Traitement automatique des langues

Cross-modal and Multimodal Natural Language Processing

sous la direction de
Gwénolé Lecorvé
John D. Kelleher

Vol. 63 - n°2 / 2022

ATALA

Cross-modal and Multimodal Natural Language Processing

ATALA
Revue de l'Association pour le Traitement Automatique des Langues

TAL
Vol.
63

n°2
2022

Cross-modal and Multimodal Natural Language Processing

ATALA

Traitement automatique des langues

Revue publiée depuis 1960 par l'Association pour le Traitement Automatique des LAngues (ATALA), avec le concours du CNRS.

©ATALA, 2022

ISSN 1965-0906

https ://www.atala.org/revuetal

Traitement automatique des langues

Comité de rédaction

Rédacteurs en chef
Cécile Fabre - CLLE, Université Toulouse 2
Emmanuel Morin - LS2N, Nantes Université
Sophie Rosset - LISN, CNRS
Pascale Sébillot - IRISA, INSA Rennes

Membres
Salah Aït-Mokhtar - Naver Labs Europe, Grenoble
Maxime Amblard - LORIA, Université Lorraine
Loïc Barrault - Meta AI
Patrice Bellot - LSIS, Aix Marseille Université
Farah Benamara - IRIT, Université Toulouse Paul Sabatier
Delphine Bernhard - LiLPa, Université de Strasbourg
Nathalie Camelin - LIUM, Université du Mans
Marie Candito - LLF, Université Paris Diderot
Vincent Claveau - IRISA, CNRS
Chloé Clavel - Télécom ParisTech
Mathieu Constant - ATILF, Université Lorraine
Géraldine Damnati - Orange Labs
Maud Ehrmann - EPFL, Suisse
Iris Eshkol - MoDyCo, Université Paris Nanterre
Dominique Estival - The MARCS Institute, University of Western Sydney, Australie
Benoît Favre - LIS, Aix-Marseille Université
Corinne Fredouille - LIA, Avignon Université
Cyril Goutte - Technologies Langagières Interactives, CNRC, Canada
Joseph Leroux - LIPN, Université Paris 13
Denis Maurel - LIFAT, Université François-Rabelais, Tours
Fabrice Maurel - GREYC, Université Caen Normandie
Adeline Nazarenko - LIPN, Université Paris 13
Aurélie Névéol - LISN, CNRS
Patrick Paroubek - LISN, CNRS
Sylvain Pogodalla - LORIA, INRIA
Fatiha Sadat - Université du Québec à Montréal, Canada
Didier Schwab - LIG, Université Grenoble Alpes
Delphine Tribout - STL, Université de Lille
François Yvon - LISN, CNRS, Université Paris-Saclay

Secrétaire
Peggy Cellier - IRISA, INSA Rennes

Traitement automatique des langues

Volume 63 – n°2 / 2022

CROSS-MODAL AND MULTIMODAL NATURAL LANGUAGE PROCESSING

Table des matières

Introduction to the Special Issue on Cross-modal and Multimodal Natural Language Processing

Gwénolé Lecorvé* — **John D. Kelleher****

* *Orange, Lannion, France*
gwenole.lecorve@orange.com
** *ADAPT Research Centre, Maynooth University, Ireland*
john.kelleher@mu.ie

ABSTRACT. *Since our communication is multimodal in terms of our ability to express ourselves via different channels and our perception of the world, the automatic production and analysis of natural language content requires the integration of these multiple modalities in order to rival human performance. However, multimodal, or cross-modal, Natural Language Processing (NLP) has long been in the minority, perhaps because it is more complex. In the wake of recent advances in artificial intelligence, which are bringing multimodality to the fore, this special issue aims to highlight, through three articles on a variety of subjects, the questions that remain, particularly with regard to data requirements, understanding the links between modalities, and the need for convergence in terms of representation and modelling.*

KEYWORDS: *Multimodality, Cross-modality, Natural Language Processing.*

TITRE. *Traitement automatique des langues intermodal et multimodal*

RÉSUMÉ. *Notre communication étant multimodale par notre capacité à nous exprimer via différents canaux et notre perception du monde, la production et l'analyse automatiques d'énoncés en langage naturel nécessitent d'intégrer ces multiples modalités pour rivaliser avec la performance de l'humain. Pourtant, le Traitement Automatique des Langues (TAL) multimodal, ou inter-modal, est longtemps resté un pan minoritaire, peut-être car plus complexe. Dans la lignée des récentes avancées en intelligence artificielle qui mettent la multimodalité sur le devant de la scène, ce numéro spécial vise à souligner, à travers trois articles aux sujets variés, les questionnements qui subsistent, notamment sur les besoins de données, la compréhension des liens entre modalités et le besoin de convergence en termes de représentation et de modélisation.*

MOTS-CLÉS : *multimodalité, intermodalité, traitement automatique des langues.*

1. Multimodality and Artifical Intelligence

Recently, Artificial Intelligence (AI) has received unprecedented media coverage. This is mostly due to the recent emergence of generative AI models. ChatGPT is currently the highest profile of these systems. However, other notable systems include Midjourney [1] and Dall-e 2 [2]—both of which are able to generate image from a text input—or MusicLM (Agostinelli *et al.*, 2023) which creates song samples from a text. Alongside these high-profile text, image and music examples progress is also being made on speech-to-text and text-to-speech. Recent models achieve performances that make the wide-scale usage of these systems feasible across a range of applications. While many of these systems are designed to transform input from one modality into another, we are also now seeing systems that can process multimodal input: for example, GPT4 can process a mixture of text and images to produce a textual response to the user (OpenAI, 2023). This shift within AI systems towards emphasising multimodal processing is also present in current debates about AI. Likewise, the recent promotion of the concept of metaverse by the major AI players testifies to this tendency to want to place natural language interactions in multimodal environments (and collect multimodal data).

The success of the generative systems, and in particular the ability of large-language models to generate fluent *English* text that is difficult to distinguish from human generated text, has brought the question of AI achieving human-like intelligence and understanding of language back to the fore in AI. A notable critic of claims attributing "understanding" to large language models is that of Bender and Koller (2020) who argue that "(linguistic) meaning" is "the relation between linguistic form and communicative intent" and, further, that "the language modelling task, because it only uses form as training data, cannot in principle lead to learning of meaning". Sahl and Carlsson (2021), however, argue that (Bender and Koller, 2020)'s critique of these AI systems is fundamentally "dualist" because it is based on distinction between form and meaning that places "understanding and meaning in a mental realm outside of language". Furthermore, they argue that even if such a distinction holds there must be a correlation between form and meaning, and that language modelling approaches can leverage this correlation to access meaning (Sahlgren and Carlsson, 2021) [3]. Consequently, Sahl and Carlsson (2021) are much more optimistic than Bender and Koller (2020) with respect to the potential for large language models, built on top of distributional representations, to be part of future natural language understanding systems (a perspective that we also share). Interestingly, Bender and Koller (2020) and Sahl and Carlsson (2021) do agree on the importance of multimodality as a future research direction for natural language understanding. Similarly, Bisk *et al.* (2020) argue that "meaning does not arise from the statistical distribution of words" and that in order to make further progress the field of Natural Language

1. `https://www.midjourney.com`.
2. `https://openai.com/product/dall-e-2`.
3. See (Kelleher and Dobnik, 2022) for more on this debate.

Processing (NLP) must move beyond training on massive mono-modal Internet-based text corpora, to consider aspects of meaning arising from perception, embodiment and social/interpersonal communication.

The argument for the importance of multimodality as the basis for "understanding" resonates with a long tradition of thought within epistemology, as Leibniz argued in the 17^{th} century "Nothing is in the intellect that was not first in the **senses**, except the intellect itself" [4]. (Note the emphasis on the plurality of the sense**s**) This is also a long history in AI. Prior to transformer models, a number of works were already interested, among others, in improving automatic speech recognition using lip movements (Bregler and Konig, 1994) or biological signals (Jou *et al.*, 2006); facilitating the use of a software interface for users by combining speech and gestures (Oviatt *et al.*, 2000); analysing TV streams by merging video and audio (speech or not) information (Duan *et al.*, 2006; Giraudel *et al.*, 2012); or producing shared semantic representations from texts and images (Bruni *et al.*, 2014).

Nowadays, recent progress in deep learning is eroding two of the most difficult challenges for the development of multimodal systems. First, the challenge of how to develop and learn multimodal representations is addressed via the representation learning of vector spaces enabled by deep learning (Kelleher, 2019). Second, learning paradigms like self-supervision, reinforcement learning or adversarial learning ease the difficulties associated with the need for massive annotated data in the supervised learning paradigm. This shift towards self-supervised learning has also led to the emergence of *foundation models* that exhibit emergent capabilities and fast adaptation to downstream tasks (Yang *et al.*, 2022), and that can be used as elementary building blocks for the construction of more complex multimodal systems (Shen *et al.*, 2023). Interestingly, the adaptation of these models via the specification of downstream tasks is also becoming easier thanks to the ability of large models to be driven by textual instructions.

2. Natural Language Processing and Multimodality

The concepts of multimodality and natural language can overlap in two different ways. In the first case, (written) natural language is used as a strategic pivot towards, from or with which other modalities interconnect. The reason is that texts are a useful and natural way to describe concepts and denote entities that are essentially multimodal (e.g., description of an image, an event, a building, etc.), and trigger natural reasoning on these objects. Such approaches can be referred to as *cross-modal* NLP. Then, in the second *multimodal* case, natural language is an ingredient of multimodal objects. This can be because natural language is not limited to the written modality but often also includes and interacts with many others (Bezemer and Jewitt, 2018; Holler and Levinson, 2019; Cohn and Schilperoord, 2022). For instance, a message can be conveyed through audio (speech) or gestures and facial expressions (sign language or

4. *"Nihil est in intellectu quod non fuerit in sensu, nisi intellectu ipse."*, (Leibniz, 1765).

completed speech). It may also be accompanied by social attitudes and non-verbal dimensions, including signs of affect, spontaneity, pathology, co-adaptation with dialogue participants, etc. Alternatively, natural language can be part of a larger object, e.g., songs or movies. In these contexts, processing natural language requires the integration of language with the whole multimodal context. NLP is thus a joint processing of multiple information channels.

Given these definitions, multimodal and cross-modal NLP covers a very large range of tasks and fields, among which:

– multimodal dialogue, multimodal question-answering;

– sign language, completed spoken language;

– speech recognition and synthesis in multimodal contexts;

– synthesis of animated emotional agents;

– handwriting recognition and analysis of handwritten documents;

– understanding, translation and summarisation of multimodal documents;

– indexing, search and mining of multimedia and/or multimodal documents;

– biological signal processing, computational psychology or sociology, for NLP;

– inter-/multimodal human-computer interface for NLP;

– other multimodal or inter-modal applications (image captioning, image-to-text generation, generation/analysis of songs and lyrics, etc.).

Unfortunately, most of the research carried out in these areas are spread over different specific communities, conferences and journals where one modality is dominant. This makes it more difficult to exchange ideas and slows down the development of interdisciplinary work. The objective of this special issue of the TAL journal is to promote NLP in multimodal contexts (several modalities contribute to the resolution of a problem) or cross-modal contexts (the transformation from one modality to another).

3. Accepted Papers

It was important for the journal to highlight the specificities (benefits, difficulties, perspectives, etc.) linked to cross- or multimodality, as well as the challenges posed by multimodality, like understanding the interactions between modalities, the harmonisation or compatibility of representations, the development of joint models or transfer from one modality to another, or the constitution (or even annotation) of multimodal resources. As detailed below, the papers accepted to this special issue clearly contribute to these research directions.

A Dataset to Answer Visual Questions about Named Entities (*Un jeu de données pour répondre à des questions visuelles à propos d'entités nommées*) by Paul Lerner, Salem Messoud, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, Jose G. Moreno, and Jesús Lovón Melgarejo: The first paper presents a new dataset, named ViQuAE, dedicated to Knowledge-based Visual Ques-

tion Answering about named Entities (KVQAE). This dataset consists of 3,700 questions paired with images, annotated using a semi-automatic method. It includes a wide range of entity types, associated with a knowledge base composed of Wikipedia articles paired with images. The paper presents the dataset, along with baseline models for the KVQAE task (decomposed into a pipeline of three sub-tasks) and an analysis of what each modality (text and images) brings.

