes*, *50*, 40-63.

O'Donnell, Michael. (2009). The UAM CorpusTool: software for corpus annotation and exploration. *Proceedings of the XXVI Congreso de AESLA*, *3*. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.159.7393&rep=rep1&type=pdf

Ostern, Nadine, & Rosemann, Michael. (2021). A Framework for Digital Affordance. *Twenty-Ninth European Conference on Information Systems* , *1*.

Paveau, Marie-Anne. (2013). Genre de discours et technologie discursive. Tweet, twittécriture et twittérature. *Pratiques. Linguistique, Littérature, Didactique*, 7–30. https://journals.openedition.org/pratiques/3533

Paveau, Marie-Anne. (2015). Ce qui s'écrit dans les univers numériques. *Itinéraires*, *2014–1*. https://doi.org/10.4000/itineraires.2313

Paveau, Marie-Anne. (2017). *L'analyse du discours numérique. Dictionnaire des formes et des pratiques*. https://journals.openedition.org/lectures/24355

Pecman, Mojca. (2018). *Langue et construction des connaissances : énergie lexico-discursive et potentiel sémiotique des sciences* (Issue 65). L'Harmattan. https://doi.org/10.4000/lidil.10508

Rundell, Michael. (2008). The Corpus Revolution Revisited. *English Today*, *24*(1), 23–27. https://doi.org/10.1017/S0266078408000060

Rychlý, Pavel. (2008). A Lexicographer-Friendly Association Score. *RASLAN*, 6–9.

Sinclair, John. (1991). *Corpus, Concordance, Collocation*. Oxford University Press.

Stanfill, Mel. (2015). The interface as discourse: The production of norms through web design. *New Media and Society*, *17*(7), 1059–1074. https://doi.org/10.1177/1461444814520873

Swales, J. M. (2004). Research genres: Explorations and applications. Cambridge University Press.

Vitali-Rosati, Marcello. (2014). *Égarements. Amour, mort et identités numériques* (ermann). Hermann. www.editions-hermann.fr

La préparation des données pour une typologie et une ontologie de l'expression du déplacement

Karl Seifen

Dynamique du Langage, Université Lumière – Lyon 2

karl.seifen@univ-lyon2.fr

RESUME

Cet article a pour but de présenter la méthodologie utilisée dans le cadre d'une étude en cours sur l'expression du déplacement. Nous montrons comment des données de langues diverses typologiquement et géographiquement peuvent être formatées et codées pour créer une ontologie (un réseau de concepts). Le codage des données se fait à deux niveaux : le premier est morphosyntaxique (propriétés syntaxiques et morphologiques des formes linguistiques) et le second sémantique (catégories conceptuelles relevant du domaine spatial). Nous illustrons notre méthodologie par une recherche sur le type d'unités encodant le concept sémantique de Trajectoire dans nos données et montrons que différentes langues encodent la Trajectoire dans la Tête de la phrase (prédicat), d'autres dans un Relateur (cas ou adposition), ou encore dans de multiples *loci* : Tête et Relateur, Tête et Satellite (particule ou affixe verbal), ou Relateur et Satellite.

***Mots-clés** : Typologie de l'espace – Trajectoire – Ontologie – Déplacement*

ABSTRACT

This paper aims to show the methodology employed in an on-going study of motion expression. We show how data from typologically and geographically diverse languages can be formatted and annotated to create an ontology (a network of concepts). Data are annotated at two different levels: (i) morpho-syntactic (syntactic and morphological properties of the linguistic units) and (ii) semantic (conceptual categories of the spatial domain). We illustrate our methodology with a query on the types of units encoding the semantic concept of Path in our data and show that language families encode Path in the Head of the sentence (predicate), in a Relator (case or

adposition) or in multiple loci: Head and Relator, Head and Satellite (particle or verbal affix), or Relator and Satellite.

Keywords *: Typology of space –Path – Ontology – Motion*

Notre étude porte sur l'expression du déplacement dans le cadre du projet SpOTy (Spatial Ontology and Typology, Labex ASLAN, Université de Lyon)[5]. Ce projet vise à créer d'une part une typologie de l'expression de la Trajectoire dans les langues du monde, et d'autre part une ontologie des relations spatiales. Cet article a pour but de présenter notre méthodologie. Plus précisément, nous montrons comment le formatage et le codage des données nous permet de créer un réseau de concept interconnecté (*i.e*. une ontologie) que nous pouvons interroger.

Dans un premier temps, nous définissons les concepts d'ontologie (section 1.) et d'évènement de déplacement (section 2.). Nous présentons ensuite la méthodologie employée pour mettre en forme et coder nos données (section 3.). Enfin, nous formulerons une typologie préliminaire de la Trajectoire à partir de notre base de données (section 4.).

1. ONTOLOGIE

Dans le domaine de l'intelligence artificielle et de la représentation des connaissances, une ontologie est un réseau interconnecté de concepts et de relations entre ces concepts, explicitant et organisant un domaine spécifique (Guarino *et al*. 2009: 2, Schalley *et al*. 2014: 141). Gruber (1993: 199) définit une ontologie comme la « spécification explicite d'une conceptualisation », *i.e*. une certaine vision du monde ou d'un domaine d'intérêt, dont les concepts sont explicitement donnés (Sowa 2003).

> « A body of formally represented knowledge is based on a conceptualization: the objects, concepts, and other entities that are assumed to exist in some area of interest and the relationships that hold among them. A conceptualization is an abstract, simplified view of the world that we wish to represent for some purpose. Every knowledge base, knowledge-based system, or knowledge-level agent is committed to

[5] L'auteur remercie le LABEX ASLAN (ANR-10-LABX-0081) de l'Université de Lyon pour son soutien financier dans le cadre du programme français "Investissements d'Avenir" géré par l'Agence Nationale de la Recherche (ANR).

some conceptualization, explicitly or implicitly. » (Genesereth & Nilsson 1987, cités par Guarino *et al*. 2009: 3)

En linguistique, et plus particulièrement en typologie, les ontologies facilitent la gestion et l'analyse de données translinguistiques, qui peuvent être complexes, éparses, ou incomplètes (Schalley 2019 : 11). Les ontologies peuvent également servir à combler le manque d'interopérabilité entre différents projets et bases de données (Farrar & Langendoen 2003: 100).

Concrètement, les ontologies sont des réseaux où différents nœuds (concepts) sont reliés entre eux par des relations spécifiques. La figure 1 schématise le concept d'ontologie tel qu'il peut être utilisé en linguistique, avec comme exemple les clitiques et affixes pronominaux du français. On y retrouve les concepts de l'ontologie (encadrés), divisé en trois types : (i) les formes (*i.e*. signifiants saussuriens : 'je', 'tu', 'nous', '-ons', '-ez'), (ii) les sens (*i.e*. signifiés : singulier, pluriel, 1ère personne, 2ème personne), et (iii) les catégories morphosyntaxiques (clitiques et suffixes). Ces concepts sont reliés par des relations spécifiques (en italique) : les formes et leurs sens par la relation *a le sens de*, ainsi que les formes et leurs catégories par la relation *appartient à la catégorie morphosyntaxique*. Les concepts de même type peuvent être reliés par la relation hyperonymique *type de* : un suffixe est un type d'affixe, le singulier et le pluriel sont des instances de nombre grammatical, et les première et deuxième personnes sont des instances de personne grammaticale.

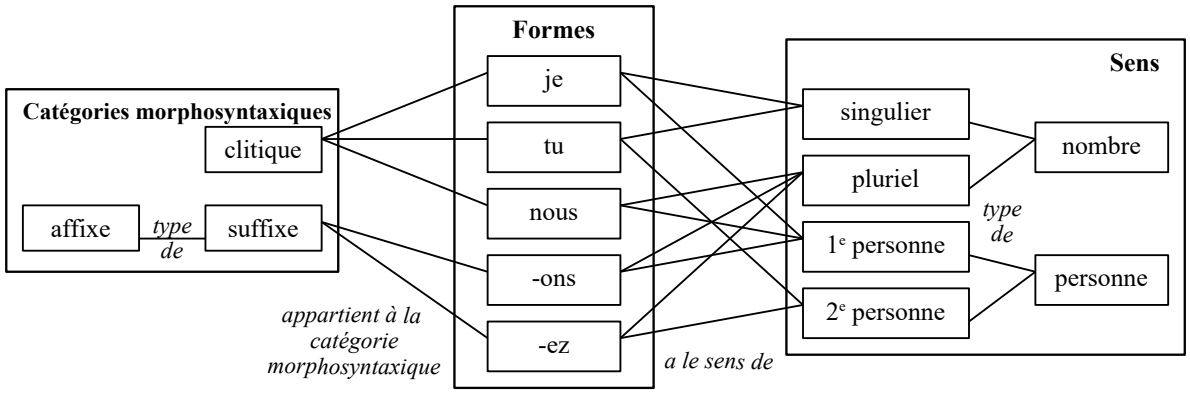


Figure 9 : Ontologie des formes pronominales du français

2 EXPRESSION DU DEPLACEMENT

Notre étude porte sur l'expression du déplacement et les évènements dits translocatifs (aussi appelés : cislocatifs, *motion events*, *translatory situations*, *displacement*). Ce type d'évènement

exprime un changement de localisation d'une entité. Talmy (1972: 1) définit les évènements translocatifs comme le déplacement ou la localisation d'un objet par rapport à un autre objet[6].

Outre une définition des évènements translocatifs, Talmy (1972, 1985, 2000) propose également une décomposition de ces évènements en constituants sémantiques (section 2.1), ainsi qu'une typologie des langues selon leur encodage du déplacement (section 2.2).

2.1 Constituants des évènements translocatifs

Talmy (1972, 1985) propose d'analyser ces événements translocatifs comme contenant quatre constituants principaux : (i) la Figure, *i.e* l'entité se déplaçant, (ii) le Site (*Ground*), *i.e*. l'entité par rapport à laquelle la Figure se déplace, (iii) le Moteur (*Fact-of-Motion*), *i.e.* l'état de la Figure, pouvant avoir les valeurs dynamique (déplacement) ou statique (localisation), et (iv) la Trajectoire (*Path*), *i.e*. la relation spatiale entre la Figure et le Fond. Outre ces quatre constituants, les évènements translocatifs peuvent également incorporer des co-évènements comme la Manière ou la Cause du mouvement (exemple 1).

1. The man ran into the room.
   Figure Manière/Moteur Trajectoire Site

Toujours selon Talmy (2000: 53-56), le composant de Trajectoire est lui-même complexe et se décompose en trois sous-éléments principaux :

1) le Vecteur qui correspond à la position du Site sur le parcours de la Figure (point initial, point médian, étendue médiane, point final).
2) la Conformation qui indique la relation topologique entre la Figure et le Site (dans, sur, *etc*.)[7].
3) la Deixis qui précise la direction du déplacement par rapport à un centre déictique (s'en approchant ou s'en éloignant). Son statut de sous-composant de la Trajectoire, plutôt que de composant constitutif des évènements translocatifs, est cependant discuté (Morita 2011: §72, Matsumoto *et al*. 2017: 121, Seifen *et al*. 2021).

Outre ces trois sous-composants, Slobin et Hoiting (1994: 498) en se basant sur les travaux de Aske (1989), montrent l'importance du composant sémantique du Franchissement de frontière (ex. « entrer », « sortir », et « traverser » par opposition à « atteindre », « quitter », et « passer »).

[6] « One object moves or is located with respect to another object » (Talmy 1972:1).