Neural Speech Synthesis Techniques for Non-Dedicated Training Data: Amateur Audiobooks in French (*Techniques de synthèse vocale neuronale à l'épreuve des données d'apprentissage non dédiées : les livres audio amateurs en français*) by Aghilas Sini, Lily Wadoux, Antoine Perquin, Gaëlle Vidal, David Guennec, Damien Lolive, Pierre Alain, Nelly Barbot, Jonathan Chevelu, and Arnaud Delhay: The second paper reports on how non-professional speech data can be used to perform neural text-to-speech. This topic is of particular interest because the usual training data for such systems rely on very clean data, with high quality and consistency, which limits the development of new voices. To overcome these challenges the authors collect non-professional data, feed it to three different speech synthesis techniques, namely single-speaker speech synthesis, voice cloning and voice conversion, and discuss the impact for each of them.

Signing Avatar – Synthesis of French Sign Language from Text (*Avatar signeur – Synthèse de la langue des signes française à partir de texte*) by Sylvie Gibet: The last paper focuses on French sign language (LSF) and presents a system that translates text to LSF by means of a 3D avatar. The paper first carefully presents the peculiarities of sign languages, introducing the key concepts and their analogy with those of the usual linguistics. Then, the author details her text-to-LSF system and its evolution along the years, from its original form (based on the composition of multichannel information) to the most recent extensions (e.g., facial animation, hands movements, etc.). Finally, the paper sets out some of the remaining difficulties, both in terms of linguistic, animation and deep learning models.

Acknowledgment

We thank the editorial committee of the TAL journal for promoting multimodality in NLP and inviting us to coordinate this special issue. In particular, thanks to Pascale Sébillot who, as editor-in-chief of the journal, guided us through the various stages until the publication of this issue. We finally also thank the reviewers and members of the scientific committee who agreed to join us for this special issue and who gave their time to help us select the articles (in alphabetical order): Loïc Barrault (Meta, France), Marion Blondel (CNRS, France), Quentin Brabant (Orange, France), Géraldine Damnati (Orange, France), Florence Encrevé (Université Paris 8, France), Antoine Gourru (Université de Saint-Étienne, France), Benjamin Lecouteux (Université Grenoble Alpes, France), Fabrice Maurel (Université de Caen Normandie, France), Slim Ouni (Université de Lorraine, France), Olivier Perrotin (CNRS, France), Jérémie Segouat (Université Toulouse, France), François Yvon (Université Paris-Saclay, France).

4. References

Agostinelli A., Denk T., Borsos Z., Engel J., Verzetti M., Caillon A., Huang Q., Jansen A., Roberts A., Tagliasacchi M., Sharifi M., Zeghidour N., Frank C., "MusicLM: Generating Music From Text", *arXiv:2301.11325*, 2023.

Bender E. M., Koller A., "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data", *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Bezemer J., Jewitt C., "Multimodality: A Guide for Linguists", *Research methods in linguistics*, 2018.

Bisk Y., Holtzman A., Thomason J., Andreas J., Bengio Y., Chai J., Lapata M., Lazaridou A., May J., Nisnevich A. *et al.*, "Experience Grounds Language", *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

Bregler C., Konig Y., ""Eigenlips" for Robust Speech Recognition", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1994.

Bruni E., Tran N.-K., Baroni M., "Multimodal Distributional Semantics", *Journal of artificial intelligence research (JAIR)*, 2014.

Cohn N., Schilperoord J., "Reimagining Language", *Cognitive Science*, 2022.

Duan L.-Y., Wang J., Zheng Y., Jin J. S., Lu H., Xu C., "Segmentation, Categorization, and Identification of Commercial Clips from TV Streams using Multimodal Analysis", *Proceedings of the ACM International Conference on Multimedia*, 2006.

Giraudel A., Carré M., Mapelli V., Kahn J., Galibert O., Quintard L., "The REPERE Corpus : a Multimodal Corpus for Person Recognition", *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 2012.

Holler J., Levinson S. C., "Multimodal Language Processing in Human Communication", *Trends in Cognitive Sciences*, 2019.

Jou S.-C., Schultz T., Walliczek M., Kraft F., Waibel A., "Towards Continuous Speech Recognition using Surface Electromyography", *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, 2006.

Kelleher J. D., *Deep learning*, MIT Press, 2019.

Kelleher J. D., Dobnik S., "Distributional Semantics for Situated Spatial Language? Functional, Geometric and Perceptual perspectives", *Probabilistic Approaches to Linguistic Theory*, CSLI Publications, 2022.

Leibniz G. W., *Nouveaux essais sur l'entendement (New essays on human understanding)*, 1765.

OpenAI, "GPT-4 Technical Report", *arXiv:2303.08774*, 2023.

Oviatt S., Cohen P., Wu L., Duncan L., Suhm B., Bers J., Holzman T., Winograd T., Landay J., Larson J. *et al.*, "Designing the User Interface for Multimodal Speech and Pen-based Gesture Applications: State-of-the-Art Systems and Future Research Directions", *Human-Computer Interaction*, 2000.

Sahlgren M., Carlsson F., "The Singleton Fallacy: Why Current Critiques of Language Models Miss the Point", *Frontiers in Artificial Intelligence*, 2021.

Shen Y., Song K., Tan X., Li D., Lu W., Zhuang Y., "HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in HuggingFace", *arXiv:2303.17580*, 2023.

Yang M. S., Du Y., Parker-Holder J., Karamcheti S., Mordatch I., Gu S. S., Nachum O. (eds), *Workshop on Foundation Models for Decision Making*, Neural Information Processing Systems, 2022.

Un jeu de données pour répondre à des questions visuelles à propos d'entités nommées

Paul Lerner* — **Salem Messoud*** — **Olivier Ferret**** — **Camille Guinaudeau*** — **Hervé Le Borgne**** — **Romaric Besançon**** — **Jose G. Moreno***** — **Jesús Lovón Melgarejo*****

* *Université Paris-Saclay, CNRS, LISN, 91400, Orsay, France*
** *Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France*
*** *IRIT, UMR 5505 CNRS, Université Paul Sabatier, Toulouse, France*
paul.lerner@lisn.upsaclay.fr, olivier.ferret@cea.fr

RÉSUMÉ. *Dans le contexte des approches multimodales, nous nous intéressons à la tâche de réponse à des questions visuelles à propos d'entités nommées en utilisant des bases de connaissances (KVQAE). Nous mettons à disposition ViQuAE, un nouveau jeu de données de 3 700 questions associées à des images, annoté à l'aide d'une méthode semi-automatique. C'est le premier jeu de données de KVQAE comprenant des types d'entités variés associé à une base de connaissances composée de 1,5 million d'articles Wikipédia, incluant textes et images. Nous proposons également un modèle de référence de KVQAE en trois étapes : recherche d'information initiale, réordonnancement, puis extraction des réponses. Les résultats de nos expériences démontrent empiriquement la difficulté de la tâche et ouvrent la voie à une meilleure représentation multimodale des entités nommées.*

MOTS-CLÉS : *jeu de données, question-réponse visuelle, bases de connaissances, multimodalité.*

ABSTRACT. *In the context of multimodal processing, we focus our work on Knowledge-based Visual Question Answering about named Entities (KVQAE). We provide ViQuAE, a novel dataset of 3,700 questions paired with images, annotated using a semi-automatic method. It is the first KVQAE dataset to cover a wide range of entity types, associated with a knowledge base composed of 1.5M Wikipedia articles paired with images. To set a baseline on the benchmark, we address KVQAE as a three-stage problem: initial Information Retrieval, Re-Ranking, and Reading Comprehension. The experiments empirically demonstrate the difficulty of the task and pave the way towards better multimodal entity representations.*

KEYWORDS: *Dataset, Knowledge-based Visual Question Answering, Multimodality.*

Requête (entrée)		Article pertinent dans la base de connaissances	
	*« Which constituency did this man represent when he was Prime Minister ? »*		« Macmillan indeed lost Stockton in the landslide Labour victory of 1945, but returned to Parliament in the November 1945 by-election in **Bromley**. »
	*« In which year did this ocean liner make her maiden voyage ? »*		« Queen Elizabeth 2, often referred to simply as QE2, is a floating hotel and retired ocean liner built for the Cunard Line which was operated by Cunard as both a transatlantic liner and a cruise ship from **1969** to 2008. »

FIGURE 1. *Exemple de questions du jeu de données ViQuAE avec leur image contextuelle et la source de la réponse (issue de la base de connaissances)*

1. Introduction

La fusion de modalités telles que l'image et le texte pour rechercher des informations est un problème reconnu comme difficile du fait de la différence de niveau de leurs sémantiques respectives (Srihari *et al.*, 2000). Ce constat est particulièrement vrai pour répondre à des questions visuelles à propos d'entités nommées (KVQAE[1]), où différents types de relations peuvent lier une question et l'image qui lui est associée en tant que contexte (cf. figure 1).

Dans la tâche classique de réponse à des questions visuelles (VQA), le contenu de l'image associée, par exemple la couleur d'un objet ou le nombre d'objets, est le sujet de la question (Antol *et al.*, 2015). La VQA fondée sur les connaissances (Wang *et al.*, 2017 ; Wang *et al.*, 2018 ; Marino *et al.*, 2019) utilise quant à elle l'image comme contexte pour poser des questions et trouver des réponses dans des bases de connaissances (BC), structurées ou non. Cependant, ces deux champs de recherche se focalisent principalement sur des catégories d'objets génériques en s'appuyant sur un prétraitement de détection d'objets (Anderson *et al.*, 2018 ; Gardères *et al.*, 2020). Dans cette optique, la seconde question de la figure 1 pourrait typiquement porter sur le type de bateau en prenant la forme : « Est-ce un bateau de pêche ? » Au contraire, notre travail se concentre sur des questions nécessitant des connaissances à propos des entités nommées, comme le *Queen Elizabeth 2* dans le cas présent. Nous avons conçu et publions le jeu de données ViQuAE dans ce but[2]. Notre jeu de données a été conçu comme un cadre d'évaluation pour diagnostiquer et suivre les progrès des systèmes de KVQAE. Nous pensons en effet que la KVQAE est une tâche bien définie et facilement évaluable. Elle est ainsi bien appropriée pour rendre compte des progrès de

1. Pour le sigle anglais de *Knowledge-based Visual Question Answering about named Entities*.
2. Disponible *via* `https://github.com/PaulLerner/ViQuAE`.

(a) Personne (Louis-Philippe I^{er})

(b) Personne (Louis-Philippe I^{er}) ou tableau (de Franz Xaver Winterhalter)

(c) Attraction touristique artificielle (tour Eiffel)

(d) Attraction touristique naturelle (mont Saint Helens)

(e) Entreprise (Apple) ou bâtiment (*Apple Fifth Avenue*)

(f) Entreprise (Apple) ou produit (iPhone 1)

FIGURE 2. *Quelques exemples d'images de différents types d'entités et différentes images du même type d'entité considérées dans notre travail*

la qualité des représentations multimodales d'entités nommées sur un plan plus général. La représentation multimodale des entités ouvre aussi la voie vers différentes applications, permettant par exemple de rendre les interactions homme-machine plus naturelles : en regardant un film, on peut se demander « Où ai-je déjà vu cette actrice ? » ou « Est-ce qu'elle a déjà gagné un Oscar ? ». Les questions sur les entités nommées sont très difficiles car les BC actuelles en contiennent des millions. De ce point de vue, utiliser chaque modalité indépendamment n'est pas suffisamment discriminant pour répondre au besoin de l'utilisateur. À titre d'exemple, dans les images de la figure 1, il est assez complexe de reconnaître *Harold Macmillan* au sein d'une BC contenant des millions de *personnes*. Cependant, on peut déduire de la question qu'il était *Premier ministre* et réduire ainsi les candidats à quelques centaines.

Shah *et al.* (2019) ont déjà travaillé sur la KVQAE mais se sont limités aux entités nommées de type personne. Au contraire, ViQuAE comprend divers types d'entités. Cette diversité est une question centrale dans la KVQAE, notamment en raison de l'hétérogénéité des représentations visuelles qui en résulte. Entre autres entités, les entreprises peuvent être ainsi représentées par un bâtiment (par exemple leur siège), un produit manufacturé qu'elles vendent ou simplement leur logo (cf. figure 2). La KVQAE nécessite donc une représentation multimodale des connaissances, ce qui la distingue clairement de la recherche d'image par le contenu. Cette diversité implique également

la nécessité d'étudier d'autres types d'entités que les personnes, qui peuvent assez bien être reconnues visuellement à partir de leur seul visage. Par ailleurs, Shah *et al.* (2019) utilisent une BC structurée et donc des méthodes assez différentes des nôtres, qui avons opté pour une BC multimodale constituée de textes non structurés et d'images (cf. section 3.4).

Sur un autre plan, ViQuAE, avec ses 3 700 questions, s'inscrit dans le courant des travaux sur l'apprentissage sans (*zero-shot)* ou avec peu d'exemples (*few-shot*), avec une double idée : d'une part, la diversité des tâches unissant texte et image ne permet pas de développer des jeux de données d'une taille suffisante pour entraîner de gros modèles à partir de zéro ; d'autre part, les percées des travaux reposant sur les *Foundation Models* (Bommasani et *et al.*, 2021) permettent de s'affranchir d'un tel entraînement. Nous espérons ainsi que ViQuAE encouragera les études vers des modèles transférables ou vers des techniques d'apprentissage sans ou avec peu d'exemples, nécessaires pour la KVQAE.

Cet article est une version étendue de Lerner *et al.* (2022b)[3]. En plus de présenter le jeu de données ViQuAE (section 3), cette version plus détaillée contient une revue des travaux connexes actualisée (section 2), des analyses supplémentaires (sections 5 et 7) ainsi qu'une nouvelle contribution concernant le réordonnancement des résultats de la recherche d'information initiale (section 6 et par conséquent, section 7).

2. Travaux connexes

La KVQAE est intrinsèquement une tâche complexe mêlant à la fois des problématiques de recherche d'information (RI) et d'extraction d'information en faisant intervenir plusieurs médias tout à la fois au niveau de la requête et des documents cibles. Elle se retrouve donc à l'interface de plusieurs domaines. Le fait d'utiliser le texte comme source de réponse dans notre cas la rapproche d'abord des systèmes de question-réponse (QA) textuels se situant dans la lignée de Voorhees et Tice (2000) et traitant la tâche en deux étapes, avec une phase initiale de RI suivie d'une extraction de la réponse (*reading comprehension*). Au cours des dernières années, une attention particulière a été accordée à l'extraction de la réponse, avec des jeux de données de plus en plus grands (Rajpurkar *et al.*, 2016 ; Joshi *et al.*, 2017 ; Kwiatkowski *et al.*, 2019). Nous profitons de ces derniers pour bâtir notre propre jeu de données, comme expliqué à la section suivante, avec le même tropisme pour les questions factuelles que la plupart des travaux dans le domaine (Chen *et al.*, 2017).

De son côté, bien qu'initialement axée sur le texte, la RI s'est rapidement étendue aux documents multimodaux. Srihari *et al.* (2000) et Clough *et al.* (2004), par exemple, partageaient déjà un certain nombre de problèmes avec la KVQAE, comme la fusion d'informations multimodales. Cependant, les modalités en RI multimodale sont souvent redondantes alors qu'elles sont complémentaires avec la KVQAE.

3. Également résumé et traduit en français dans Lerner *et al.* (2022a).

L'usage de plusieurs modalités en QA prend quant à elle souvent une forme cross-modale (Kembhavi *et al.*, 2017 ; Sampat *et al.*, 2020 ; Talmor *et al.*, 2021 ; Chang *et al.*, 2022 ; Reddy *et al.*, 2021) assimilable à de l'extraction de réponse par le biais de plusieurs modalités (texte, tableaux ou images). La source de la réponse, quelle que soit la modalité, est fournie en même temps que la question contextuelle et les deux sont interdépendantes. Ainsi, Reddy *et al.* (2021) construisent leur corpus à partir d'articles de presse, où le système a accès aux métadonnées des images, telles que leur légende. Par conséquent, la tâche relève davantage du raisonnement logique que de la RI, contrairement aux questions de KVQAE, qui sont autosuffisantes.

Pour sa part, la VQA fondée sur la connaissance (Wang *et al.*, 2017 ; Wang *et al.*, 2018 ; Marino *et al.*, 2019 ; Jain *et al.*, 2021 ; Schwenk *et al.*, 2022) se concentre sur des questions de sens commun concernant des catégories d'objets génériques. En outre, les jeux de données VQA (fondés sur la connaissance ou pas) sont généralement construits à partir des images du jeu de données *Common Objects in Context* (COCO, Lin *et al.* (2014)). Pour ces deux raisons, la VQA fondée sur la connaissance a été largement traitée en s'appuyant sur des détecteurs d'objets entraînés sur COCO, ce qui facilite la RI (Gardères *et al.*, 2020).

Le premier jeu de données KVQAE a quant à lui été présenté par Shah *et al.* (2019) : il s'agit de KVQA, fondé sur Wikidata et focalisé sur les entités de type personne. Malgré sa grande taille, ce jeu de données présente plusieurs limites : (i) il est restreint aux entités de type personne, avec une RI se réduisant à la reconnaissance faciale ; (ii) les questions sont générées automatiquement à partir de patrons et de Wikidata. De ce fait, elles sont assez répétitives et limitées par le schéma de Wikidata : la plupart des questions portent sur l'identité de la personne, son lieu de naissance, sa date de naissance ou son emploi. Au contraire, nous visons à construire un jeu de données couvrant divers types d'entités avec une expression riche et des questions couvrant de nombreux sujets.

3. Jeu de données et base de connaissances ViQuAE

3.1. *Annotation automatique*

Pour limiter les efforts d'annotation manuelle, nous nous sommes appuyés sur des jeux de données de question-réponse existants, qui comprennent des questions couvrant différents sujets et entités. Nous avons ainsi décidé d'utiliser le jeu de données textuel TriviaQA en raison de sa taille et de la typologie de ses questions (Joshi *et al.*, 2017). L'idée principale de notre processus est de remplacer la mention de l'entité dans la question par une représentation visuelle de l'entité. Celle-ci est alors référencée par une mention ambiguë (par exemple « cet homme »). De cette façon, il n'est pas possible de répondre à la question sans s'appuyer sur l'image contextuelle. Dans le premier exemple de la figure 1, la mention de l'entité nommée « *Harold Macmillan* » de la question originale est ainsi remplacée par la mention ambiguë « *this man* ».

Notre processus débute par une analyse syntaxique et une identification des entités nommées dans les questions à l'aide de `spaCy`[4], ce qui permet d'obtenir environ 0,9 mention valide par question. L'analyse des dépendances permet de ne conserver que certaines mentions d'entités, par exemple le sujet de la question. À partir de ces mentions d'entité, puisque la réponse à la question est connue, la désambiguïsation peut être effectuée en vérifiant si la réponse est présente dans l'article Wikipédia de l'entité candidate. Cette étape a en fait été réalisée par Joshi *et al.* (2017) avec TAGME (Ferragina et Scaiella, 2010) lorsqu'ils ont initialement conçu TriviaQA pour l'extraction de réponse ; nous avons simplement fait correspondre nos mentions d'entités avec leurs entités désambiguïsées. Environ 55 % des mentions d'entités (donc de questions potentielles) ont été désambiguïsées, laissant 45 % de côté. Wikidata permet de recueillir des informations sur les entités désambiguïsées : leur type, leur profession, leur genre et leur catégorie Commons. Nous avons utilisé cette dernière pour trouver une image pertinente tandis que les autres sont nécessaires pour générer une mention ambiguë. Les personnes sont référencées par leur profession (par exemple « cet écrivain ») et les autres entités par leur type (par exemple « cette attraction touristique »). De plus, si le genre était disponible, nous avons également utilisé « *this man/woman* » et « *he-him-his/she-her-hers* » selon la dépendance syntaxique de la mention originale. Étant donné que certaines entités abstraites, telles que les pays ou les nationalités, sont souvent mentionnées dans les questions mais ne sont pas pertinentes pour la KVQAE, le type d'entité est restreint à une liste de types et de sous-types construite manuellement, disponible avec le jeu de données. De plus, pour se conformer à la RGPD[5], et étant donné que beaucoup de questions portent sur des personnes, nous avons conservé seulement les questions portant sur des personnes décédées. Cette étape écarte 31 % de questions supplémentaires. Les images sont récupérées à partir de la catégorie Commons de l'entité. 3 % des questions n'avaient pas d'images disponibles et ont donc été écartées. Grâce aux contributeurs de Wikimedia Commons, toutes les images du jeu de données sont soit sous licence libre[6], soit dans le domaine public, ce qui nous permet de les redistribuer pour assurer la reproductibilité de notre travail. Nous décrivons comment filtrer cette annotation automatique dans la section suivante.

3.2. *Annotation manuelle*

L'annotation automatique décrite ci-dessus présente quelques inconvénients. Les deux principales sources d'erreurs sont : (i) l'image sélectionnée, qui peut être inappropriée ; (ii) la trop grande spécificité de la question, qui permet parfois de répondre sans avoir besoin de l'image[7]. Pour remédier à ce problème, une interface d'annotation

4. `https://spacy.io/`
5. `https://eur-lex.europa.eu/eli/reg/2016/679/oj`
6. `https://freedomdefined.org/Definition`
7. Par exemple, « *Which constituency did this man represent when he was Prime Minister,* ***succeeding Sir Edward Campbell*** *?* » contient trop d'informations. Il faudrait la reformuler pour retrouver l'exemple de la figure 1.

a été conçue à l'aide de Label Studio[8]. L'annotateur peut reformuler librement la question (des mentions complémentaires sont suggérées) tant que la réponse n'est pas modifiée. Il doit également choisir parmi huit images candidates, si celle sélectionnée n'est pas appropriée, en utilisant comme référence l'image de référence de la BC (cf. section 3.4) tout en s'assurant qu'il ne s'agit pas d'un quasi-doublon. En dernier recours, l'annotateur peut simplement rejeter la question. L'interface et les instructions d'annotation sont partagées avec le reste de notre code.

Cette annotation manuelle a été réalisée par sept annotateurs internes (les auteurs de l'article, à l'exception de Salem Messoud). L'interface a permis de traiter environ 120 questions par heure. La proportion de questions à propos de personnes a été équilibrée pour assurer la diversité du jeu de données. Nous avons annoté 5 700 questions générées, dont 2 000 ont été écartées, principalement parce qu'elles étaient surspécifiées ou que l'image n'était pas pertinente. Finalement, le jeu de données ViQuAE est constitué de 3 700 questions, réparties aléatoirement en ensembles de taille égale pour l'entraînement, la validation et le test, sans recouvrement entre les images. La majorité (55 %) des questions valides ont été éditées par les annotateurs, avec une distance de Levenshtein moyenne de cinq mots entre la version initiale et la version éditée.

Pour mesurer l'accord inter-annotateur, un sous-ensemble de 103 questions ont été annotées par au moins 3 annotateurs différents. L'accord a ensuite été calculé en utilisant le Kappa de Fleiss (Fleiss, 1971). Les annotateurs se sont mis d'accord pour rejeter ou non la question avec $\kappa = 0{,}33$, montrant un accord léger. En effet, déterminer si une question est surspécifiée ou non peut être assez subjectif[9]. De plus, la reformulation de certaines questions surspécifiées peut être subtile. Cependant, il faut rappeler que, dans notre cas, les désaccords entre annotateurs ne concernent pas la *réponse* à la question mais seulement le filtrage du jeu de données généré automatiquement puisque les questions et les réponses sont définies dans TriviaQA et qu'un annotateur *ne peut pas changer la réponse*.

3.3. *Analyse des données*

Le jeu de données ViQuAE se compose de 3 700 questions contextualisées par 3 300 images uniques, dont deux exemples sont présentés en figure 1. Les questions comportent en moyenne 12 mots, pour un vocabulaire de 4 700 mots. Sur les 3 700 réponses, les plus fréquentes, « France » et « Turquie », n'apparaissent que 13 fois, soit 0,3 % du total, ce qui montre la quasi-absence de biais *a priori* sur les réponses pour un classifieur indépendant de la question. De plus, il n'y a qu'un chevauchement de 25 % des réponses et de 18 % des entités entre les ensembles d'entraînement et de test. Ces trois points soulignent la différence entre la KVQAE et la VQA (fondée sur la connaissance ou pas) et démontrent que traiter la KVQAE comme une tâche de classification serait inefficace.