[7] Vandeloise (1986: 22-30) montre que les prépositions spatiales du français telles que « dans » ont des caractéristiques fonctionnelles plutôt que topologiques.

2.2 Typologie(s) de la Trajectoire

A partir de sa description des évènements translocatifs, Talmy (1985) propose une typologie des langues, selon le *locus* du composant Trajectoire, c'est-à-dire selon l'unité morphosyntaxique encodant la Trajectoire. Deux types de langues sont ainsi décrits : les langues à cadrage verbal (*verb-framed languages*) et les langues à cadrage satellitaire (*satellite-framed languages*). Les langues à cadrage verbal encodent la Trajectoire dans le verbe (langues romanes, sémitiques, polynésiennes). Ainsi dans l'exemple 2a, la Trajectoire est encodée dans le verbe *entrer*. Les langues à cadrage satellitaire (langues germaniques, slaves, sinitiques) encodent quant à elles la Trajectoire dans un satellite. Celui-ci peut être défini comme un constituant de surface, autre qu'un syntagme nominal ou prépositionnel, dépendant du verbe (Talmy 1985: 102), incluant les particules verbales, les affixes verbaux, ou les noms incorporés[8]. L'exemple 2b illustre ce deuxième type : la particule verbale *in* porte l'information de Trajectoire.

2. Cadrages verbal et satellitaire :

   a. L'enfant **entra**. (cadrage verbal)
   b. The child walked **in**. (cadrage satellitaire)

Cette typologie a fait l'objet de nombreuses révisions. Wälchli (2001: 301) propose trois types de langues : verbale (la Trajectoire est encodée dans un verbe), adverbale (la Trajectoire est encodée dans une particule ou un affixe verbal) et adnominale (la Trajectoire est encodée dans un cas ou une adposition)[9]. Matsumoto (2003: 410) propose de remplacer les notions de verbe et de satellite par celle de tête et de non-tête (*head* et *non-head*), d'une part parce que les verbes non-finis (gérondifs, connectifs, participes) ont un comportement semblable aux satellites, et d'autre part parce que les adpositions et les cas ne sont pas des satellites.

> « Talmy's typology of verb vs. satellite-framed languages suffers from the misleading use of the term "verb". What is meant by the term verb is in fact the head of a clause. Satellites can also be a verb […]. For this reason, a better name for verb-framed languages is head-framed languages. Satellite-framed languages, on the other hand, can be termed as nonhead-framed languages. Note that satellites and nonheads are slightly

[8] Bien que la définition que donne Talmy du satellite exclue les adpositions, ces dernières sont paradoxalement inclues dans le cadrage satellitaire (voir Imbert *et al*. (2011) sur la problématique de la notion de satellite dans les travaux de Talmy).

[9] « According to the locus of highest differentiation of displacement one may distinguish three gross types for the encoding of displacement: **(V) verbal encoding** (i.e. by the verb stem), **(AN) adnominal encoding** (i.e. by prepositions, postpositions or case marking), and **(AV) adverbal encoding** (i.e. by verb affixes or verb particles) » (Wälchli 2001: 301).

different notions: all satellites are nonheads by definition […] but not all nonheads are satellites. » (Matsumoto 2003: 410)

Enfin, Fortis et Vittrant (2011, 2016)[10] proposent une typologie des constructions, plutôt que des langues (les langues possédant généralement plusieurs types de constructions). Celle-ci est basée sur les différents *loci* possibles de la Trajectoire, à savoir :

1) la Tête (*i.e.* le prédicat, qu'il soit verbal ou non comme dans l'exemple 3a),
2) le Satellite (particule, affixe verbal, adverbial),
3) le Relateur (cas ou adposition),
4) l'Argument (généralement un nom déverbal, comme en 3b),
5) la construction elle-même ; la Trajectoire n'est alors pas portée par un marqueur mais par la construction elle-même, comme dans l'exemple 3c. Dans cet énoncé, le sens de « traverser » est véhiculé par la construction plutôt que par un morphème spécifique),
6) plusieurs *loci* possibles (combinaison de plusieurs stratégies, comme dans l'exemple 3d).

« Nous avons choisi de répartir les éléments susceptibles d'exprimer la trajectoire en quatre catégories. Ces quatre catégories sont : la tête de la phrase (notée T), le satellite de la tête […] ou du nom, noté S, les adnominaux (A), terme qui subsume les cas et adpositions, et enfin les noms (N). Le cadrage multiple montre que ces loci d'encodage sont cumulables dans bon nombre de langues. » (Fortis & Vittrant 2011: 83)

3. a. Indonésien (Malayo-polynésien, austronésien, Indonésie) : Tête non-verbale

Məreka **kə** bioskop
3PL **to** movies
'They (are going) to the movies.' (Fortis & Vittrant 2016: 358)

b. Français (Roman, indo-européen, France) : Argument et Relateur

L'**acheminement jusqu'à** l'aéroport se fait en navette. (Fortis & Vittrant 2016: 355)

c. Japonais (Japonique, Japon) : Construction et Tête verbale

Taroo-ga Omekaidoo-o hashit-te it-ta.
Taro-NOM Ome_road-ACC run-CVB go-PST
'Taro ran across the road.'(lit. Taro went the road running) (Morita 2009: 237, cité par Fortis & Vittrant 2016: 368)

[10] Dans le cadre du projet de recherche Typologie de la Trajectoire (2006-2011, Fédération Typologie et Universaux du Langage).

d. Japonais (Japonique, Japon) : Tête, Satellite et Relateur

onnanohito-ga dookutu-no naka-**kara** **de**-te **iki**-masi-ta
woman-NOM cave-GEN inside-**ABL** **exit**-CVB **go**-POL-PST
'The woman exited from the inside of the cave away from me.' (Fortis & Vittrant 2016: 367)

3. DONNEES

Notre étude vise à produire une typologie de la Trajectoire dans les langues du monde. Nous avons donc compilé des données décrivant des évènements translocatifs. Notre corpus actuel est constitué d'énoncés de 57 langues appartenant à 20 familles différentes. Les données sont issues de deux sources principales : (i) la littérature scientifique traitant de l'espace, et plus particulièrement de l'espace dynamique (déplacement), et (ii) les corpus personnels des membres du projet 'Typologie de la Trajectoire' (2006-2011, Fédération Typologie et Universaux du Langage), ainsi que des chercheurs appartenant aux laboratoires associés.

Les données collectées au sein du projet Typologie de la Trajectoire l'ont été à l'aide du stimulus Trajectoire (Ishibashi *et al*. 2006, Vuillermet & Kopecka 2019)[11]. Ce stimulus a pour but d'éliciter des descriptions de déplacement dans des langues typologiquement différentes, en demandant aux informateurs de décrire différentes scènes. L'utilisation du stimulus Trajectoire permet d'avoir des données comparables entre différentes langues. Il contient 76 clips vidéo d'une dizaine de secondes (2 entraînements, 9 distracteurs, 65 trajectoires). Les exemples 4a (mandarin), 4b (ese ejja), et 4c (polonais) décrivent tous la même situation : une femme sort d'un champ de maïs (figure 2).

[11] Disponible en ligne : http://tulquest.huma-num.fr/en/node/132.

Figure 10 : Clip 38 du stimulus Trajectoire (038_Path_F_walk_outof_field_sideRL)

4. a. Chinois mandarin (Sinitique, sino-tibétain ; Chine)

yí	wèi	nǚshì	zǒu=chu	yùmǐ	dì
un	CLF	femme	marcher=OUT	maïs	champ

‘Une femme sort du champ de maïs.’ (Jin-Ke Song, corpus personnel)

b. Ese Ejja (Tacana ; Bolovie)

epona	kwaya'yoani	shixeiye=xe	eshaja mexi
femme	sort	champ.de.maïs=PERL	avec un panier

‘Une femme sort du champ de maïs avec un panier.’ (Vuillermet 2012: 574)

c. Polonais (Slave, indo-européen ; Pologne)

kobieta	wy-chodzi	z	pola	kukurydzy
femme	OUT-marche	de	champ.GEN	maïs.GEN

‘Une femme sort du champ de maïs en marchant.’ (Fagard & Kopecka 2021: 161)

3.1 Formatage

Les données collectées auprès de différentes sources dans le cadre de notre étude doivent être formatées, afin de permettre leur intégration dans une base de données (dont la création est un de nos objectifs). Ce formatage suit deux contraintes : d’une part, il doit être lisible par la base de données qui en récupère les informations, et d’autre part, il doit être suffisamment intuitif pour permettre, à terme, aux chercheurs de formater et d’intégrer eux-mêmes leurs données dans la base.

Nous avons donc opté pour un format de type tableur balisé, où les données et les métadonnées sont divisées en ligne, chacune introduite par des balises (*i.e.* mots-clés). Les balises et les informations correspondantes sont données dans le tableau 1.

Balise	**Information**
ID	Identifiant unique de l'énoncé
LGE	Langue de l'énoncé (suivant les conventions ISO 639-3)
TRANS	Transcription de l'énoncé
ORTHO	Enoncé en orthographe native (si applicable)
TRAD	Traduction de l'énoncé en anglais
TRADFR	Traduction de l'énoncé en français (si disponible)
TRAJ	Numéro du stimulus Trajectoire associé à l'énoncé (si applicable)
SOURCE	Référence bibliographique ou corpus personnel
TOKEN	Transcription tokénisée de l'énoncé
GLOSE	Ligne de glose interlinéaire
CLAUSE	Division de l'énoncé en propositions
MS	Codage morphosyntaxique (cf. infra)
CC	Codage des catégories conceptuelles (cf. infra)

Tableau 1: Balises et informations correspondantes

Les énoncés sont associés à un identifiant (ID), à une langue au format ISO (LGE), à une référence bibliographique ou un corpus personnel (SOURCE) et éventuellement à une scène issue du stimulus Trajectoire (TRAJ). Chaque énoncé contient également une transcription (TRANS), une traduction en anglais (TRAD), voire dans d'autres langues de travail (français : TRADFR, espagnol : TRADES), ainsi qu'une version en orthographe de la langue, si la langue est écrite (ORTHO). Les données elles-mêmes contiennent une transcription tokénisée en morphèmes (TOKEN), accompagnée pour chaque morphème d'une glose interlinéaire (GLOSE), du numéro de la proposition à laquelle ils appartiennent (CLAUSE), et d'informations morphosyntaxiques (MS) et sémantiques (CC) lorsque le morphème relève du domaine spatial. Ce formatage est illustré en figure 3, avec un exemple de thaï central (taï-kadaï, Thaïlande).

ID	tha_50					
LGE	THA					
TRANS	pʰûːjĭŋ dɤːn kʰûn paj bon nɤːn					
ORTHO	ผู้หญิงเดินขึ้นไปบนเนิน					
TRADFR	La femme monte sur le monticule [en marchant en s'éloignant du CD].					
TRAD	The woman walks up the knoll [away from DC].					
TRAJ	71					
SOURCE	Seifen p.c.					
TOKEN	pʰûːjĭŋ	dɤːn	kʰûn	paj	bon	nɤːn
GLOSE	woman	walk	ascend	go	top	knoll
CLAUSE	1	1	1	1	1	1
MS	An	Hv	Hv	Hg	Rn	An
CC	Fh	Mm	Pvu	Ddf	Cto	Gg

Figure 11 : Exemple de formatage des données

3.2 Codage

Chaque morphème relevant de la spatialité reçoit deux codes. Le premier rend compte de ses propriétés morphosyntaxiques (ligne MS) et le second de ses propriétés sémantiques (ligne CC, catégorie conceptuelle). Le codage de ces informations a plusieurs avantages. Tout d'abord, il permet de normaliser la notation des éléments spatiaux issus de sources diverses. Par exemple, le sens de « déplacement à l'intérieur de » pouvant être glosé par *enter*, *go_in*, *in*, ou *into* (sans compter les gloses non-anglaises) est rendu par « Pbi » (Trajectoire avec franchissement de frontière vers l'intérieur du Site, en anglais : ***P**ath with **b**oundary-crossing **i**nto a Ground*) dans notre codage, quelle que soit la glose utilisée par l'auteur d'origine.