8. `https://labelstud.io/`

9. Par exemple « This inner planet and which other planet in our solar system has no moon ? »

	ViQuAE	KVQA
# Questions	3 700	183 000
# Questions par image	1,1	7,4
Vocabulaire	4 700	600*
Taille moyenne des questions	12,4	10,1
Réponse la plus probable *a priori*	0,3 %	15,9 %
Chevauchement entre les réponses	25,3 %	89,4 %
Chevauchement entre les entités	18,1 %	40,6 %
# Questions par entité	1,5	9,7
# Types d'entités	980	1

TABLEAU 1. *Statistiques du jeu de données par rapport à KVQA (Shah* et al.*, 2019). **En moyenne, sur 49 sous-ensembles aléatoires de la même taille que ViQuAE, le vocabulaire du jeu de données KVQA entier se compose de 8 400 tokens.*

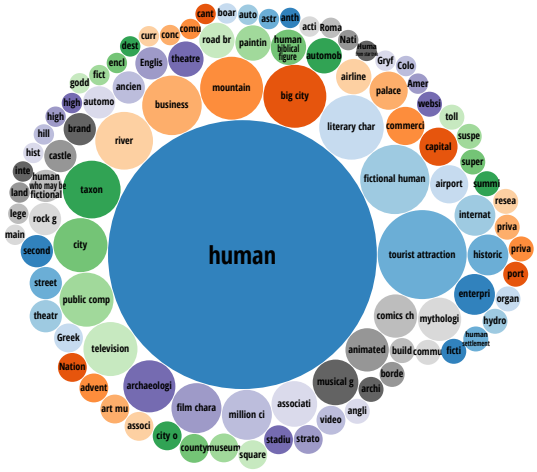

FIGURE 3. *Les 100 types d'entités (non-exclusifs) les plus fréquents dans le jeu de données*

Une contribution importante du jeu de données est sa diversité de types d'entités, qui est l'un des principaux défis pour les représentations multimodales (cf. section 1). ViQuAE comprend près de mille types d'entités différents (issus de l'ontologie Wikidata) parmi ses 2 400 entités uniques. Toutefois, ces types ne sont pas exclusifs : une moyenne de 1,6 type est attribuée à chaque entité. Le jeu de données comporte 43 % de personnes, sans prendre en compte d'autres entités semblables, comme les personnages fictifs ou mythologiques, ou des groupes de personnes, par exemple les groupes de musique. La figure 3 présente un graphique à bulles des 100 types d'entités les plus

fréquents. Un résumé des statistiques comparées avec le jeu de données KVQA de Shah *et al.* (2019) est rapporté dans le tableau 1. Nous pouvons constater que, malgré sa petite taille, ViQuAE est plus diversifié sous certains aspects.

Cependant, le jeu de données ViQuAE présente aussi certaines limites. L'un des inconvénients de notre processus d'annotation, et plus précisément de la désambiguïsation des entités nommées, est que les réponses sont systématiquement présentes dans la page Wikipédia de l'entité. Ainsi, les questions sont *mono-hop* au niveau de l'article. Bien sûr, la question peut toujours nécessiter un raisonnement sur plusieurs phrases ou paragraphes de l'article. En revanche, Shah *et al.* (2019) comprennent plusieurs questions *multi-hop* qui, même si elles ne semblent pas très naturelles, permettent d'évaluer les capacités de raisonnement du modèle.

Toujours dans la thématique du multi-hop, Shah *et al.* (2019) comprennent des images avec plusieurs personnes où les questions contiennent alors des expressions référentielles (par exemple « la personne sur la droite »). Dans le cadre de l'annotation automatique, nous avons au contraire visé à avoir une seule entité représentée de manière proéminente par image. Dans le cas où une telle image n'existait pas, l'annotateur a pu utiliser une expression référentielle en reformulant la question mais cela reste marginal par rapport à Shah *et al.* (2019).

3.4. *La base de connaissances ViQuAE*

La BC ViQuAE est construite à partir de la sauvegarde du 01/08/2019 de Wikipédia, disponible dans KILT (Petroni *et al.*, 2021) et comprenant 5,9 millions d'articles. Chacun d'eux est associé à une entité Wikidata. Pour obtenir une représentation visuelle de l'entité, une image unique est extraite de Wikidata, dans l'ordre suivant de préférence des propriétés Wikidata : (i) P18 « image » ; (ii) P154 « image du logotype » ; (iii) P41 « image du drapeau » ; (iv) P94 « image du blason » ; (v) P2425 « ruban de médaille ». Les articles sans image sont écartés, ce qui aboutit à une BC de 1,5 million d'articles, dont 542 000 à propos de personnes, chacun associé à une image. La BC obtenue est donc cent fois plus grande que celle des expériences de Shah *et al.* (2019). 95 % des images de la base de connaissances sont uniques.

4. Approche de base pour la KVQAE et expérimentations

Nous traitons le problème de la KVQAE en trois étapes successives : recherche d'information initiale (cf. section 5), réordonnancement (cf. section 6) et extraction des réponses (*reading comprehension* ; cf. section 7), avec des métriques d'évaluation dédiées pour chacune. Nous reprenons ainsi une décomposition classiquement adoptée en QA textuelle (Chen *et al.*, 2017 ; Karpukhin *et al.*, 2020).

L'évaluation de cette approche de base a été réalisée sur notre jeu de données ViQuAE. Plus précisément, l'évaluation finale de la tâche est toujours effectuée sur l'ensemble de test de 1 257 questions tandis que les hyperparamètres sont optimisés

sur l'ensemble de validation de 1 250 questions et uniquement pour les modèles *few-shot*, l'apprentissage est effectué sur l'ensemble d'entraînement de 1 190 questions. Conformément à Joshi *et al.* (2017), les alias Wikipédia d'une réponse donnée sont considérés comme des réponses valides.

Nous n'avons pas expérimenté notre approche sur KVQA (Shah *et al.*, 2019) puisque, ce jeu de données ayant été généré automatiquement à partir de Wikidata, rien ne garantit que les réponses se trouvent dans notre BC[10]. De plus, il comprend 29 % de questions booléennes (réponse oui/non) pour lesquelles on ne peut pas évaluer la pertinence du passage de texte/de l'article Wikipédia automatiquement.

Bien que certains détails soient omis dans cette section et les suivantes en raison des contraintes d'espace, toutes les expériences peuvent être reproduites en utilisant notre code[11].

5. Recherche d'information initiale

La recherche d'information initiale a pour but de filtrer la base de connaissances afin d'obtenir des passages de texte candidats pertinents par rapport à la requête (question et image). Contrairement au réordonnancement (cf. section suivante), la RI initiale est contrainte du point de vue calculatoire par la grande taille de la BC.

Nous adoptons une approche de fusion tardive au niveau des modalités : la recherche est effectuée indépendamment avec la question et l'image puis les résultats sont fusionnés au niveau des scores. Notre implémentation s'appuie sur Elasticsearch[12] et Faiss (Johnson *et al.*, 2019), respectivement pour la recherche parcimonieuse et dense, toutes deux *via* la bibliothèque Datasets de Hugging Face (Lhoest et *et al.*, 2021).

5.1. *Recherche de texte initiale*

En amont de la recherche, nous filtrons les données semi-structurées des articles, comme les tableaux et les listes (Karpukhin *et al.*, 2020 ; Wang *et al.*, 2019). Chaque article est ensuite divisé en passages disjoints de 100 mots tout en préservant les limites des phrases, ce qui produit 12 millions de passages (environ 8 passages par article). Le titre de l'article est concaténé au début de chaque passage. Comme modèle *zero-shot*[13], nous utilisons BM25 (Robertson *et al.*, 1995) et optimisons ses hyperparamètres sur l'ensemble de validation en utilisant une recherche par dichotomie. Pour

10. On peut estimer grossièrement que 37 % des questions (hors booléennes) de KVQA n'ont pas de réponse dans notre BC en vérifiant si la réponse est incluse dans l'article de l'entité-sujet.
11. `https://github.com/PaulLerner/ViQuAE`
12. `https://www.elastic.co/`
13. La notion de *zero-shot* renvoie ici au fait qu'il n'y a pas d'optimisation des paramètres d'un modèle mais seulement de ses hyperparamètres.

définir également une référence *few-shot*, nous utilisons DPR (Karpukhin *et al.*, 2020). DPR est un modèle de recherche dense fondé sur deux modèles BERT (Devlin *et al.*, 2019) : un pour la question et un pour le passage. DPR est entraîné à minimiser l'entropie croisée des similarités entre les questions et les passages (avec un seul passage pertinent par question). La sélection des passages négatifs utilisés lors de l'entraînement est faite à l'aide de BM25 afin de garantir sa difficulté et sa qualité. DPR est préentraîné sur TriviaQA, filtré de toutes les questions utilisées dans ViQuAE, avant d'être ajusté sur ViQuAE. Nous considérons également le modèle sans ajustement, entraîné uniquement sur TriviaQA, comme une autre référence *zero-shot*. La validation est effectuée sur les questions TriviaQA utilisées pour générer l'ensemble de validation ViQuAE. Pour l'entraînement, nous utilisons les mêmes hyperparamètres que Karpukhin *et al.* (2020).

5.2. *Recherche d'image initiale*

Pour la recherche d'images, nous utilisons deux représentations différentes de manière alternative : ArcFace (Deng *et al.*, 2019) pour les visages, si au moins un visage est détecté ; ImageNet-ResNet (He *et al.*, 2016) et CLIP (Radford *et al.*, 2021) pour l'image complète. Par conséquent, la BC est divisée en deux parties : les personnes avec un visage détecté et les non-personnes, en faisant l'hypothèse que les visages ne sont pertinents que pour les personnes. Comme Deng *et al.* (2019), nous utilisons MTCNN (Zhang *et al.*, 2016) pour la détection des visages. Les cinq points de repère du visage (les yeux, le nez et les coins de la bouche) sont adoptés pour effectuer une transformation de similarité afin qu'ils soient toujours à la même position dans l'image, quelle que soit la pose originale de la personne. Si plusieurs visages sont détectés, seul celui associé à la plus forte probabilité est conservé. 6,6 % des personnes de la BC n'ont pas de visage détecté et ont donc été écartées.

ArcFace est une méthode d'apprentissage de représentation pour la reconnaissance et la vérification des visages très efficace. Il est préentraîné sur MS-Celeb (Guo *et al.*, 2016), composé de photos de célébrités. Ses entités ont un certain chevauchement avec ViQuAE, qui est analysé dans la section suivante. Cette approche est assez comparable à celle utilisée par Shah *et al.* (2019), bien qu'ils aient opté pour FaceNet (Schroff *et al.*, 2015) et ne donnent pas de détails sur le jeu de données d'entraînement[14].

Le modèle ResNet, dont les connexions résiduelles permettent de construire des réseaux très profonds, est très utilisé pour l'apprentissage de représentations visuelles, par exemple dans ArcFace. Nous désignons par « ImageNet-ResNet » le modèle entraîné sur mille catégories d'objets d'ImageNet (Deng *et al.*, 2009), le jeu de données de préentraînement le plus populaire pour la classification d'images. Les caractéristiques extraites de la dernière couche convolutive d'ImageNet-ResNet se sont en effet avérées être efficaces pour la recherche d'images (Sharif Razavian *et al.*, 2014 ; Ra-

14. C'est ce qui a motivé notre choix pour ArcFace car FaceNet a originellement été proposé dans une version entraînée avec un jeu de données propriétaire de Google.

denović *et al.*, 2018). Nous utilisons le *max-pooling* pour en réduire la carte de caractéristiques (*feature map*), compte tenu des résultats rapportés dans Radenović *et al.* (2018).

CLIP (Radford *et al.*, 2021) est une architecture permettant d'apprendre des représentations visuelles à partir d'une faible supervision textuelle. L'objectif d'apprentissage est similaire à celui de DPR, bien que CLIP associe des images à des légendes pertinentes au lieu de requêtes à des documents pertinents. CLIP a été entraîné sur un jeu de données de 400 millions de paires image-légende. Nous ne nous intéressons qu'à l'encodeur visuel de CLIP et laissons de côté son encodeur textuel.

Tous ces modèles sont gelés dans nos expériences, c'est-à-dire qu'ils ne sont pas ajustés. Dans un souci de comparaison équitable, nous utilisons systématiquement une architecture ResNet-50 pour toutes les représentations visuelles. La recherche dense est effectuée au moyen du produit scalaire, équivalent à la similarité cosinus car les représentations sont normalisées au préalable (sauf pour DPR).

5.3. *Fusion multimodale*

Les résultats de la recherche par l'image sont ensuite mis en correspondance avec les passages pour la fusion avec la recherche textuelle. Les scores des résultats de ces modèles ayant des distributions très différentes, ils sont centrés-réduits avant de les fusionner. La fusion est faite *via* une combinaison linéaire (Karpukhin *et al.*, 2020 ; Ma *et al.*, 2021) : $P = \alpha_b B + \alpha_d D + \mathbf{F}\alpha_a A + (1 - \mathbf{F})(\alpha_i I + \alpha_c C)$. On note $B,\ D,\ A,\ I,\ C$, les scores respectifs de BM25, DPR, ArcFace, ImageNet-ResNet et CLIP, chacun étant pondéré par l'hyperparamètre α_j. $\mathbf{F} \in \{0, 1\}$ dénote la détection d'un visage. Seuls les 100 premiers passages sont considérés. Par conséquent, si, compte tenu d'une requête, un passage n'est pas retrouvé par un système donné, il lui est attribué le score minimal des autres passages retrouvés par ce système (Ma *et al.*, 2021). Les passages sont ensuite réordonnés par rapport au score P. Les hyperparamètres d'interpolation α_j sont réglés sur l'ensemble de validation en utilisant une recherche par dichotomie pour maximiser le rang réciproque moyen. Pour limiter l'espace de recherche et permettre une comparaison directe entre BM25 et DPR, nous contraignons $\sum_j \alpha_j = 1$ et n'utilisons qu'un seul modèle pour la recherche texte : nous avons donc $\alpha_b = 0$ ou $\alpha_d = 0$.