D'autre part, le codage utilisé participe à une ontologie, où les formes linguistiques sont ainsi liées à des catégories de deux types : (i) catégorie sémantique avec une relation de type *a le sens de X* et (ii) catégorie morphosyntaxique avec une relation de type *a la propriété morphosyntaxique X*. Le codage choisi note également des relations hyperonymiques (concepts emboîtés). Ainsi, le code « Pbi » ('déplacement à l'**i**ntérieur de') est un sous-type de « Pb » (déplacement avec franchissement de frontière : ***b**oundary*), lui-même sous-type de « P » (Trajectoire : ***P**ath*).

Les codages morphosyntaxiques ont donc la forme « Xy(z) ». La première lettre du codage (« X ») indique la fonction syntaxique de l'unité : Tête (prédicat), Satellite (particules et affixes verbaux), Relateur (cas et adpositions), Argument, ou Satellite de l'argument (modifieur du nom). 'H' indique la Tête, 'S' le Satellite, 'R' un Relateur et 'A' un Argument et 'Z' un Satellite de l'argument. La deuxième lettre (« y ») indique la catégorie morpho-syntaxique : verbe fini (v), verbe non-fini (f), nom (n), pronom (p), affixe ou clitique (x),

adposition (a), *etc*. Enfin, la troisième lettre (« z ») est optionnelle et marque une catégorie grammaticale pertinente pour notre travail comme la diathèse (e) ou mouvement associé (m)[12,13]. Les trois éléments du code morphosyntaxique « X », « y », et « z » sont indépendants. Il n'existe aucune limite combinatoire pour l'ontologie (mais en réalité certaines combinaisons semblent improbables).

Le deuxième codage (CC : catégories conceptuelles) est sémantique et de type « Abc », où « c » est un type de « Ab », lui-même type de « A ». Par exemple, la catégorie « D » (Deixis) possède les sous-types « Ds » (Deixis statique : *ici*, *là-bas*) et « Dd » (Deixis dynamique : déplacement par rapport au centre déictique), lui-même possédant les sous-types « Ddf » (centrifuge : s'éloignant du centre déictique *en allant*), « Ddp » (centripète : s'approchant du centre déictique *en venant*), et « Ddt » (transverse : sans s'approcher ni s'éloigner du centre déictique, *en passant*), comme montré en tableau 2. Le codage des catégories conceptuelles concerne l'ensemble des constituants sémantiques présentés en 2.1. (Figure, *Fact-of-Motion*, Site, Manière, Cause, Trajectoire et ses sous-composants : Vecteur, Conformation, Deixis).

<table>
<tr><th>Niveau 1</th><th>Niveau 2</th><th>Niveau 3</th></tr>
<tr><td rowspan="4">« D » :
Deixis</td><td rowspan="3">« Dd » :
Deixis dynamique</td><td>« Ddf » : centrifuge</td></tr>
<tr><td>« Ddp » : centripète</td></tr>
<tr><td>« Ddt » : transverse</td></tr>
<tr><td colspan="2">« Ds » : Deixis statique</td></tr>
</table>

Tableau 2 : Codage sémantique de la Deixis

La figure 4 exemplifie le codage des propriétés morphosyntaxiques (en bleu) et sémantiques (en vert) des morphèmes de la phrase *La femme sort de la grotte*. Ici, *femme* est un Argument nominal (codage morphosyntaxique : « An ») désignant une Figure humaine (codage sémantique : « Fh »). Le verbe *sort* est la Tête de la phrase et un verbe fini (« Hv ») ; il décrit une Trajectoire avec franchissement de frontière vers l'extérieur (« Pbo »). La préposition *de* est un Relateur de type adposition (« Ra ») ; elle indique un déplacement depuis un point initial,

[12] Le mouvement associé (Guillaume & Koch 2021: 3) indique que l'action exprimée par le verbe est associée à un déplacement, qui peut la précéder, la suivre, ou lui être concomitant.

[13] Un tel marquage permet de retrouver et d'interroger les morphèmes concernés plus facilement. Ces deux catégories nous ont semblé, lors de la conception du codage, comme particulièrement intéressantes à faire ressortir ; le mouvement associé étant un sujet d'actualité, et la diathèse étant relativement peu décrite dans les études sur la spatialité.

une source (« Ps »). Enfin, *grotte* est un Argument nominal (« An ») désignant l'entité servant de référence au déplacement, ici le point initial (« Gs ») de ce déplacement.

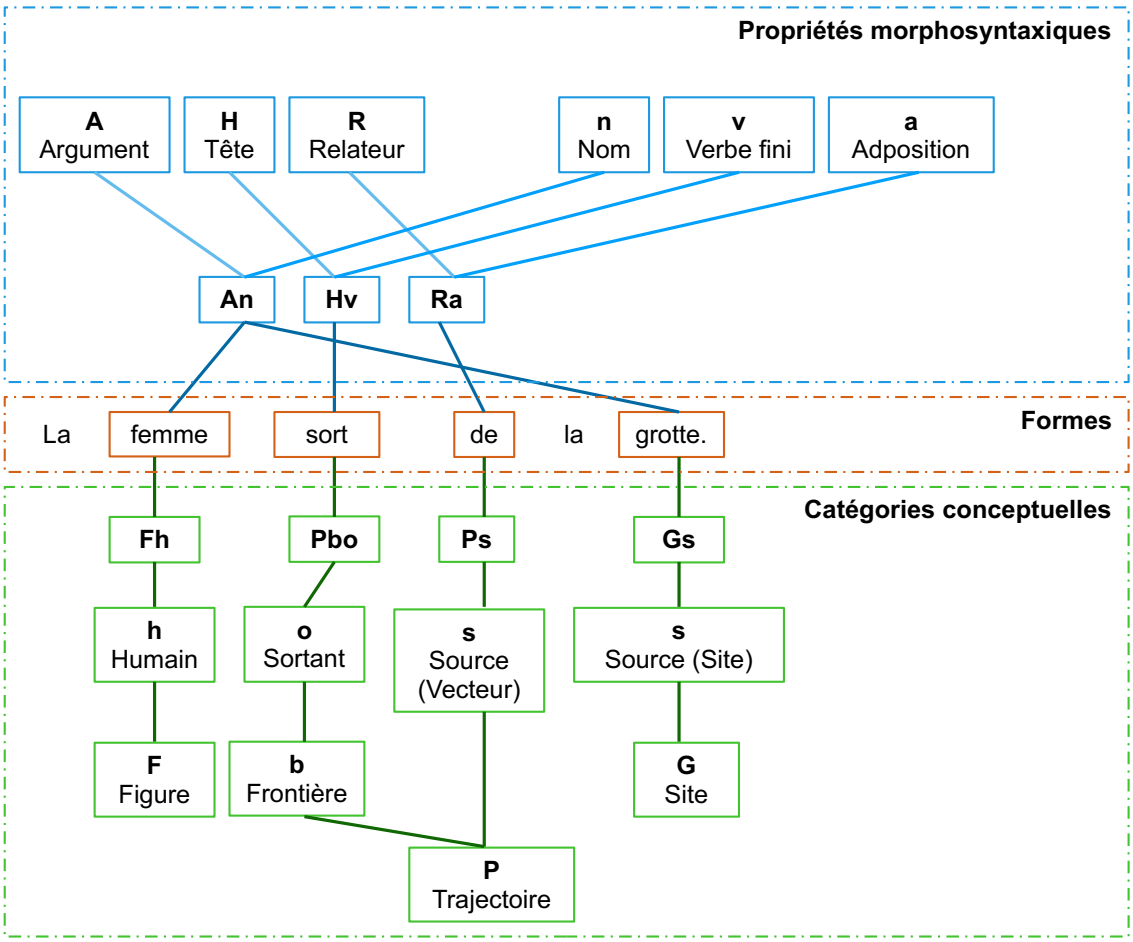


Figure 12 : Exemple de codage morphosyntaxique et sémantique

4. ANALYSE PRELIMINAIRE

Le codage des données permet leur interrogation sous plusieurs angles. Il est par exemple possible de rechercher les correspondances entre catégories conceptuelles et catégories morphosyntaxiques, *i.e.* comment est encodé un concept spécifique, ou à l'inverse, qu'encode une forme particulière. Dans cette section, nous explorons l'encodage du composant sémantique de Trajectoire, dans la lignée des travaux présentés en 2.2.

Concrètement, nous recherchons dans les données les unités avec le code « P » (Trajectoire) en catégorie conceptuelle (quels que soient les sous-types), et examinons tous les codes morphosyntaxiques correspondants. Nous obtenons le résultat schématisé dans le tableau 3, où chaque ligne correspond à une unité encodant de la Trajectoire. Pour chaque unité, le tableau indique l'énoncé d'origine, la langue et le code morphosyntaxique correspondant (la forme linguistique elle-même n'est pas indiquée). Le tableau 3 est limité aux deux premières

unités de notre recherche (sur un total de 821 occurrences), qui appartiennent aux énoncés aho_01 et aho_02 de la langue AHO (Ahom, famille Taï-Kadaï) et codées « Hv » (tête verbale).

sentence	language	ms_code
aho_01	aho	Hv
aho_02	aho	Hv
…		

Tableau 3 : Résultats de la recherche des unités encodant la Trajectoire

Les données ainsi obtenues nous permettent d'établir le tableau 4, qui indique pour chaque sous-famille de langues la proportion de *loci* syntaxiques (têtes, relateurs, et satellites) encodant le composant sémantique de Trajectoire.