5.4. *Résultats*

Puisqu'il est fondé sur TriviaQA (Joshi *et al.*, 2017), ViQuAE n'est supervisé que de façon distante, c'est-à-dire qu'un document est jugé pertinent s'il contient la réponse. Nous évaluons la RI avec la précision à K (P@K) et le rang réciproque moyen (MRR) ainsi que Hits@K. Hits@K représente la proportion de questions pour lesquelles la RI récupère *au moins un* document pertinent parmi les K premiers. Les résultats sont présentés dans les tableaux 2 et 3. Les tests de significativité statistique

#	Modèle	MRR	P@1	P@20	Hits@20
-	$\mathbf{F}A$ (ArcFace, visage détecté)	54,3	50,2	5,5	65,3
a	$(1-\mathbf{F})I$ (ImageNet, *pas* de visage détecté)	17,5	11,9	4,9	36,1
b	$(1-\mathbf{F})C$ (CLIP, *pas* de visage détecté)	**27,5**^{a}	**20,5**^{a}	**9,5**^{a}	**53,1**^{a}
a	B (BM25, texte seulement)*	23,2	16,5	7,1	45,3
b	D_0 (DPR *zero-shot*, texte seulement)*	35,5^{a}	24,9^{a}	**17,6**^{ac}	66,5^{ac}
c	$\mathbf{F}0,3A+(1-\mathbf{F})(0,1I+0,3C)$	**41,4**^{ab}	**35,6**^{abd}	7,4	59,5^{a}
d	D_f (DPR *few-shot*, texte seulement)*	38,2^{ab}	27,8^{ab}	17,5^{ac}	**66,7**^{ac}

TABLEAU 2. *Résultats de la RI initiale évaluée au niveau de l'article : sur deux sous-ensembles (visage détecté ou pas) et sur le test complet. *Chaque article se voit assigner le score maximal de ses passages. Les exposants dénotent des différences significatives selon le test de randomisation de Fisher avec $p \leq 0{,}01$. Hits@1 est omis car il est équivalent à P@1.*

sont effectués à l'aide du test de randomisation de Fisher (Fisher, 1937 ; Smucker *et al.*, 2007). Nous présentons également comme référence les performances de BM25 et de DPR utilisant seulement le texte.

Pour étudier l'apport de chaque modalité séparément, nous évaluons les résultats au niveau de l'article. Cette comparaison suppose deux subtilités :

– les différentes représentations visuelles (section 5.2) sont d'abord évaluées séparément sur des sous-ensembles du jeu de données selon la détection des visages, puis en combinaison sur l'ensemble du jeu de test ;

– DPR fonctionne au niveau du passage car il est fondé sur BERT, qui ne peut pas traiter directement de longs articles. De plus, le passage sert comme unité par la suite pour l'extraction des réponses, elle aussi fondée sur BERT. Dans le tableau 2 nous avons donc simplement assigné à l'article le score maximal de ses passages pour avoir un point de comparaison, mais la performance au niveau du passage est étudiée dans le tableau 3.

Dans le tableau 2, nous constatons que la recherche *via* l'image obtient d'assez bons résultats, notamment avec ArcFace quand un visage est détecté. Comparativement, Shah *et al.* (2019) obtiennent une P@1 de 73,5 avec FaceNet mais leur BC est cent fois plus petite. Par ailleurs, CLIP surpasse largement ImageNet quand aucun visage n'est détecté. Ces résultats concordent avec ceux de Radford *et al.* (2021) et pourraient motiver de futurs travaux sur l'utilisation de CLIP pour la recherche d'image par le contenu. Enfin, nous remarquons une dynamique très différente entre les modèles visuels et textuels, notamment entre ArcFace et DPR : ArcFace est très précis mais avec un rappel relativement mauvais tandis que c'est l'inverse pour DPR. Ceci s'explique par l'objet des questions, qui permet de deviner la réponse avec un nombre suffisamment grand d'essais sans regarder l'image. Pour l'exemple de la figure 1, DPR pourrait ainsi retourner toutes les circonscriptions (*constituency*) du Royaume-Uni, ordonnées aléatoirement.

#	Modèle	MRR	P@1	P@20	Hits@20
a	B (BM25, texte seulement)	19,0	13,1	5,9	39,5
b	D_0 (DPR *zero-shot*, texte seulement)	30,5^{a}	21,2^{a}	16,2^{ac}	60,5^{ac}
c	$0,3(B + \mathbf{F}A) + (1 - \mathbf{F})(0,1I + 0,3C)$	27,9^{a}	20,4^{a}	10,1^{a}	50,5^{a}
d	$0,3(D_0 + \mathbf{F}A) + (1 - \mathbf{F})(0,1I + 0,3C)$	36,0^{abce}	26,7^{abce}	17,1^{ac}	65,2^{abce}
e	D_f (DPR *few-shot*, texte seulement)	32,8^{abc}	22,8^{a}	16,4^{ac}	61,2^{ac}
f	$0,3(D_f + \mathbf{F}A) + 0,2(1 - \mathbf{F})(I + C)$	**37,9**^{abcde}	**27,8**^{abce}	**17,5**^{ac}	**65,7**^{abce}

TABLEAU 3. *Résultats de la RI évaluée au niveau du passage avec les* baseline *textuelles et la fusion de la recherche multimodale, dans les deux configurations d'apprentissage : sans ou avec peu d'exemples*

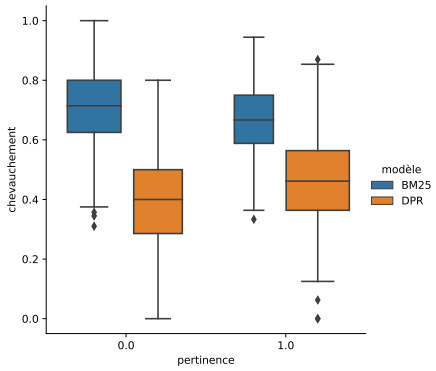

FIGURE 4. *Chevauchement entre les lemmes de la question et du premier passage retourné par BM25 et DPR en fonction de la pertinence du passage. Chaque boîte montre les quartiles tandis que ses moustaches s'étendent pour montrer le reste de la distribution, à l'exception des valeurs extrêmes.*

Le gain de performance de DPR par rapport à BM25 est important, y compris dans sa version *zero-shot* où il surpasse significativement BM25 et même la recherche multimodale fondée sur BM25, pour P@20 et Hits@20. Contrairement à BM25, DPR est capable de trouver des passages pertinents, même avec très peu de chevauchement lexical, grâce à ses représentations sémantiques distribuées, comme on peut le voir sur la figure 4. Toutefois, ses passages pertinents tendent tout de même à avoir un chevauchement supérieur avec la question par rapport à ses passages non pertinents. DPR pose par ailleurs le problème d'être plus sensible aux biais des jeux de données que BM25. Par exemple, pour une question telle que « Dans quel pays est-ce que cette personne est née ? », le modèle peut être biaisé par la distribution selon le jeu d'entraînement si une nationalité est surreprésentée.

Par ailleurs, il faut noter que la fusion multimodale apporte des gains de performance significatifs. Ce gain diffère selon le type de l'entité-sujet de la question. Pour

#	Modèle	MRR	P@1	P@20	Hits@20
a	B (BM25, texte seulement)	19,8	14,4	6,1	37,6
b	D_0 (DPR *zero-shot*, texte seulement)	28,0^{a}	19,2^{a}	14,4^{a}	57,9^{a}
c	$0,3(B + \mathbf{F}A) + (1 - \mathbf{F})(0,1I + 0,3C)$	32,4^{a}	24,4^{a}	11,9^{a}	56,0^{a}
d	$0,3(D_0 + \mathbf{F}A) + (1 - \mathbf{F})(0,1I + 0,3C)$	37,9^{abce}	28,9^{abe}	17,4^{abce}	67,4^{abce}
e	D_f (DPR *few-shot*, texte seulement)	31,1^{ab}	21,7^{a}	15,2^{ac}	57,5^{a}
f	$0,3(D_f + \mathbf{F}A) + 0,2(1 - \mathbf{F})(I + C)$	**40,4**^{abce}	**29,8**^{abce}	**18,4**^{abcde}	**67,8**^{abce}
a	B (BM25, texte seulement)	18,3	12,1	5,8	41,0
b	D_0 (DPR *zero-shot*, texte seulement)	32,7^{ac}	22,9^{ac}	**17,7**^{ac}	62,6^{ac}
c	$0,3(B + \mathbf{F}A) + (1 - \mathbf{F})(0,1I + 0,3C)$	24,1^{a}	17,1^{a}	8,5^{a}	45,9^{a}
d	$0,3(D_0 + \mathbf{F}A) + (1 - \mathbf{F})(0,1I + 0,3C)$	34,3^{ac}	24,7^{ac}	16,9^{ac}	63,4^{ac}
e	D_f (DPR *few-shot*, texte seulement)	34,1^{ac}	23,8^{ac}	17,4^{ac}	**64,3**^{ac}
f	$0,3(D_f + \mathbf{F}A) + 0,2(1 - \mathbf{F})(I + C)$	**35,7**^{abc}	**26,0**^{ac}	16,8^{ac}	64,0^{ac}

TABLEAU 4. *Résultats de la RI évaluée au niveau du passage pour les questions à propos de personnes (partie supérieure) et de non-personnes (partie inférieure)*

les questions à propos de personnes, la P@1 passe de 14,4 avec BM25 seul à 24,4 en fusionnant BM25 et la recherche d'images, soit une amélioration de 70 %. En comparaison, l'amélioration est plus faible, seulement 41 %, en termes de P@1 pour les questions sur les non-personnes (cf. tableau 4). En outre, sur le sous-ensemble d'entités qui se chevauchent avec MS-Celeb (le jeu de données de préentraînement d'ArcFace), la valeur de P@1 monte jusqu'à 25,7, ce qui représente une amélioration de 5 % par rapport au score mesuré sur toutes les personnes. De manière similaire, la fusion multimodale apporte un gain significatif avec DPR, même en tenant compte du fait que sa *baseline* textuelle est meilleure.

Plus globalement, ces premiers résultats montrent des tendances intéressantes mais également une marge d'amélioration importante, laissant la place à de futurs travaux sur la fusion multimodale. En attendant, nous présentons une *baseline* pour le réordonnancement des résultats de cette étape à la section suivante.

6. Réordonnancement

La relative faiblesse de la précision de la RI initiale est une conséquence, en particulier pour les approches denses, d'une modélisation limitée par la taille de la BC à considérer. Par conséquent, il est intéressant de réordonnancer les passages issus de cette première phase, beaucoup plus restreints en termes de volume, afin d'améliorer la précision. Nous adoptons une approche de fusion tardive similaire à celle de la RI initiale. Le réordonnancement est effectué indépendamment avec le texte et l'image puis les résultats sont fusionnés.

#	Modèle	MRR	P@1	P@20	Hits@20
a	RRT (visage détecté)	37,4	25,2	13,2^{b}	73,7
b	$\mathbf{F}A$ (ArcFace, visage détecté)	49,6^{a}	42,9^{a}	12,2	76,3
c	$\mathbf{F}A + \mathrm{RRT}$ (visage détecté)	**52,4**^{ab}	**43,0**^{a}	**13,3**^{b}	**77,4**^{a}
a	RRT	39,2	27,5	14,1	73,7
b	$\mathbf{F}A + \mathrm{RRT}$	**47,0**^{a}	**36,7**^{a}	**14,2**	**75,6**^{a}

TABLEAU 5. *Évaluation au niveau de l'article du réordonnancement de la RI initiale, selon la détection d'un visage. Les exposants dénotent des différences significatives dans le test de randomisation de Fisher avec $p \leq 0,01$.*

6.1. *Réordonnancement pour l'image*

Pour le réordonnancement d'image, nous utilisons deux représentations d'image différentes : ArcFace pour les visages, si au moins un visage est détecté par MTCNN, comme pour la RI initiale ; Re-Ranking Transformers (RRT) (Tan *et al.*, 2021) pour l'image complète. Si aucun visage n'est détecté, nous utilisons uniquement RRT pour réordonnancer les images ; sinon, nous combinons ArcFace et RRT en utilisant encore une fois la technique du « minimum par défaut » de Ma *et al.* (2021).