	Tête	Relateur	Satellite
Afro-Asiatique, Tchadien	0,81	0,19	0,00
Austro-Asiatique, Bahnarique	0,63	0,37	0,00
Austro-Asiatique, Viet-Muong	1,00	0,00	0,00
Austronésien, Océanique	0,27	0,60	0,13
Coréanique	0,17	0,50	0,33
Hmong-Mien, Hmongique	1,00	0,00	0,00
Indo-Européen, Germanique	0,11	0,53	0,37
Indo-Européen, Hellénique	0,00	0,29	0,71
Indo-Européen, Indo-Aryen	0,33	0,67	0,00
Indo-Européen, Roman	0,49	0,51	0,00
Indo-Européen, Slave	0,04	0,51	0,45
Japonique	0,00	0,67	0,33
Kx'a, ǂHoan	0,50	0,50	0,00
Kx'a, Ju-Kung	0,56	0,16	0,28
Maya	0,60	0,30	0,10
Nadahup	0,52	0,48	0,00
Niger-Congo, Bantou	0,57	0,34	0,09
Niger-Congo, Edoïde	0,78	0,22	0,00

Légende
1,00
0,90
0,80
0,70
0,60
0,50
0,40
0,30

Niger-Congo, Jukunoïde	0,78	0,22	0,00	**0,20**
Niger-Congo, Kwa	0,60	0,36	0,04	**0,10**
Niger-Congo, Senufo	0,78	0,22	0,00	**0,00**
Nilo-Saharien, Saharien	0,54	0,46	0,00	
Ouralien, Fennique	0,14	0,86	0,00	
Ouralien, Ougrien	0,00	0,45	0,55	
Pama-Nyungan	0,14	0,57	0,29	
Sandawe	0,50	0,50	0,00	
Sino-Tibétain, Bodique	0,22	0,33	0,44	
Sino-Tibétain, Lolo-Birman	0,44	0,56	0,00	
Sino-Tibétain, Sinitique	0,12	0,44	0,44	
Tacana	0,41	0,59	0,00	
Tai-Kadai, Kadaï	1,00	0,00	0,00	
Tai-Kadai, Taï	0,72	0,27	0,01	
Tuu, !Ui	0,60	0,40	0,00	

Tableau 4 : Proportion de Têtes, Relateurs et Satellites encodant de la Trajectoire pour chaque famille de langues

On observe dans ce tableau que l'encodage de la Trajectoire varie grandement entre les familles de langues. Certaines langues comme les langues viet-muong (austro-asiatique) encodent exclusivement, dans nos données, la trajectoire dans la Tête, comme l'indique la couleur rouge, signe que la totalité des formes encodant de la Trajectoire sont des Têtes. De même, les langues slaves (famille indo-européenne) privilégient plusieurs *loci* possibles tels que le Satellite et le Relateur. Ces multiples stratégies sont indiquées en jaune, signe que la moitié des formes encodant de la Trajectoire sont des Satellites, mais aussi des relateurs.

Cinq principales tendances sont observées dans nos données. Tout d'abord, l'encodage de la Trajectoire dans la **Tête** (H) est très présent dans les langues hmongiques (famille hmong-mien), kadaï (famille taï-kadaï) et viet-muong (famille austro-asiatique). Les langues fenniques (famille ouralienne), quant à elles, encodent principalement la Trajectoire dans un **Relateur** (R).

Concernant les langues privilégiant plusieurs stratégies, les langues ju-kung (famille kx'a, anciennement khoïsan) encodent la Trajectoire dans la **Tête** et le **Satellite** (HS). On recense également des combinaisons **Tête** et **Relateur** (HR) dans les langues niger-congo, tchadiennes (famille afro-asiatique), sahariennes (famille nilo-saharienne), ǂhoan (famille kx'a, anciennement khoïsan) et sandawe (isolat, anciennement khoïsan), !ui (famille tuu,

anciennement khoïsan), nadahup, maya, bahnariques (famille austro-asiatique), taïes (famille taï-kadaï), océaniques (famille austronésienne), indo-aryennes, romanes (famille indo-européenne), tacana, et lolo-birmanes (famille sino-tibétaine). Enfin, la dernière tendance observée est la stratégie **Relateur** et **Satellite** (RS), majoritaire en japonais et en coréen, ainsi que dans les langues slaves, germaniques, helléniques (famille indo-européenne), pama-nyunga, ougriennes (famille ouralienne), sinitiques et bodiques (famille sino-tibétaine).
On note que les types S (satellite) et HRS (tête-relateur-satellite) ne sont dominants dans aucune famille de langues, bien qu'attestés dans différentes langues (respectivement exemples 5a et 5b).

5. a. Type S (satellite) : Anglais (Indo-Européen, Germanique)

He climbs **down**. (Fortis & Vittrant 2016)

b. Type HRS : Grec ancien (Indo-Européen, Hellénique ; Grèce)

énth'	hé:	g'	**eis-elthoûsa**.
to.there	she	LNK	**to-go_to**.PART.AOR.NOM

'Therein she entered' (Fortis & Vittrant 2016)

5. CONCLUSION

Dans cette présentation de notre méthodologie, nous avons souhaité montrer comment utiliser des données linguistiques pour créer une ontologie des relations spatiales. Cette ontologie simplifie l'analyse de l'expression de la Trajectoire à partir de données provenant de langues différentes (typologiquement, génétiquement, et géographiquement). *Via* un système de codage, nous avons associé des formes linguistiques à leurs caractéristiques morphosyntaxiques (fonction syntaxique, partie du discours, catégorie grammaticale), et à leurs caractéristiques sémantiques (catégories conceptuelles du domaine spatial).

Type	**Sous-type**	**Exemple de familles où le type est dominant**
H		Langues hmong, kadaï, viet-muong
R		Langues fennique
S		*Non-attesté*
HR	**Plus de H**	Langues niger-congo, tchadiennes, nadahup, maya, taï
	Plus de R	Langues océaniques, romanes, tacana, lolo-birmanes

	Autant de H et R	Langues ǂhoan, sandawe
HS		Langues ju-kung
RS	**Plus de R**	Japonais, coréen, langues slaves, germaniques, pama-nyungan
	Plus de S	Langues ougriennes, hélleniques
	Autant de R et S	Langues sinitiques, bodiques
HRS		*Non-attesté*

Tableau 5 : Récapitulatif des *loci* de la Trajectoire

En utilisant cette ontologie, nous avons également établi une typologie préliminaire de l'encodage du composant de Trajectoire. Nous avons montré que cet encodage varie selon les familles de langues : les stratégies préférées par les langues pour encoder la Trajectoire (H : Tête, R : Relateur, S : Satellite, ou stratégies multiples) différent selon leur appartenance à une famille. Ceci est résumé dans le tableau 5.

ABBREVIATIONS

3 troisième personne, **ABL** ablatif, **ACC** accusatif, **AOR** aoriste, **CLF** classificateur, **CVB** coverbe, **GEN** génitif, **LNK** linker, **NOM** nominatif, **PART** participe, **PERL** perlatif, **PL** pluriel, **PST** passé

BIBLIOGRAPHIE

Aske, Jon (1989). Path predicates in English and Spanish : A closer look. *Proceedings of the fifteenth annual meeting of the Berkeley Linguistic Society*, 1-14.

Fagard, Benjamin, & Kopecka, Anetta (2021). Source/Goal (a)symmetry : A comparative study of German and Polish. In A. Kopecka & M. Vuillermet (Éds.), *Source-Goal (a)symmetries across languages.* Amsterdam: John Benjamins. 130-171.

Farrar, Scott, & Langendoen, D. Terence (2003). A linguistic ontology for the semantic web. *GLOT international*, *7*(3), 97-100.

Fortis, Jean-Michel, & Vittrant, Alice (2011). L'organisation syntaxique de l'expression de la trajectoire : Vers une typologie de la construction. *Faits de Langues - Les Cahiers, 3*, 71-98.

Fortis, Jean-Michel, & Vittrant, Alice (2016). On the morphosyntax of path-expressing constructions : Toward a typology. *STUF Language Typology and Universals*, *69*(3), 341-374.

Genesereth, Michael, & Nilsson, Nils (1987). *Logical Foundations of Artificial Intelligence*. Burlington: Morgan Kaufmann.

Gruber, Thomas R. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition*, *5*(2), 199-220.

Guarino, Nicola, Oberle, Daniel, & Staab, Steffen (2009). What is an ontology? In S. Staab & R. Studer (Éds.), *Handbook on ontologies*. Berlin: Springer-Verlag Berlin Heidelberg. 1-17.

Guillaume, Antoine, & Koch, Harold (2021). Introduction : Associated Motion as a grammatical category in linguistic typology. In A. Guillaume & H. Koch (Éds.), *Associated motion*. Berlin, New York : De Gruyter. 3-30.

Imbert, Caroline, Grinevald, Colette, & Söres, Anna (2011). Pour une catégorie de « satellite » de Trajectoire dans une approche fonctionnelle-typologique. *Faits de Langues - Les Cahiers, 3*, 99-116.

Ishibashi, Miyuki, Kopecka, Anetta, & Vuillermet, Marine (2006). *Trajectoire : Matériel visuel pour élicitation des données linguistiques*. Fédération de Recherche en Typologie et Universaux Linguistiques.

Matsumoto, Yo (2003). Typologies of Lexicalization Patterns and Event Integration : Clarifications and Reformulations. In S. Chiba (Éd.), *Empirical and theoretical investigations into language : A festschrift for Masaru Kajita*. Tokyo: Kaitakusha. 403-418.

Matsumoto, Yo, Akita, Kimi, & Takahashi, Kiyoko (2017). The functional nature of deictic verbs and the coding patterns of Deixis : An experimental study in English, Japanese, and Thai. In I. Ibarretxe-Antunano (Éd.), *Motion and Space across Languages : Theory and applications*. Amsterdam: John Benjamins. 95-122.

Morita, Takahiro (2009). *La catégorisation des verbes de déplacement en japonais et en français* [Thèse de doctorat]. EHESS.

Morita, Takahiro (2011). Intratypological Variations in Motion Events in Japanese and French. *CogniTextes*. *Revue de l'Association Française de Linguistique Cognitive, 6*, 6. [En ligne].

Schalley, Andrea (2019). Ontologies and ontological methods in linguistics. *Language and Linguistics Compass*, *13*(11), 1.

Schalley, Andrea, Musgrave, Simon, & Haugh, Michael (2014). Accessing phonetic variation in spoken language corpora through non-standard orthography. *Australian Journal of Linguistics*, *34*(1), 139-170.

Seifen, Karl, Vittrant, Alice, Song, Jinke, & Bunkham, Nichuta (2021, 2/07). Path and Deixis as distinct concepts in Burmese, Thai, and Chinese. *NAMED 2021*, Paris, ENS.

Slobin, Dan I., & Hoiting, Nini (1994). Reference to movement in spoken and signed languages : Typological considerations. *Proceedings of the Berkeley Linguistics Society*, *20*, 487-505.

Sowa, John F. (2003). *Ontology* [Blog]. http://www.jfsowa.com/ontology/

Talmy, Leonard (1972). *Semantic Structures in English and Atsugewi* [Thèse de doctorat]. University of California.

Talmy, Leonard (1985). Lexicalization patterns : Semantic structure in lexical form. In T. Shopen (Éd.), *Language Typology and Syntactic Description—Grammatical Categories and the Lexicon* (Vol. 3). Cambridge: Cambridge University Press. 57-149.

Talmy, Leonard (2000). *Toward a Cognitive Semantics : Typology and Process in Concept Structuring* (Vol. 2). Cambridge: MIT Press.

Vandeloise, Claude (1986). *L'Espace en français. Sémantique des propositions spatiale*. Paris : Seuil.

Vuillermet, Marine (2012). *A grammar of Ese Ejja, a Takanan language of the Bolivian Amazon* [Thèse de Doctorat]. Université Lumière Lyon 2.

Vuillermet, Marine, & Kopecka, Anetta (2019). Trajectoire : A methodological tool for eliciting Path of motion. *Methodological Tools for Linguistic Description and Typology, Language Documentation & Conservation Special Publication, 16*, 340-353.

Wälchli, Bernhard (2001). A typology of displacement (with special reference to Latvian). *STUF - Language Typology and Universals, 54*, 298-323.

Les ressources numériques au service de la collaboration internationale : Étude des gestes produits lors d'un discours narratif par des personnes avec aphasie, avant et après la thérapie POEM

Victoria VALENTIN, Ana-Inès ANSALDO, Edith DURAND

LaboAnsaldo, Laboratoire de plasticité cérébrale, communication et vieillissement

victoria-valentin@outlook.fr

RESUME

Étudier le langage à l'ère numérique : une gageure à la fois méthodologique, pratique et théorique. Dans le contexte de l'époque contemporaine, les chercheurs doivent composer avec de nouvelles ressources en apprenant à maîtriser leurs potentiels. Notre étude s'inscrit dans ce processus en ciblant un objectif d'étude de la communication de personnes présentant une aphasie à l'aide d'une ressource numérique et un objectif de formation en recherche clinique à distance (France-Québec) dans le respect des normes éthiques et d'utilisation des données des deux pays, à travers l'étude des effets d'une thérapie en orthophonie sur les gestes produits lors d'un discours narratif de personnes avec aphasie.

L'aphasie est un trouble du langage acquis à la suite d'une lésion cérébrale. Le symptôme principal de l'aphasie est l'anomie ou manque du mot (Chomel-Guillaume, Leloup, Bernard, Bakchine, 2010). La présente étude fait partie d'une étude plus large incluant 10 personnes avec aphasie qui ont bénéficié de la thérapie Personalized Observation, Execution and Mental imagery (POEM) ayant déjà montré des résultats significatifs avec une amélioration de la récupération de la dénomination de verbes à la suite de la thérapie (Durand 2018, 2020, 2021). Or, les effets de POEM sur la communication non verbale n'ont pas encore été étudiés. C'est donc l'objet de cette étude. Ainsi, à l'aide du logiciel d'annotations EUDICO Linguistic Annotator (ELAN), une étude des types et des fonctions de gestes a été réalisée à partir de l'enregistrement vidéo d'une tâche de discours narratif chez 5 des 10 participants, avant et après POEM. ELAN est un outil utilisable dans les sciences du langage pour étudier la communication non verbale et est apparu pertinent pour étudier les effets d'une thérapie en orthophonie. Enfin, pour assurer la fiabilité de ces annotations, un protocole d'annotation des

données basé sur les articles de Sekine et al (2013,2021) et Kong et al. (2015) ainsi qu'un protocole de vérification de ces annotations (accord inter-juge) ont été créés.

Les premiers résultats recueillis montrent chez 5 participants une utilisation accrue des gestes du type pantomimes, métaphores et déictiques concrets ainsi que l'utilisation accrue de gestes ayant pour fonction la récupération lexicale et les moyens de communication alternatifs descriptifs. Ces changements observés rejoignent le rationnel de POEM, à savoir l'utilisation de stratégies motrices pour faciliter la dénomination d'actions. Ces premières données vont dans le sens du développement de nouvelles stratégies après POEM, à savoir l'utilisation de gestes pour retrouver les mots et compenser le manque du mot. Par ailleurs, la formation en recherche clinique à distance a pu se réaliser aisément avec une acquisition de compétences pour l'annotation des gestes à l'aide du logiciel ELAN notamment et ce, tout en respectant les normes éthiques et d'utilisation des données. Il s'avère donc pertinent de pouvoir utiliser toutes les ressources numériques pour contribuer à la recherche en sciences du langage.
***Mots-clés** : aphasie - anomie - POEM - observation d'action - exécution d'action - imagerie mentale - ELAN*

ABSTRACT

Studying language in the digital age is a methodological, practical and theoretical challenge. In today's context, researchers must deal with new resources by learning to master their potential. Our study is part of this process by targeting a study of the communication of people with aphasia using a digital resource and a training objective in distance clinical research (France-Quebec) in the respect of the ethical standards and the use of the data of the two countries, through the study of the effects of a therapy in speech therapy on the gestures produced during a narrative speech of people with aphasia

Aphasia is a language disorder acquired as a result of brain damage. The main symptom of aphasia is anomia or lack of words (Chomel-Guillaume, Leloup, Bernard, Bakchine, 2010). The present study is part of a larger study including 10 people with aphasia received Personalized Observation, Execution and Mental imagery (POEM) therapy having already shown significant results with an improvement in verb naming recovery following therapy (Durand 2018, 2020, 2021). However, the effects of POEM on nonverbal communication have not yet been studied. Therefore, this is the focus of this study. Thus, using the EUDICO Linguistic Annotator (ELAN) annotation software, a study of gesture types and functions was conducted based on video recordings of a narrative speech task in 5 participants, before and after POEM. ELAN is a tool that can be used in the language sciences to study nonverbal

communication and has been found to be relevant for studying the effects of therapy in speech therapy. Finally, to ensure the reliability of these annotations, a data annotation protocol based on the articles by Sekine et al (2013,2021) and Kong et al (2015) as well as a protocol for verifying these annotations (inter-judge agreement) were created.

Initial results collected show for the 5 participants an increased use of pantomime, metaphor, and concrete deictic type gestures as well as increased use of gestures with the function of lexical retrieval and descriptive alternative means of communication. These observed changes are consistent with the rationale of POEM, namely the use of motor strategies to facilitate the naming of actions. These initial data support the development of new strategies after POEM, namely the use of gestures to retrieve words and compensate for the lack of the word. Moreover, training in remote clinical research was easily achieved with the acquisition of skills in annotating gestures using ELAN software, in particular, while respecting ethical standards and the use of data. It is therefore relevant to be able to use all digital resources to contribute to research in language sciences.

Key words*: aphasia - anomia - POEM - action observation - gesture execution - mental imagery - ELAN*

1. INTRODUCTION

Le contexte actuel de développement des échanges internationaux en matière de recherche et l'évolution constante des outils technologiques et du Web 2.0 (Eysenbach, 2008) posent la question des méthodes et du matériel utilisés lors d'une étude. En sciences du langage, les études à travers les langues et plus spécifiquement en orthophonie sur la communication interpersonnelle dans une langue donnée, amènent les chercheurs à se doter d'outils de performance, avec notamment l'utilisation de certaines ressources numériques. Parmi elles, le logiciel d'annotations EUDICO Linguistic Annotator (ELAN) est un outil permettant d'analyser la communication non verbale, notamment les gestes : étant liés aux représentations verbales des mots, cette analyse est pertinente pour recueillir des informations sur le fonctionnement des personnes neurotypiques, mais surtout des personnes avec des troubles de la communication de type aphasie. En plus, un autre avantage de cette ressource est d'être un support d'étude et de collaboration *in situ* mais aussi internationale, ce qui est d'un grand intérêt pour la recherche clinique en orthophonie.

L'aphasie est un trouble du langage acquis entraînant des conséquences délétères sur la communication et dont le symptôme le plus fréquent et le plus persistant est l'anomie, c'est-à-dire le manque du mot (Chomel-Guillaume, Leloup, Bernard, Bakchine, 2010). Il a été démontré que les personnes avec aphasie utilisaient davantage les gestes que les personnes neurotypiques (de Beer et al. 2019, Feyereisen 1983 ; Sekine et al. 2013). Cela pourrait s'expliquer par le besoin de libérer la mémoire de travail verbale (Goldin-Meadow et al., 2001) et par le gain pouvant être obtenu dans la communication, tant pour celui qui s'exprime que pour celui qui écoute (Kita et al. 2017). En effet, la production des gestes (désignant ici les mouvements spontanés liés au discours et le suppléant parfois) se retrouve très intriquée dans celle du langage. Ils peuvent être de différentes natures et revêtir des fonctions diverses en fonction du message à transmettre

Chez les personnes avec aphasie, des études ont montré que les gestes aidaient à clarifier les propos lors de la productions d'erreurs (des paraphasies), à retrouver un mot (Akhavan et al., 2018 ; Kistner, Dipper & Marshal 2019 ; Lanyon & Rose, 2009) et à améliorer et compléter le discours (Dipper et al 2015 ; van Nispen et al. 2017).

Avec pour objectif d'améliorer la communication verbale chez ces patients, les études sur la création et le développement de la thérapie en orthophonie POEM (Durand 2018, 2020, 2021) ont permis de mesurer des effets d'amélioration sur la dénomination des verbes. Notre étude s'inscrit dans la continuité de ce travail en proposant de mesurer les effets sur la communication

non verbale. Ces derniers peuvent être mesurés à l'aide d'ELAN, en annotant les types et les fonctions des gestes produits par les participants aphasiques à partir d'enregistrements vidéo. Sur 10 participants, les productions gestuelles vont être analysées pour les 5 premiers participants au cours d'une tâche de discours narratif, avant et après la thérapie.

Dans l'optique d'utiliser les méthodes les plus avancées pour étudier différents aspects de la communication entre les personnes avec aphasie, les ressources numériques utilisées dans cette étude sont plurielles, requises à la fois pour la création, le stockage et l'échange des données de l'étude avec la plateforme NextCloud, ainsi que leur analyse avec le logiciel ELAN.
Hypothèses générales :

(1) Les ressources numériques permettent le développement de collaboration internationale de recherche dans le respect des normes éthiques et de la confidentialité des données.
(2) Les effets sur la communication non verbale d'une thérapie en orthophonie peuvent être mesurés à l'aide d'un logiciel d'annotations des gestes.

2. MATERIEL ET METHODE

3.1. A propos de la collaboration internationale

3.1.1 Origine de la collaboration

La première auteure Victoria Valentin a réalisé son mémoire de master sous la direction d'Edith Durand (Valentin-Nguyen-Dao, 2021) et son souhait était de le poursuivre dans le cadre de son projet doctoral autour de la création d'une thérapie en orthophonie. Dans le but d'élargir ses expériences dans la recherche et d'atteindre les objectifs précités, le laboratoire du Pre Ansaldo m'a permis d'effectuer un stage à distance. Les objectifs initiaux de cette collaboration sont la formation à l'analyse des gestes et à l'utilisation du logiciel ELAN.

3.1.2 Respect du cadre éthique de la recherche

L'article de Eysenbach (2008) évoque les changements impliqués par l'utilisation d'Internet ayant transformé la recherche biomédicale et ouvert de nouvelles perspectives, notamment pour la collaboration entre les chercheurs et l'utilisation d'outils et d'approches spécifiques.
Dans le cadre de la présente étude, les auteures devant travailler à distance (France – Quebec), différents outils ont été mis en place pour permettre cette collaboration :

- En amont de cette collaboration, un engagement de confidentialité relatif à la confidentialité des informations et à leur utilisation a été signé, dans le cadre du projet de recherche intitulé « Récupération de l'aphasie : application de la recherche clinique

pour optimiser l'intervention auprès de personnes présentant une aphasie après un Accident Vasculaire Cérébral (AVC)." du Laboratoire Plasticité cérébrale, Communication et Vieillissement.

- Au préalable, les participants de l'étude ont signé un formulaire d'information et de consentement d'utilisation des données, telles que les vidéos, dans le cadre de ce projet de recherche, approuvé par le Comité d'éthique de la recherche Vieillissement-neuroimagerie (CMER RNQ 13-14 – 003).

- Dans cette perspective de respect de la confidentialité et de l'utilisation des données des participants, nous avons choisi la plateforme Nextcloud pour l'ensemble des échanges de données. Cette ressource permet de stocker les données sur un serveur crypté et a donc permis de sécuriser nos partages dans le respect des recommandations de chaque pays (la France et le Québec). Ce serveur a été choisi pour sa conformité à l'HIPAA (Health Insurance Portability and Accountability Act), en vigueur outre-Atlantique et à la RGPD (Règlement Général sur la Protection des Données - RGPD/CNIL, 2018) en vigueur sur le territoire de l'Union Européenne.