RRT utilise l'architecture Transformer (Vaswani *et al.*, 2017) et son mécanisme d'auto-attention pour combiner les caractéristiques globales et locales obtenues à partir de DELG (Cao *et al.*, 2020), un extracteur de caractéristiques, les deux méthodes étant remarquablement efficaces. Le modèle RRT prend en entrée une séquence de représentations globales et locales obtenues à partir d'une paire d'images (associées à la question et au passage) et utilise le plongement du token spécial `[CLS]` pour apprendre une métrique de similarité. RRT et DELG sont entraînés sur Google Landmarks v2 (GLDv2) (Weyand *et al.*, 2020), composé de photos de monuments, et ne sont pas ajustés dans nos expériences. Les URLs de ces photos ont un chevauchement de 9 % avec les images des questions de ViQuAE, qui est analysé à la section 6.4.

Les scores d'ArcFace et RRT n'étant pas comparables, nous les normalisons d'abord en fonction du rang puis nous utilisons PosFuse (Lillis *et al.*, 2010) pour fusionner les scores normalisés. La normalisation par le rang est très utile pour minimiser l'effet des valeurs extrêmes. Elle consiste à attribuer à chaque passage le score $1 - \frac{r-1}{K}$, où r est le rang du passage et K, le nombre total de passages à ordonner (100 dans nos expériences). PosFuse est quant à elle une méthode supervisée qui apprend la probabilité qu'un passage apparaissant à une position donnée soit pertinent. Elle est optimisée sur le jeu de validation en utilisant une recherche par dichotomie.

6.2. *Réordonnancement pour le texte*

Comme Wang *et al.* (2019), notre réordonnanceur de texte prend en entrée la concaténation d'une paire question-passage et l'encode au moyen de BERT. De façon

#	Modèle	MRR	P@1	P@20	Hits@20
a	RI initiale	37,9	27,8	17,5	65,7
b	Texte	47,4^{a}	37,8^{a}	23,7^{a}	73,1^{a}
c	Texte + $\mathbf{F}A$ + RRT	**52,7**^{ab}	**43,7**^{ab}	**25,1**^{ab}	**75,3**^{ab}

TABLEAU 6. *Évaluation au niveau du passage du réordonnancement de la RI initiale, fondé seulement sur le texte ou en fusionnant texte et image*

comparable à RRT, la représentation associée au token [CLS] est introduite dans un perceptron pour prédire un score de pertinence unique pour chaque passage. Le modèle est entraîné sur 24 passages (avec un seul passage pertinent) échantillonnés parmi les 100 meilleurs passages retournés par la RI. Comme à la section précédente, le modèle est d'abord préentraîné sur notre sous-ensemble de TriviaQA, avec une RI effectuée avec BM25 sur les 5,9 millions d'articles de la Wikipédia de KILT au lieu de notre BC multimodale. Le modèle est ensuite ajusté sur ViQuAE en utilisant les mêmes hyperparamètres, la RI étant alors effectuée avec le modèle multimodal fondé sur DPR.

6.3. *Fusion multimodale*

Les scores du réordonnanceur textuel sont ensuite fusionnés avec ceux du réordonnanceur d'image *via* une simple combinaison linéaire, comme pour la RI initiale, après une normalisation min-max. Les méthodes de normalisation, de fusion et les hyperparamètres de ces dernières ont été choisis sur le jeu de validation, avec une recherche par dichotomie.

6.4. *Résultats*

Les performances des réordonnanceurs sont présentées dans les tableaux 5 et 6 de façon comparable à la section précédente. Le modèle pour le texte améliore l'ordonnancement initial de 10 points de MRR et P@1, ce qui représente des améliorations de 25 % et 36 % respectivement. La fusion avec le modèle d'image augmente encore les performances de 5,9 points de P@1, soit une amélioration totale de 57 %. Le chevauchement de 9 % avec GLDv2 (le jeu de données de préentraînement de RRT) est probablement trop faible pour en tirer des conclusions significatives : l'amélioration relative entre les réordonnanceurs texte et multimodal est similaire avec ou sans chevauchement.

L'amélioration de l'ordonnancement de la RI initiale est sans doute due pour l'essentiel au mécanisme d'auto-attention des Transformers appliqué aux paires question-passage, avec BERT dans le cas du texte et avec RRT dans le cas de l'image. L'auto-attention est un mécanisme coûteux mais qui permet de capturer des interactions plus riches qu'un simple produit scalaire (Karpukhin *et al.*, 2020).

Requête	1er résultat	2ème résultat	3ème résultat
			
« *This arch bridge spans what river ?* »	« Marlow Bridge [SEP] [...] The Széchenyi Chain Bridge, spanning the River *Danube* in Budapest [...] »	« Hudson River [SEP] The width of the Lower *Hudson River* required major feats of engineering to cross [...]	« Pont de la Tournelle [SEP] [...] This bridge connected the Eastern bank of the **Seine** (le quai Saint-Bernard) to l'île Saint-Louis. [...]
			
« *What was the last film directed by this film producer ?* »	« David Lean [SEP] [...] responsible for large-scale epics such as "The Bridge on the River Kwai" (1957), [...] and "**A Passage To India**" (1984). »	« Bernard Herrmann [SEP] [...] is particularly known for [...] "*Psycho*", "*North by Northwest*", "*The Man Who Knew Too Much*", and "*Vertigo*". »	« David Lean [SEP] [...] Lean recruited long-time collaborators for the cast and crew, [...] John Box, the production designer for "*Dr. Zhivago*". »

FIGURE 5. *Requêtes accompagnées des trois premiers résultats de la RI multimodale initiale. La réponse (dans le passage pertinent) est imprimée en caractères gras et les réponses plausibles dans les passages non pertinents sont imprimées en italique. Les visages détectés sont indiqués en rouge. Le passage de texte a été raccourci pour la mise en page.*

7. Extraction des réponses

7.1. *Méthodes*

Le but de cette étape est d'extraire une réponse concise à partir d'un passage de texte candidat (provenant par exemple des étapes précédentes de RI). Pour établir notre référence sur ViQuAE, nous nous limitons à un modèle textuel car nous faisons l'hypothèse qu'*une fois le passage pertinent retrouvé en associant texte et image*, il est *possible* de répondre à la question sans utiliser l'image (cf. exemple de la figure 1). L'extraction des réponses est réalisée avec le modèle BERT multipassage de Wang *et al.* (2019). Ce modèle prend en entrée la concaténation de la question et du passage et les encode avec BERT, comme le réordonnanceur. Les représentations sont ensuite données à deux perceptrons différents, entraînés indépendamment à prédire les positions de début et de fin de la réponse. Lors de l'inférence, la probabilité de la position de la réponse est le produit des probabilités de début et de fin. Afin de rendre les scores de réponse comparables d'un passage à l'autre, BERT multipassage exploite la technique de la normalisation globale de Clark et Gardner (2018) afin que

# Exemples	Entrée	F1	Appariement exact (EM)
Aucun	Top 5 RI initiale	22,1	18,5
Aucun	Top 5 réordonnancement	27,7	24,3
Aucun	+ pondération	29,4	26,1
Peu	Top 5 RI initiale	25,5 $\pm$ 0,7	21,4 $\pm$ 0,8
Peu	Top 5 réordonnancement	32,7 $\pm$ 0,4	28,9 $\pm$ 0,4
Peu	+ pondération	33,8 $\pm$ 0,5	30,1 $\pm$ 0,6
Peu	Mi-oracle	43,1 $\pm$ 0,2	39,1 $\pm$ 0,4
Peu	Oracle complet	66,5 $\pm$ 0,7	60,7 $\pm$ 0,9

TABLEAU 7. *Résultats de l'extraction de réponses sur l'ensemble de test de ViQuAE. Pour le modèle few-shot, moyennes sur 5 entraînements avec des graines aléatoires différentes. En inférence, les modèles zero et few-shot prennent les 5 premiers passages en entrée.*

tous les passages partagent la même normalisation softmax. Pour les passages non pertinents, le modèle est entraîné à prédire la première position, c'est-à-dire celle du token spécial [CLS]. De plus, puisque la réponse peut apparaître plusieurs fois dans le même passage, l'objectif d'entraînement, à l'instar de Karpukhin *et al.* (2020), est de maximiser la log-vraisemblance marginale de toutes les positions de réponse dans le passage. Pour prendre en compte les scores P associés aux passages par l'étape précédente de réordonnancement, nous pondérons le score de réponse a tel que $a \leftarrow a \cdot P$ (Wang *et al.*, 2019).

Le modèle est implémenté et entraîné en utilisant la bibliothèque Transformers de Hugging Face (Wolf *et al.*, 2020), elle-même fondée sur PyTorch (Paszke *et al.*, 2019). Les mêmes hyperparamètres que Karpukhin *et al.* (2020) sont utilisés, à l'exception du ratio de passages pertinents et non pertinents par question, qui est fixé à $8{:}16$. Nous avons également étudié sur le jeu de validation l'effet de la variation du nombre de passages sur l'extraction de réponses lors de l'inférence. Avec les 5 premiers passages, les modèles sont nettement meilleurs. Dans la suite, l'extraction est ainsi appliquée sur les 5 premiers résultats de la RI, initiale ou réordonnancée.

Comme pour le réordonnanceur de texte, le modèle est préentraîné sur TriviaQA et ensuite ajusté sur ViQuAE. Bien que le modèle soit préentraîné, étant donné la petite taille de ViQuAE, l'entraînement a été effectué 5 fois avec des graines aléatoires différentes pour tenir compte de la variabilité causée par l'ordre des questions et le choix aléatoire des passages pertinents et non pertinents parmi leurs ensembles respectifs.

7.2. *Résultats*

Conformément à Joshi *et al.* (2017) ainsi qu'à Petroni *et al.* (2021), nous utilisons l'appariement exact (EM) et le score F1 pour évaluer l'extraction de la réponse après un prétraitement standard (normalisation en minuscule, suppression des articles et de la ponctuation). Les résultats sont présentés dans le tableau 7. Sans surprise,

l'ajustement du modèle sur l'ensemble d'entraînement améliore les performances : + 16 % d'EM. Dans les deux cas, le réordonnancement apporte une amélioration notable de plus de 32 % et la pondération réalisée avec son score est également bénéfique.

Toutefois, les résultats sont globalement assez faibles par rapport à l'état de l'art en QA textuelle. Nous pouvons les comparer aux performances du modèle sur le sous-ensemble de TriviaQA qui a servi à générer le test de ViQuAE : 62,9 de F1 et 59,2 d'EM en prenant en entrée le top 24 de BM25[15], ce qui est du même ordre de grandeur que les résultats obtenus par Wang *et al.* (2019) et par Karpukhin *et al.* (2020) sur les sous-ensembles officiels de validation et de test respectivement. On observe ainsi une amélioration relative de 147 % (F1) et 177 % (EM) en passant de ViQuAE à TriviaQA pour le même ensemble initial de questions.

Pour mieux comprendre ces chiffres, nous avons étudié deux configurations différentes. Premièrement, *mi-oracle*, où les 5 premiers résultats du réordonnancement sont filtrés pour ne contenir que des passages pertinents (s'il y en a ; sinon, la réponse extraite sera fausse). Cette configuration se traduit par une amélioration significative de 35 % d'EM par rapport à la référence et montre ainsi que le modèle ne fait pas bien la distinction entre un passage pertinent et non pertinent, même si le réordonnancement permet de réduire l'écart[16]. Par exemple, dans la figure 5, deux passages sur trois ne sont pas pertinents mais fournissent une réponse plausible à la question. De futurs travaux pourraient se focaliser sur une meilleure intégration de l'image dans l'extraction de la réponse. Enfin, nous avons considéré la configuration *oracle complet*, où le modèle ne reçoit que des passages pertinents[17]. L'écart de performance continue de se creuser : + 55 % en EM par rapport à *mi-oracle*, qui souffre des résultats modestes de la RI. Ce constat corrobore les résultats de la section 5 : la KVQAE est très difficile pour les représentations d'images actuelles et de futurs travaux devraient porter sur une meilleure fusion des informations multimodales. De plus, ces chiffres assez élevés, comparables aux résultats sur TriviaQA, confirment notre hypothèse : une fois que le passage pertinent a été retrouvé, il est possible de répondre à la question sans regarder l'image. Ces résultats *oracle* pourraient servir de référence haute aux futures études.

8. Conclusion et perspectives

Nous présentons un nouveau jeu de données, ViQuAE, conçu comme un cadre d'évaluation pour suivre les progrès des systèmes de KVQAE. ViQuAE a été annoté selon une procédure semi-automatique que nous fournissons également. Ses questions ont pour cible une base de connaissances librement disponible de 1,5 million d'articles

15. 63,3 de F1 et 59,7 d'EM en pondérant avec le score de BM25. BM25 a un MRR de 70,6 et une P@1 de 60,2 sur ce sous-ensemble.