3.2 A propos de l'étude

3.2.1 La thérapie POEM

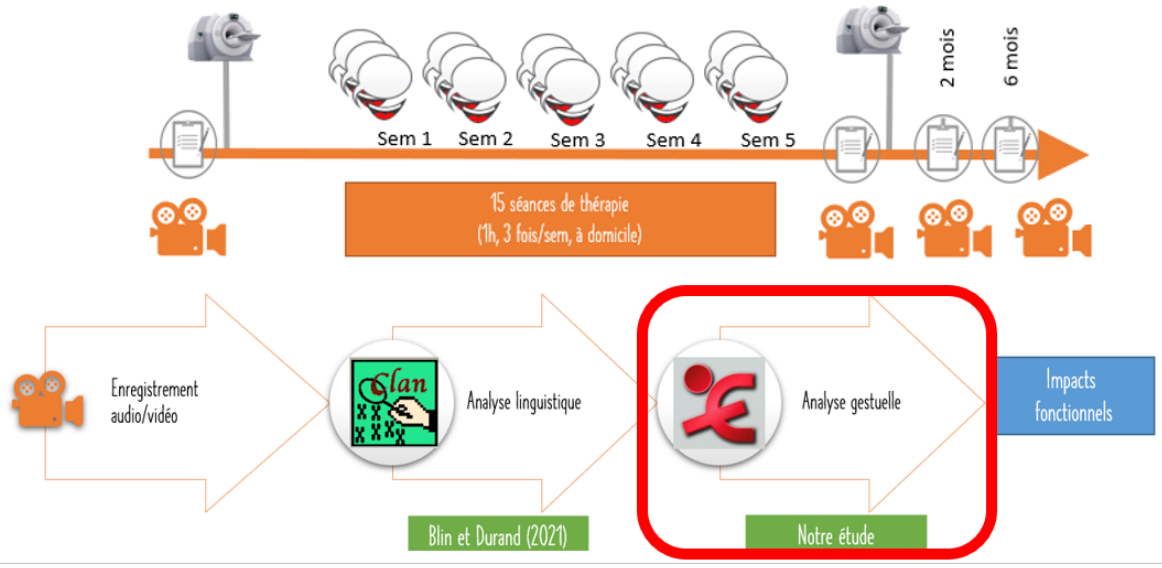


Figure 1 : Déroulement de la thérapie POEM et études successives

Créée et développée par Durand et al (2018, 2020, 2021), la thérapie POEM est la première à combiner trois stratégies sensori-motrices dans un but de réhabilitation du langage, à savoir

l'observation d'action, l'exécution d'action et l'imagerie mentale. Les études préalables ont montré des améliorations significatives sur la dénomination des verbes, jouant un véritable rôle de pivot dans la construction morphosyntaxique (Conroy et al., 2006) et concernant les items traités comme non traités.

Nous précisons que l'ensemble des auteurs est familiarisé avec la thérapie POEM. Dans un premier temps, un membre de l'équipe d'Ana Inès Ansaldo a réalisé l'analyse verbatim du discours avec le logiciel CLAN (Blin, 2021), puis notre étude a suivi pour l'analyse des gestes à l'aide du logiciel ELAN.

3.2.2 Présentation des participants

Une population de cinq participants avec aphasie a été inclue dans cette étude, suivant leur accord préalable de participation à la recherche. Ils sont âgés de 55 à 74 ans et présentent différents types d'aphasies.

Concernant les critères d'inclusion, les participants devaient :

- présenter une aphasie chronique (impliquant que l'AVC devait dater d'un an au moins), afin de ne pas attribuer à tort les effets de la thérapie à la période de récupération spontanée survenant dans l'année qui suit l'AVC (Carey et Seitz, 2007).
- avoir une lésion unique dans l'hémisphère gauche ;
- présenter une anomie des verbes modérée à sévère ;
- être droitier avant l'AVC ;
- pratiquer le français comme langue maternelle ;
- être âgé de 60 ans et plus ;
- ne présenter aucune autre pathologie d'origine neurologique, cognitive ou psychiatrique.

Le tableau ci-dessous présente les informations relatives à leur profil :

Participant	Age (en années)	Sexe	Années d'études	Temps après l'AVC (en mois)	Type d'aphasie
P1	74	H	18	49	Broca
P2	55	F	12	36	Broca
P3	59	H	12	56	Broca
P4	72	F	11	408	Transcortical motrice
P5	56	F	11	252	Broca

Tableau 1 : Profil des participants de l'étude.

Nous nous sommes assurées que le groupe était le plus homogène possible, au regard des critères d'éligibilité mentionnés. Ces critères ont été décidés selon les plus hauts standards en cours en aphasiologie, incluant les standards impliquant des investigations en neuroimagerie (Cf. Thèse Edith Durand 2020). Des références peuvent être trouvées dans les articles déjà publiés et revus par des pairs en 2021 et 2023. L'étude présentée lors de cette conférence fait partie d'une étude de plus grande envergure. Les lecteurs sont dirigés vers les publications relatives à cette étude avec les références suivantes (Durand et al, 2021, Durant et al. 2023).

3.2.3 Enregistrements vidéo

Chacun des 10 participants est enregistré à l'aide d'un enregistrement vidéo au cours de l'évaluation ayant lieu avant et après la thérapie pour mesurer les effets de la thérapie. Nous disposons donc d'un total de 10 enregistrements vidéo d'une tâche de discours narratif (Cendrillon, tâche classique en orthophonie), 5 avant et 5 après la thérapie POEM. Ils ont été effectués avec une caméra FaceTime HD 720p, intégrée dans un MacBook Pro (13 pouces, 2015).

La durée moyenne des vidéos avant la thérapie est de 5 minutes et 7 secondes et la durée moyenne des vidéos après la thérapie est de 4 minutes et 40 secondes.

Participant	Durée de la vidéo PRE	Durée de la vidéo POST
P1	07 min 15s	06 min 50 s
P2	02 min 44 s	04 min 04 s
P3	05 min 35 s	02 min 50 s
P4	05 min 39 s	03 min 04 s
P5	04 min 23 s	07 min 53 s
Durée moyenne	05 min 07 s	04 min 40 s

Tableau 2 : Durée des vidéos avant et après la thérapie, pour chaque participant.

Selon Hilverman et al (2016), une personne produira davantage de gestes au cours d'une tâche de discours narratif spontané pour une histoire ayant une structure épisodique, et selon Stark et al. (2021), les gestes faciliteront davantage l'accès lexical si les deux personnes sont à même de reconnaître les mots-cibles à partir d'une histoire faisant partie d'une culture commune. De plus, comme conseillé dans cette dernière étude, la personne ayant le rôle d'expérimentateur

était présente pour écouter et offrir une expérience d'échange le plus naturel possible au participant, ponctuant parfois l'écoute de petites phrases et de gestes spontanés.

3.3 Analyse des gestes

3.3.1 Formation au logiciel ELAN

ELAN (EUDICO Linguistic ANnotator) est un outil d'annotation permettant de créer, d'éditer, de visualiser et de rechercher des annotations sur fiches multimodaux (audio, vidéo) et se trouve particulièrement adapté à l'annotation des gestes. Ce logiciel est téléchargeable depuis le site https://archive.mpi.nl/tla/elan/download. Dans l'onglet Documentation de la page de téléchargement, on trouve également un manuel complet consultable en ligne ou au format PDF. Les deux grands avantages facilitant l'utilisation de cette ressource numérique sont son accès gratuit et la maintenance de sa plateforme à jour.

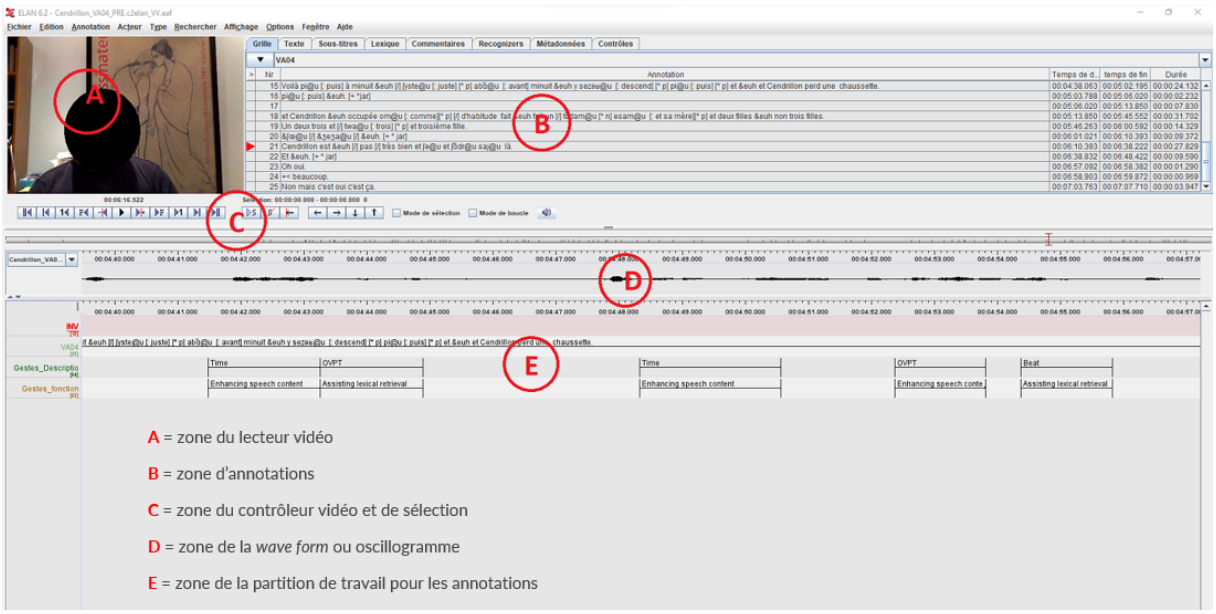


Figure 2: Interface du logiciel ELAN avec un enregistrement vidéo de l'étude

Ci-dessus, l'interface du logiciel ELAN avec l'un des enregistrements de l'étude ouvert.

(1) La zone A correspond au lecteur vidéo, où s'affiche l'entrée visuelle du fichier.

(2) La zone B est l'espace où s'affichent les annotations effectuées au fur et à mesure du travail, sur la nature des gestes et leur fonction.

(3) La zone C est celle du contrôleur vidéo et de sélection.

(4) La zone D permet de visualiser l'entrée audio sous forme d'un oscillogramme, ou *wave form*.

(5) La zone E correspond à la zone de partition de travail pour les annotations à effectuer. Pour ce faire, on utilise la fonction de vocabulaire contrôlé sur ELAN, qui consiste à

avoir un nombre limité de natures et de fonctions de gestes, disponibles sur une liste déroulante.

Trois fonctions d'ELAN ont été utiles dans nos travaux :

1) La fonction *chat2elan* a permis d'insérer le verbatim édité avec CLAN[14] (Blin et Durand, 2021) sur notre interface de travail, comme suggéré dans l'étude de Stark et al. (2021) afin de permettre un recueil d'informations sur les relations entre le discours et les gestes. Cette commande a été retrouvée sur le site https://archive.mpi.nl/forums/.
2) La fonction *Statistic for Multiple* Files nous a permis d'extraire les données d'annotations relatives aux types et aux fonctions des gestes vers un fichier Excel, nous permettant ainsi de faire un comparatif entre les gestes produits avant versus après la thérapie, dans le but d'en mesurer les effets.
3) La fonction *Calculate Inter-Annotator Reliability* a participé à mesurer l'accord inter-juge pour une vérification des annotations le plus objective possible.

3.3.2 Protocole d'annotation des gestes

L'objectif des analyses de ces enregistrements vidéo a été de comptabiliser les gestes produits par les participants et leurs fonctions au cours d'une tâche de récit. Pour favoriser la bonne attribution des annotations, E. Durand a élaboré un guide d'annotation appuyé sur le codage proposé par Kong et al. (2015) et de Sékine et al. (2013, 2021), qui se sont intéressés spécifiquement aux gestes produits par les personnes avec aphasie. Ces articles ont déjà été publiés et éprouvés auprès d'une telle population.
Le guide d'annotations permet donc de préciser le type de gestes produits, à savoir :

(1) Un geste référentiel (**referential**) est utilisé pour assigner des entités à des référents.

(2) Un geste déictique concret (**concrete deictic**) indique une référence concrète dans l'environnement physique.

(3) Un geste pointant vers soi (**pointing to self**) permet au locuteur de pointer son propre corps pour parler de lui-même.

[14] Le programme CLAN (Computerized Language Analysis) permet de transcrire des corpus audio ou vidéo de discours spontanés et facilite leur analyse avec l'ajout d'annotations linéaires (par exemple pour représenter des catégories syntaxiques, morphologiques et phonétiques).

(4) Un geste du point de vue l'observateur (**OVPT - Observer View Point**) montre une action concrète, un événement ou un objet observé de loin.

(5) Un geste du point de vue du personnage (**CVPT - Character View Point**) utilise le corps du locuteur lui-même pour dépeindre une action concrète, un événement ou un objet, comme s'il était le personnage ou l'objet lui-même.

(6) Une pantomime (**pantomime**) est constituée de deux gestes du point de vue du personnage (CVPT) ou plus, qui s'enchaînent en continu dans la même séquence de gestes.

(7) Un geste métaphorique (**métaphoric**) présente une image d'un concept abstrait (ex : la justice).

(8) Un emblème (**emblem**) est un geste dont la forme et la signification ont été établis par convention (ex : faire « ok » avec le pouce levé).

(9) Un geste relatif au temps (**time**) indiquant qu'un espace correspond à un temps (ex : le passé serait représenté vers l'arrière du corps et le futur vers l'avant du corps).

(10) Un battement (**beat**), c'est-à-dire un ou plusieurs mouvements qui n'ont pas de signification particulière mais sont répétitifs et en rythme avec la production du discours.

(11) Une lettre (**letter**) réalisée par des mouvements associés, comme pour écrire dans l'espace ou sur une surface.

(12) Un nombre (**number**) montré par un geste sur les doigts du locuteur.

(13) Un geste sans lien avec la tâche (**no link**) (ex : le participant essuie son nez parce qu'il goutte).

Du point de vue de la rééducation orthophonique, il est intéressant d'analyser quels sont les types de gestes produits, mais surtout la fonction associée à chacun de ces gestes. Ci-dessous, voici le protocole d'annotations pour les fonctions possibles des gestes (ce guide est également appuyé sur les articles Kong et al (2015) et de Sékine et al (2013, 2021) :

(1) Le geste donne des informations additionnelles (**provide additional info**) (ex : le participant dit « ouvrir » et le geste est un mouvement de torsion dans l'air, représentant l'action d'ouvrir un bocal).

(2) Le geste améliore le contenu du discours (**enhancing speech content**) en donnant le même sens que le contenu discours.

(3) Le geste donne un moyen de communication alternatif (**providing alternative means of communication)** en portant une signification qui n'est pas incluse dans le contenu du discours (ex : le pouce en l'air voulant dire « OK » à la question : êtes-vous prêt ?).

(4) Le geste guide et contrôle le flux du discours (**guiding and controlling the flox of speech**) en synchronisant le rythme des mouvements avec le rythme du discours.

(5) Le geste renforce l'intonation ou la prosodie du discours (**reinforcing the intonation or prosody of speech**), en mettant par exemple en relief un mot en particulier.

(6) Le geste destiné à faciliter l'accès lexical (**assisting lexical retrieval)**, pouvant se produire lorsque le participant fait une pause ou une prolongation notable dans son discours.

(7) Le geste destiné à faciliter la reconstruction d'une phrase (**assisting sentence reconstruction**) quand le participant veut montrer la modification de la structure syntaxique.

(8) Pas de fonction spécifique déduite (**no specific function deduced**).

3.3.3 Protocole de vérification des annotations (accord inter-juge)

L'analyse sur corpus vidéo pouvant revêtir une certaine part de subjectivité, une procédure de double évaluation a également été créée afin d'assurer la fiabilité des résultats de l'accord inter-juge (ED et VV). Le manuel d'ELAN indique qu'un accord inter-juge est valable si le kappa est de 0,6 (60% d'accord).

1) Nous avons d'abord annoté la même vidéo de façon indépendante et comparé nos annotations avec la fonction *Calculate Inter-Annotator Reliability*.

2) A la suite de ce premier essai d'annotations, la mesure kappa n'atteignant pas encore le seuil d'acceptabilité, nous nous sommes rencontrées pour discuter des annotations attribuées et nous avons annoté une nouvelle fois une vidéo de façon indépendante.

3) Les mesures kappas avaient dépassé le seuil d'acceptabilité pour les types comme pour les fonctions de gestes.

3. RÉSULTATS

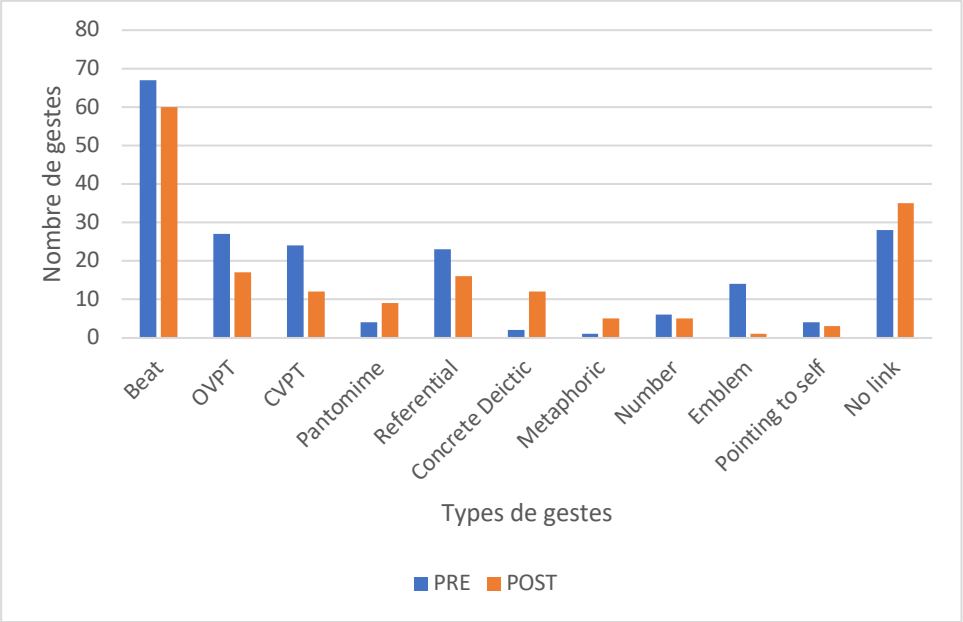


Figure 3 : Evolution des types de gestes avant et après la thérapie

L'axe des abscisses correspond aux types de gestes et l'axe des ordonnés donne le nombre de gestes d'un type donné, produits par les participants. Les données en bleu correspondent aux types de gestes produits avant la thérapie et les données en rouge correspondent à ceux qui ont été produits après la thérapie. L'indice Kappa validant l'accord inter-juge est de 0.6216.
Plusieurs évènements retiennent notre attention :

1) Les OVPT, les CVPT, les gestes référentiels et les battements sont moins nombreux.

2) Le nombre de pantomimes, de déictiques concrets et de métaphores ont augmenté.

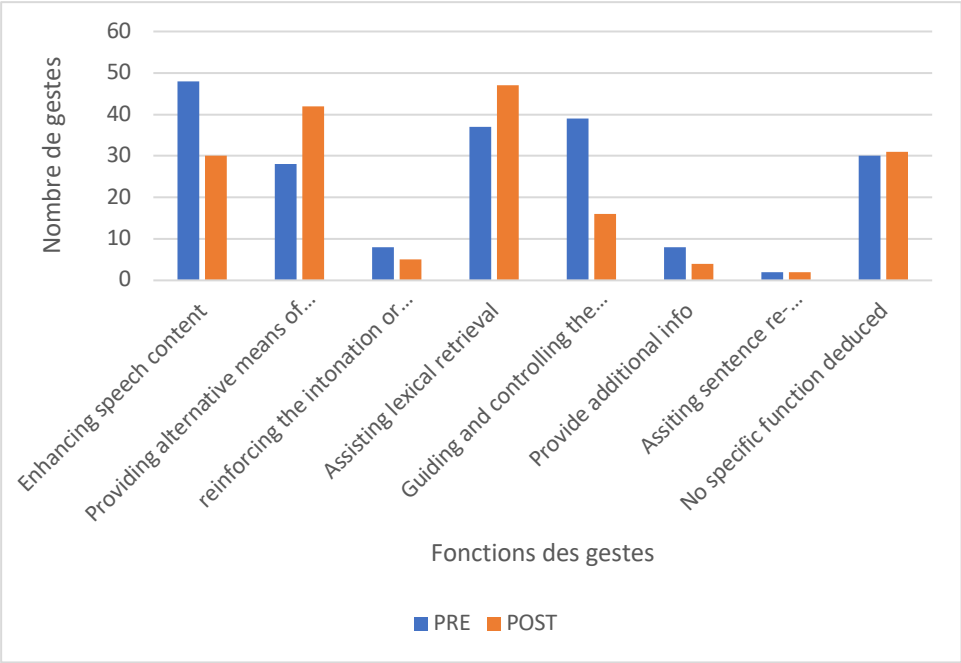


Figure 4 : Evolution des fonctions de gestes avant et après la thérapie

L'axe des abscisses correspond aux fonctions de gestes et l'axe des ordonnés donne le nombre de gestes d'une fonction donnée, produits par les participants. A nouveau, les données en bleu correspondent aux types de gestes produits avant la thérapie et les données en rouge correspondent à ceux qui ont été produits après la thérapie. L'indice Kappa validant l'accord inter-juge est de 0,7812.

Trois événements notables retiennent notre attention :

1) Il y a davantage de gestes qui servent en tant que moyens de communication alternative, c'est-à-dire que de nouvelles informations du discours sont données à travers le geste, après la thérapie.

2) Il y a davantage de gestes qui soutiennent la récupération lexicale.

3) Les gestes aidant à améliorer le contenu du discours transmis ont baissé.

4. DISCUSSION

Pour rappel, nous nous intéressons à l'utilisation des ressources numériques pour l'étude des gestes produits lors d'un discours narratif filmé. Les objectifs de cette étude sont doubles : formation en recherche clinique et étude des gestes chez les patients aphasiques. La discussion portera d'abord sur les résultats obtenus après l'analyse des vidéos des participants, puis sur la part des ressources numériques dans une visée formative.

D'abord sur le plan du matériel et de la méthode utilisée, tout comme suggéré dans l'article de Hilverman et al. (2016) et en tant que tâche classique en orthophonie, la tâche de discours narratif proposée aux participants s'est déroulée à partir d'une histoire ayant une structure épisodique (ici le conte de Cendrillon). Ce choix se conformait également au conseil de Stark et al. (2021) indiquant qu'une histoire issue de la culture commune facilitait la reconnaissance des mots-cibles pour l'évaluateur.