16. Les résultats *mi-oracle* sont similaires, qu'ils proviennent de la RI initiale ou réordonnancée.

17. Pour *oracle complet*, les résultats de la RI sont filtrés de la même manière que pour *mi-oracle* mais s'il n'y en a aucun, on utilise ceux liés à l'article Wikipédia de l'entité-sujet.

Wikipédia associés à des images. Par rapport au jeu de données existant KVQA (Shah *et al.*, 2019), ViQuAE couvre notamment différents types d'entités et de sujets. Cependant, il ne contient que des questions *mono-hop* au niveau de l'article et ses images représentent une seule et unique entité, sauf dans certains cas exceptionnels. Nos résultats suggèrent que cette configuration fournit déjà de nombreux défis mais une future version du jeu de données pourrait introduire des questions *multi-hop* ou plusieurs entités par image.

Nous proposons aussi une approche de la KVQAE en trois étapes, distinguant recherche d'information initiale, réordonnancement et extraction des réponses, avec des méthodes d'apprentissage sans ou avec peu d'exemples. Un résultat notable de cette première référence est l'apport positif de l'association du texte et de l'image dans ces différentes configurations. Sans négliger l'extraction des réponses, les évaluations soulignent par ailleurs la nécessité d'une meilleure RI. En effet, notre stratégie de fusion tardive néglige l'interaction entre les modalités. Les travaux futurs devront se concentrer sur une meilleure représentation multimodale, idéalement en intégrant le texte et l'image dans le même espace, tant du côté de la requête que du côté de la BC. Une attention particulière devra être accordée à la représentation des entités non-personnes. Ces représentations multimodales pourront aussi bénéficier à l'étape d'extraction des réponses car nos expériences montrent que l'utilisation d'un modèle textuel seul est insuffisante si la RI est bruitée, bien que le réordonnancement permette en partie de pallier ce cas de figure. D'autre part, bien que nous ayons démontré l'efficacité de notre BC, un système de KVQAE pourrait tirer bénéfice d'une BC plus riche visuellement, avec plusieurs images par entité, afin de prendre en compte la diversité des représentations. Nous espérons plus globalement que ce travail encouragera la recherche vers une meilleure représentation multimodale des entités nommées.

Remerciements

Nous remercions les relecteurs anonymes pour leurs retours constructifs. Ce travail a été financé par le projet ANR-19-CE23-0028 MEERQAT. Il a en outre bénéficié d'un accès aux moyens de calcul de l'IDRIS au travers de l'allocation de ressources 2021-AD011012846 attribuée par GENCI.

9. Bibliographie

Anderson P., He X., Buehler C., Teney D., Johnson M., Gould S., Zhang L., « Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering », *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June, 2018.

Antol S., Agrawal A., Lu J., Mitchell M., Batra D., Zitnick C. L., Parikh D., « VQA : Visual Question Answering », *2015 IEEE International Conference on Computer Vision (ICCV)*, IEEE, Santiago, Chile, December, 2015.

Bommasani R., *et al.*, « On the Opportunities and Risks of Foundation Models », *arXiv :2108.07258 [cs]*, August, 2021. arXiv : 2108.07258.

Cao B., Araujo A., Sim J., « Unifying deep local and global features for image search », *European Conference on Computer Vision*, Springer, 2020.

Chang Y., Narang M., Suzuki H., Cao G., Gao J., Bisk Y., « WebQA : Multihop and Multimodal QA », *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 16495-16504, June, 2022.

Chen D., Fisch A., Weston J., Bordes A., « Reading Wikipedia to Answer Open-Domain Questions », *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, 2017.

Clark C., Gardner M., « Simple and Effective Multi-Paragraph Reading Comprehension », *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, Association for Computational Linguistics, Melbourne, Australia, July, 2018.

Clough P., Sanderson M., Müller H., « The CLEF Cross Language Image Retrieval Track (ImageCLEF) 2004 », *in* P. Enser, Y. Kompatsiaris, N. E. O'Connor, A. F. Smeaton, A. W. M. Smeulders (eds), *Image and Video Retrieval*, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2004.

Deng J., Dong W., Socher R., Li L.-J., Li K., Fei-Fei L., « ImageNet : A large-scale hierarchical image database », *2009 IEEE Conference on Computer Vision and Pattern Recognition*, June, 2009. ISSN : 1063-6919.

Deng J., Guo J., Xue N., Zafeiriou S., « ArcFace : Additive Angular Margin Loss for Deep Face Recognition », *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June, 2019.

Devlin J., Chang M.-W., Lee K., Toutanova K., « BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding », *Proceedings of the 2019 NAACL, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, June, 2019.

Ferragina P., Scaiella U., « TAGME : on-the-fly annotation of short text fragments (by wikipedia entities) », *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM'10)*, Association for Computing Machinery, New York, NY, USA, October, 2010.

Fisher R. A., *The Design of Experiments*, 2ème edn, Oliver & Boyd, Edinburgh & London., 1937.

Fleiss J. L., « Measuring nominal scale agreement among many raters », *Psychological Bulletin*, 1971.

Gardères F., Ziaeefard M., Abeloos B., Lecue F., « ConceptBert : Concept-Aware Representation for Visual Question Answering », *Findings of the Association for Computational Linguistics : EMNLP 2020*, Association for Computational Linguistics, Online, November, 2020.

Guo Y., Zhang L., Hu Y., He X., Gao J., « MS-Celeb-1M : A Dataset and Benchmark for Large-Scale Face Recognition », *in* B. Leibe, J. Matas, N. Sebe, M. Welling (eds), *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2016.

He K., Zhang X., Ren S., Sun J., « Deep residual learning for image recognition », *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

Jain A., Kothyari M., Kumar V., Jyothi P., Ramakrishnan G., Chakrabarti S., « Select, Substitute, Search : A New Benchmark for Knowledge-Augmented Visual Question Answering », *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, Association for Computing Machinery, New York, NY, USA, 2021.

Johnson J., Douze M., Jégou H., « Billion-scale similarity search with GPUs », *IEEE Transactions on Big Data*, 2019.

Joshi M., Choi E., Weld D., Zettlemoyer L., « TriviaQA : A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension », *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, Association for Computational Linguistics, Vancouver, Canada, July, 2017.

Karpukhin V., Oguz B., Min S., Lewis P., Wu L., Edunov S., Chen D., Yih W.-t., « Dense Passage Retrieval for Open-Domain Question Answering », *Proceedings of the 2020 EMNLP (EMNLP)*, Association for Computational Linguistics, Online, November, 2020.

Kembhavi A., Seo M., Schwenk D., Choi J., Farhadi A., Hajishirzi H., « Are You Smarter Than a Sixth Grader ? Textbook Question Answering for Multimodal Machine Comprehension », *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July, 2017.

Kwiatkowski T., Palomaki J., Redfield O., Collins M., Parikh A., Alberti C., Epstein D., Polosukhin I., Devlin J., Lee K., Toutanova K., Jones L., Kelcey M., Chang M.-W., Dai A. M., Uszkoreit J., Le Q., Petrov S., « Natural Questions : A Benchmark for Question Answering Research », *Transactions of the Association for Computational Linguistics*, March, 2019.

Lerner P., Ferret O., Guinaudeau C., Le Borgne H., Besançon R., Moreno J. G., Lovón Melgarejo J., « Un jeu de données pour répondre à des questions visuelles à propos d'entités nommées en utilisant des bases de connaissances », *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) 2022.*, ATALA, Avignon, France, 2022a.

Lerner P., Ferret O., Guinaudeau C., Le Borgne H., Besançon R., Moreno J. G., Lovón Melgarejo J., « ViQuAE, a Dataset for Knowledge-based Visual Question Answering about Named Entities », *Proceedings of The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, Association for Computing Machinery, New York, NY, USA, 2022b.

Lhoest Q., *et al.*, « Datasets : A Community Library for Natural Language Processing », *Proceedings of the 2021 EMNLP : System Demonstrations*, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, November, 2021.

Lillis D., Zhang L., Toolan F., Collier R. W., Leonard D., Dunnion J., « Estimating probabilities for effective data fusion », *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 2010.

Lin T.-Y., Maire M., Belongie S., Hays J., Perona P., Ramanan D., Dollár P., Zitnick C. L., « Microsoft COCO : Common Objects in Context », *in* D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (eds), *Computer Vision – ECCV 2014*, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2014.

Ma X., Sun K., Pradeep R., Lin J., « A Replication Study of Dense Passage Retriever », *arXiv :2104.05740 [cs]*, April, 2021.

Marino K., Rastegari M., Farhadi A., Mottaghi R., « OK-VQA : A visual question answering benchmark requiring external knowledge », *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

Paszke A., *et*, *al.*, « PyTorch : An Imperative Style, High-Performance Deep Learning Library », *Advances in Neural Information Processing Systems*, 2019.

Petroni F., Piktus A., Fan A., Lewis P., Yazdani M., De Cao N., Thorne J., Jernite Y., Karpukhin V., Maillard J., Plachouras V., Rocktäschel T., Riedel S., « KILT : a Benchmark for Knowledge Intensive Language Tasks », *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, Association for Computational Linguistics, Online, June, 2021.

Radenović F., Iscen A., Tolias G., Avrithis Y., Chum O., « Revisiting Oxford and Paris : Large-Scale Image Retrieval Benchmarking », *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June, 2018.

Radford A., Kim J. W., Hallacy C., Ramesh A., Goh G., Agarwal S., Sastry G., Askell A., Mishkin P., Clark J. *et al.*, « Learning transferable visual models from natural language supervision », *International Conference on Machine Learning*, PMLR, 2021.

Rajpurkar P., Zhang J., Lopyrev K., Liang P., « SQuAD : 100,000+ Questions for Machine Comprehension of Text », *Proceedings of the 2016 EMNLP*, Association for Computational Linguistics, Austin, Texas, November, 2016.

Reddy R. G., Rui X., Li M., Lin X., Wen H., Cho J., Huang L., Bansal M., Sil A., Chang S.-F., Schwing A., Ji H., « MuMuQA : Multimedia Multi-Hop News Question Answering via Cross-Media Knowledge Extraction and Grounding », December, 2021.

Robertson S. E., Walker S., Jones S., Hancock-Beaulieu M. M., Gatford M., « Okapi at TREC-3 », *in* D. K. Harman (ed.), *Third Text REtrieval Conference (TREC-3)*, vol. 500-225 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 1995.

Sampat S. K., Yang Y., Baral C., « Visuo-Linguistic Question Answering (VLQA) Challenge », *Findings of the Association for Computational Linguistics : EMNLP 2020*, Association for Computational Linguistics, Online, November, 2020.

Schroff F., Kalenichenko D., Philbin J., « FaceNet : A Unified Embedding for Face Recognition and Clustering », *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June, 2015.

Schwenk D., Khandelwal A., Clark C., Marino K., Mottaghi R., « A-OKVQA : A Benchmark For Visual Question Answering Using World Knowledge », *Computer Vision – ECCV 2022 : 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, Springer-Verlag, Berlin, Heidelberg, p. 146–162, 2022.

Shah S., Mishra A., Yadati N., Talukdar P. P., « KVQA : Knowledge-Aware Visual Question Answering », *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019.

Sharif Razavian A., Azizpour H., Sullivan J., Carlsson S., « CNN Features Off-the-Shelf : An Astounding Baseline for Recognition », *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June, 2014.

Smucker M. D., Allan J., Carterette B., « A comparison of statistical significance tests for information retrieval evaluation », *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, Association for Computing Machinery, New York, NY, USA, November, 2007.

Srihari R. K., Zhang Z., Rao A., « Intelligent Indexing and Semantic Retrieval of Multimodal Documents », *Information Retrieval*, May, 2000.

Talmor A., Yoran O., Catav A., Lahav D., Wang Y., Asai A., Ilharco G., Hajishirzi H., Berant J., « MultiModalQA : Complex Question Answering over Text, Tables and Images », *ICLR 2021*, 2021.

Tan F., Yuan J., Ordonez V., « Instance-Level Image Retrieval Using Reranking Transformers », *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, p. 12105-12115, October, 2021.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L. u., Polosukhin I., « Attention is All you Need », *in* I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (eds), *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017.

Voorhees E. M., Tice D. M., « Building a question answering test collection », *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'00)*, ACM Press, Athens, Greece, 2000.

Wang P., Wu Q., Shen C., Dick A., Van Den Henge A., « Explicit knowledge-based reasoning for visual question answering », *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2017.

Wang P., Wu Q., Shen C., Dick A., van den Hengel A., « FVQA : Fact-Based Visual Question Answering », *IEEE transactions on pattern analysis and machine intelligence*, 2018.