Concernant les résultats, le fait qu'on observe une augmentation des gestes ayant pour fonction la récupération lexicale va dans le sens du rationnel de la thérapie POEM qui est de faciliter l'accès lexical. De plus, les évènements relevés pour les fonctions de gestes peuvent être intéressants à mettre en perspective : on peut faire l'hypothèse que l'énergie qui était dédiée à améliorer le discours par des gestes est mise au service d'autres fonctions par rapport à avant la thérapie (ici la récupération lexicale et la mise en place d'un moyen de communication alternatif). De plus, concernant les types de gestes produits, on peut faire l'hypothèse que le nombre de CVPT relevés n'aurait baissé qu'au bénéfice de l'emploi plus fréquent des pantomimes, constituées d'un enchaînement de plusieurs CVPT, ce qui montrerait une transition vers une expression plus riche en informations. L'augmentation des métaphores pourrait indiquer que les participants atteindraient plus facilement un niveau d'abstraction dans leur discours et celle des déictiques concrets pourrait suggérer que les participants auraient un allègement du coût cognitif pour une tâche de discours narratif en utilisant davantage leur environnement concret. Les résultats d'une étude préalable ont montré que la thérapie pourrait avoir un impact de la thérapie sur le fonctionnement exécutif et de la mémoire de travail, cela pourrait être corroboré par une amélioration de l'utilisation des gestes observée dans notre étude (Durand, Valentin, Moritz-Gasser, 2022).

Cette analyse a été faite au niveau du groupe, nous avons actuellement des résultats préliminaires pour les cinq premiers participants sur les dix qui ont participé initialement. Ces résultats ne sont que descriptifs et ne permettent pas de faire des analyses statistiques, mais ils font ressortir des changements intéressants, qui sont à confirmer avec la suite de nos recherches. Ils correspondent aux premières réflexions, qui iraient dans le sens de notre seconde hypothèse générale et permettent d'envisager les effets de la thérapie POEM sur la communication non-verbale.

La collaboration internationale avec le Laboratoire de Plasticité cérébrale, Communication et Vieillissement a permis l'analyse des données grâce à l'utilisation du logiciel ELAN et à la création d'un protocole d'annotation appuyé sur les articles de Sekine et al. (2013, 2021) et de

Kong et al. (2015) et d'un protocole de vérification inter-juge (Valentin, Ansaldo & Durand, en préparation).

Étudier le langage à l'ère numérique s'est donc ici accompagné de plusieurs ressources. Notamment NextCloud, la plate-forme de collaboration à distance avec un stockage sécurisé avec un chiffrement des données sur le serveur, ce qui nous a permis d'échanger des vidéos de façon sécuritaire selon les normes des deux pays des collaborateurs (France-Québec), en addition du formulaire de confidentialité d'utilisation des données signé en amont du stage. L'ensemble des enregistrements vidéo ayant été recueillis au Québec dans le cadre d'une étude de plus grande envergure. Enfin ELAN, un logiciel d'accès gratuit est fiable et a permis d'annoter des fichiers multimodaux et d'avoir des accords inter-juge, et ce, même à un niveau de collaboration internationale. Cet outil nous a permis d'avoir un recul sur l'aspect non verbal de la communication de la personne. Grâce aux études préalables, on savait de façon sûre que POEM permettait d'avoir des résultats au niveau des aspects verbaux, et les premières pistes mises à jour relevées dans notre étude paraissent très intéressantes concernant les aspects non verbaux. Comme on l'a dit en introduction, les personnes avec aphasie utilisent plus de gestes, et si en orthophonie on est capable de qualifier les types et les fonctions de ces gestes avec ELAN, on s'oriente vers la conclusion qu'il y a aussi des changements opérés sur les aspects non verbaux. Les stratégies sensorimotrices de l'observation d'action, de l'exécution du geste de l'action et de l'imagerie mentale de cette action favoriseraient donc une modification des types et des fonctions de gestes accompagnant la production verbale. Cette ressource numérique permet donc d'avoir ce point de vue sur la communication avec des troubles acquis, et de prendre en compte les aspects fonctionnels.

5. CONCLUSION

Les analyses des résultats obtenus concernant les gestes, bien que descriptives à ce stade, nous apportent les premières pistes de réflexion sur l'utilisation des gestes par les personnes aphasiques et les effets de la thérapie POEM sur le comportement non verbal. Ceux-ci vont dans le sens de notre hypothèse générale n°2, c'est-à-dire que les effets d'une thérapie en orthophonie sur les gestes peuvent être analysés grâce à un logiciel d'annotation et invitent à approfondir l'analyse avec les vidéos restantes des autres participants de l'étude.

Construite avec un double objectif, cette collaboration internationale illustre bien la nouvelle façon d'utiliser les ressources numériques actuelles pour la recherche et s'inscrit dans l'eHealth, ou Médecine 2.0. Que ce soit pour communiquer, échanger ou pour procéder à l'analyse des

données recueillies, il a fallu recourir et se former à l'utilisation de plusieurs outils. Parmi eux la plate-forme NextCloud, pour l'échange des données et leur stockage sécurisé et le logiciel ELAN, pour l'analyse des types et des fonctions de gestes dans un corpus d'enregistrements vidéo d'une tâche de discours narratif, avant et après la thérapie POEM.

BIBLIOGRAPHIE

Akhavan, Niloofar, Göksun, Tilbe, & Nozari, Nazbanou (2018). Integrity and function of gestures in aphasia. *Aphasiology*, *32*(11), 1310-1335. https://doi.org/10.1080/02687038.2017.1396573

Blin, Elise. (2021). *Effets de la thérapie sensorimotrice "POEM" dans le discours de patients aphasiques.* **Mémoire**

Chomel-Guillaume, Sophie, Leloup, Gilles, Bernard, Isabelle (2010). *Les aphasies : Évaluation et rééducation*. Elsevier Masson

Conroy, Paul, Sage, Karen., & Ralph, Matthew. A. Lambon. (2006). Towards theory-driven therapies for aphasic verb impairments: A review of current theory and practice. *Aphasiology*, *20*(12), 1159-1185. https://doi.org/10.1080/02687030600792009

de Beer, Carola., de Ruiter, Jan-P., Hielscher-Fastabend Martina, & Hogrefe, Katharina (2019). The Production of Gesture and Speech by People With Aphasia : Influence of Communicative Constraints. *Journal of Speech, Language, and Hearing Research*, *62*(12), 4417-4432. https://doi.org/10.1044/2019_JSLHR-L-19-0020

Durand, Edith. (2020). *Validation comportementale et exploration des corrélats neurofonctionnels de ses effets dans les cas d'aphasie*. **Thèse**. http://hdl.handle.net/1866/24622

Durand, Edith, Berroir, Pierre, & Ansaldo, Ana-Inés (2018). The Neural and Behavioral Correlates of Anomia Recovery following Personalized Observation, Execution, and Mental Imagery Therapy : A Proof of Concept. *Neural Plasticity*, *2018*, e5943759. https://doi.org/10.1155/2018/5943759

Durand, Edith, Masson-Trottier, Michèle, Sontheimer, Anna, & Ansaldo, Ana-Inés (2021). Increased links between language and motor areas : A proof-of-concept study on resting-state functional connectivity following Personalized Observation, Execution and Mental imagery therapy in chronic aphasia. *Brain and Cognition*, *148*, 105659. https://doi.org/10.1016/j.bandc.2020.105659

Durand, Edith, Valentin, Victoria, Moritz-Gasser, Sylvie. (2022). Fonctions exécutives dans l'anomie des verbes: étude de cas de la thérapie POEM auprès d'un participant avec aphasie bilingue [communication orale]. Journée de Neurologie de Langue française, 2022, Strasbourg, France.

Eysenbach, Gunther (2008). Medicine 2.0 : Social Networking, Collaboration, Participation, Apomediation, and Openness. *Journal of Medical Internet Research*, *10*(3), e22. https://doi.org/10.2196/jmir.1030

Feyereisen, Pierre. (1983). Manual Activity During Speaking in Aphasic Subjects*. *International Journal of Psychology*, *18*(1-4), 545-556. https://doi.org/10.1080/00207598308247500

Goldin-Meadow, Susan, Nusbaum, Howard, Kelly, Sencer D., & Wagner, S. (2001). Explaining math : Gesturing lightens the load. *Psychological Science*, *12*(6), 516-522. https://doi.org/10.1111/1467-9280.00395

Hilliard, Caitlin, Cook, Susan Wagner, & Duff, Melissa C. (2016a). Hippocampal declarative memory supports gesture production : Evidence from amnesia. *Cortex; a journal devoted to the study of the nervous system and behavior*, *85*, 25-36. https://doi.org/10.1016/j.cortex.2016.09.015

Kistner, Judith, Dipper, Lucy T., & Marshall, Jane (2019). The use and function of gestures in word-finding difficulties in aphasia. *Aphasiology*, *33*(11), 1372-1392. https://doi.org/10.1080/02687038.2018.1541343

Kita, Sotaro, Alibali, Martha W., & Chu, Mingyuan. (2017). How do gestures influence thinking and speaking? The gesture-for-conceptualization hypothesis. *Psychological Review*, *124*(3), 245-266. https://doi.org/10.1037/rev0000059

Kong, Anthony Pak-Hin, Law, Sam-Po., Kwan, Connie Ching-Yin, Lai, Christy, & Lam, Vivian. (2015). A Coding System with Independent Annotations of Gesture Forms and Functions during Verbal Communication : Development of a Database of Speech and GEsture (DoSaGE). *Journal of nonverbal behavior*, *39*(1), 93-111. https://doi.org/10.1007/s10919-014-0200-6

Lanyon, Lucette, & Rose, Miranda L. (2009). Do the hands have it? The facilitation effects of arm and hand gesture on word retrieval in aphasia. *Aphasiology*, *23*(7-8), 809-822. https://doi.org/10.1080/02687030802642044

Le règlement général sur la protection des données—RGPD | CNIL. (2018). https://www.cnil.fr/. https://www.cnil.fr/fr/reglement-europeen-protection-donnees

Sekine, Kazuki, & Rose, Miranda L. (2013). The Relationship of Aphasia Type and Gesture Production in People With Aphasia. *American Journal of Speech-Language Pathology*, *22*(4), 662-672. https://doi.org/10.1044/1058-0360(2013/12-0030)

Stark, Brielle C., Clough, Sharice, & Duff, Melissa (2021). Suggestions for Improving the Investigation of Gesture in Aphasia. *Journal of Speech, Language, and Hearing*

Research : JSLHR, *64*(10), 4004-4013. https://doi.org/10.1044/2021_JSLHR-21-00125

Valentin Victoria, Ansaldo Ana-Inès & Durand Edith (en préparation). ***Des gestes pour la récupération des verbes. XXIIIèmes Rencontres d'orthophonie 2023, Unadreo Form.***

Valentin-Nguyen-Dao, Victoria (2021), *Création d'un matériel en langue franco-française pour une réplication de la thérapie « Personalized Observation, Execution and Mental imagery (POEM) » ciblant l'anomie des verbes dans l'aphasie post-AVC et étude de ses effets auprès d'un participant avec aphasie.* **Mémoire.** https://dumas.ccsd.cnrs.fr/dumas-03684662