Wang Z., Ng P., Ma X., Nallapati R., Xiang B., « Multi-passage BERT : A Globally Normalized BERT Model for Open-domain Question Answering », *Proceedings of the 2019 EMNLP and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, Hong Kong, China, November, 2019.

Weyand T., Araujo A., Cao B., Sim J., « Google Landmarks Dataset v2 - A Large-Scale Benchmark for Instance-Level Recognition and Retrieval », *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Wolf T., Debut L., Sanh V., Chaumond J., Delangue C., Moi A., Cistac P., Rault T., Louf R., Funtowicz M., Davison J., Shleifer S., von Platen P., Ma C., Jernite Y., Plu J., Xu C., Le Scao T., Gugger S., Drame M., Lhoest Q., Rush A., « Transformers : State-of-the-Art Natural Language Processing », *Proceedings of the 2020 EMNLP : System Demonstrations*, Association for Computational Linguistics, Online, p. 38-45, October, 2020.

Zhang K., Zhang Z., Li Z., Qiao Y., « Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks », *IEEE Signal Processing Letters*, October, 2016.

Techniques de synthèse vocale neuronale à l'épreuve des données d'apprentissage non dédiées : les livres audio amateurs en français

Aghilas Sini* — **Lily Wadoux*** — **Antoine Perquin*** — **Gaëlle Vidal*** — **David Guennec*** — **Damien Lolive*** — **Pierre Alain*** — **Nelly Barbot*** — **Jonathan Chevelu*** — **Arnaud Delhay***

* *Université de Rennes, CNRS, IRISA, France*

RÉSUMÉ. *Dans cet article, nous nous intéressons à la capacité des systèmes de synthèse vocale neuronale à tirer parti des données non dédiées en langue française. En effet, ces dernières sont abondantes mais leurs conditions d'enregistrement sont hétérogènes, alors que les données dédiées à la synthèse de parole (de meilleure qualité) sont en quantité limitée et difficiles à collecter. Leur impact est mesuré sur trois systèmes : synthèse de parole monolocuteur, clonage de voix et conversion de voix. Des évaluations objectives et subjectives sur la reproduction de la voix du locuteur et sur la qualité des échantillons synthétisés ont été menées. Elles montrent qu'il est difficile de produire une synthèse vocale de qualité comparable avec l'état de l'art dans certaines conditions d'enregistrement ou pour des voix atypiques.*

MOTS-CLÉS : *Synthèse de la parole, clonage de voix, conversion de voix, synthèse vocale neuronale.*

ABSTRACT. *In this article, we consider how neural speech synthesis systems perform with non-dedicated data in French. Indeed, these are plentiful, unlike dedicated data of better quality which are limited in their availability and difficult to collect, but are recorded in heterogeneous conditions. Their impact is measured on three systems: single-speaker speech synthesis, voice cloning and voice conversion. Speaker similarity and overall quality were measured through objective and subjective evaluations. Our results outline the difficulty of producing high-quality speech synthesis under some recording conditions, or for atypical voices.*

KEYWORDS: *Speech synthesis, Voice Cloning, Voice Conversion, Neural synthesis.*

1. Introduction

Les technologies de la synthèse vocale restent fortement contraintes par les données nécessaires à leur élaboration, à la fois en termes de qualité et de quantité. Même si les premiers systèmes de synthèse de la parole en utilisaient de très faibles quantités, l'évolution des méthodes a conduit à l'utilisation de plus grandes quantités de données. La synthèse de la parole de bout en bout (*end-to-end speech synthesis*) peut, selon les applications, nécessiter des dizaines d'heures, voire des centaines d'heures dans le cas de la synthèse massivement multilocuteur. Une constante, quelle que soit la technologie, reste que les données employées sont quasi exclusivement dédiées à la tâche de synthèse et de très haute qualité, en particulier si une application industrielle est visée. Dans la suite, on utilisera le terme « données dédiées » pour référencer les jeux de données conçus spécifiquement pour la synthèse de parole.

Les méthodes actuelles de synthèse de la parole sont généralement des variantes de la synthèse monolocuteur (Tan *et al.*, 2021). Elles impliquent une importante préparation et une quantité non triviale de données pour produire une voix de synthèse. Une approche récente, le clonage de voix, ne nécessite au contraire que quelques minutes de parole pour modéliser l'identité du locuteur (Snyder *et al.*, 2017). Une autre méthode réside dans la conversion de voix visant à transformer un énoncé vocal existant, produit par un locuteur source, afin qu'il soit perçu comme produit par un locuteur cible différent (Zhao *et al.*, 2019).

Les meilleurs résultats dans les publications en synthèse de la parole sont généralement obtenus sur la base de corpus d'apprentissage dédiés et très qualitatifs. Ils tendent à respecter un fort degré d'uniformité stylistique et une homogénéité des caractéristiques des locuteurs. On a ainsi des voix très majoritairement jeunes, plus souvent féminines que masculines, employant un style calme, posé, relativement neutre émotionnellement (et ce malgré une avancée de l'expressivité générale grâce aux méthodes de bout en bout). Ce manque de diversité s'explique souvent par les attentes présumées de la population pour laquelle la voix de synthèse a été produite et aussi par la difficulté à maîtriser des données expressives. Il existe peu d'études sur l'influence du choix de la voix sur la qualité de la synthèse produite (Hinterleitner *et al.*, 2014). À notre connaissance, il n'en existe pas sur les voix atypiques en français.

Des initiatives basées sur des données de qualité variable et non dédiées à la tâche de synthèse de la parole existent cependant, notamment en anglais. On peut par exemple citer le corpus *Librispeech* (Panayotov *et al.*, 2015), construit pour la tâche de reconnaissance de la parole, et une version améliorée (*LibriTTS* (Zen *et al.*, 2019)) où les audios les moins utilisables pour les applications de synthèse ont été exclus.

À notre connaissance, cet aspect de la synthèse de parole n'est pas exploré pour le français. Dans cet article, nous tâchons de pallier ce manque en considérant la question suivante : comment les systèmes de synthèse vocale se comportent-ils lorsqu'ils ont été entraînés sur des données de qualité inférieure aux standards du domaine et, qui plus est, non dédiées à la tâche de synthèse ?

Ne pouvant traiter le problème de manière exhaustive, nous restreignons le cadre de cette étude à l'utilisation de données disparates issues de livres audio expressifs, ayant subi un ensemble minimal de prétraitements. De même, une seule technologie majeure par méthode de synthèse est prise en compte. Nous nous basons ainsi sur l'architecture Tacotron 2 (Shen *et al.*, 2018) qui représente la base architecturale de la majorité des publications depuis son apparition. Une variante du modèle est apprise pour chacune des trois applications visées : synthèse, conversion et clonage de voix. Ces modèles sont appris sur un corpus de parole obtenu à partir de livres audio de locuteurs amateurs à l'expressivité variable. Le vocodeur est commun à toutes les approches afin d'évaluer uniquement le modèle acoustique, ce dernier étant généralement l'élément de la chaîne le plus impacté par la variabilité dans les données.

Nous commençons par présenter l'état de l'art sur les techniques de synthèse vocale utilisées dans cet article et justifions nos choix. La section 3 donne ensuite une présentation détaillée des données utilisées pour nos expériences. Les choix architecturaux et les détails concernant l'entraînement des différents modèles sont décrits en section 4. Enfin, les sections 5 et 6 détaillent le protocole expérimental et les résultats obtenus. Cette dernière section se conclut par une discussion des résultats. Nos conclusions et perspectives futures sont présentées en section 7.

2. Travaux connexes

Cette partie présente les technologies de synthèse vocale et les grands défis auxquels le domaine est confronté. Nous évoquons d'abord la synthèse de bout en bout de manière globale avant de nous focaliser sur le modèle acoustique puis sur le vocodeur. Enfin, nous discutons des défis relatifs aux données en synthèse de bout en bout ainsi que des travaux s'assimilant au nôtre sur cet aspect. Une discussion de la difficulté du processus d'évaluation de la synthèse est également abordée dans ce cadre.

2.1. *Processus de synthèse vocale*

Un système de synthèse de parole à partir du texte a pour objectif de produire, à partir d'une séquence de mots, éventuellement accompagnée de consignes, un signal de parole correspondant.

Ces systèmes présentent actuellement les meilleurs résultats en termes de rendu naturel dans l'état de l'art, au contraire des systèmes précédents (sélection d'unités, synthèse par HMM (*Hidden Markov Models*) ou DNN (*Deep Neural Network*) non de bout en bout) qui apparaissent de plus en plus rarement dans les *challenges* de comparaison (Ling *et al.*, 2021). Ils présentent de nombreux avantages, l'essentiel du savoir expert étant contenu dans le modèle lui-même. Ce dernier fait essentiellement un travail de conversion entre une séquence d'unités linguistiques (texte) ou phonétiques (parfois les deux) et une séquence cible de nature acoustique (audio). S'il existe des modèles réalisant cette conversion directement (Clarinet par exemple,

Ping *et al.* (2019)), on divise généralement le processus en deux modèles distincts. Il s'agit alors de prédire une représentation acoustique intermédiaire (généralement un mel-spectrogramme) plutôt que l'audio directement. Ce modèle est alors appelé modèle acoustique. Un second modèle, appelé vocodeur, vient alors traduire le mel-spectrogramme en un signal audio. L'avantage de ce principe est que le vocodeur peut être indépendant du locuteur et appris sur de grandes quantités de données multilocuteurs. Dans cette étude, nous utilisons un unique vocodeur pour l'ensemble des approches.

2.2. *Modélisation acoustique*

Dans cette section, nous détaillons le modèle acoustique et ses variantes pour la synthèse monolocuteur, le clonage et enfin la conversion de voix. Ils sont schématisés dans la figure 1.

2.2.1. *Modèle acoustique et synthèse monolocuteur*

Le modèle acoustique le plus populaire est Tacotron 2 (Shen *et al.*, 2018). Il s'agit d'un modèle neuronal séquence à séquence autorégressif qui suit l'architecture encodeur/décodeur et inclut un module d'attention. Des extensions ont été proposées afin d'accélérer la synthèse au prix d'un apprentissage plus complexe, tel FastSpeech2 (Ren *et al.*, 2020) ou d'ajouter du contrôle de la prosodie, par exemple en modélisant la prosodie d'un signal de parole de consigne à l'aide d'un auto-encodeur variationnel (Elias *et al.*, 2021). Elles sont néanmoins moins utilisées que Tacotron 2 dans la littérature, ce qui explique notre choix.

Le cas de la synthèse neuronale de bout en bout monolocuteur est le cas le plus simple d'utilisation du modèle acoustique. Avec cette méthode, chaque voix synthétisée est produite par un modèle acoustique distinct : un modèle par locuteur.

2.2.2. *Clonage de voix*

La synthèse de bout en bout monolocuteur requiert une quantité non négligeable de données par locuteur pour entraîner les modèles acoustiques correspondant à chaque locuteur. Ce n'est pas le cas pour la synthèse neuronale multilocuteur. Son principe est d'entraîner un unique modèle acoustique sur un corpus comprenant plusieurs locuteurs, dont ceux que l'on souhaite synthétiser. La possibilité d'utiliser plusieurs locuteurs à l'entraînement offre deux avantages. Premièrement, moins de données pour chaque locuteur sont nécessaires car l'agrégation de tous les locuteurs donne le volume de données requis. Deuxièmement, le modèle a la possibilité d'apprendre des différences entre locuteurs. Lors de la synthèse, la voix souhaitée est spécifiée au modèle acoustique par un vecteur *one-hot* (Arık *et al.*, 2017 ; Ping *et al.*, 2018). Cette approche ne peut donc synthétiser que les voix de son corpus d'entraînement.

Il est cependant possible d'utiliser ce type de modèle multilocuteur pour le personnaliser avec un locuteur absent du corpus d'entraînement. Cette approche, appelée

clonage de voix, se décline en deux méthodes : l'adaptation au locuteur et l'encodage de locuteurs. La première repose sur une étape d'adaptation, ou *fine-tuning*, du modèle multilocuteur préentraîné afin qu'il ne produise plus que la voix du locuteur cible. L'encodage de locuteurs, quant à lui, ne nécessite pas d'étape de *fine-tuning*. À la place, un second modèle, appelé encodeur de locuteurs, fournit au modèle de synthèse une représentation vectorielle des caractéristiques du locuteur, appelée plongement de locuteur. Ce modèle peut être entraîné conjointement au modèle acoustique, ou séparément (Jia *et al.*, 2018). Lors de la synthèse, le plongement du locuteur cible est nécessaire pour que le modèle acoustique génère de la parole synthétique se rapprochant de sa voix réelle. En cas de changement de locuteur, il suffit de transmettre les échantillons audio de la nouvelle cible à l'encodeur de locuteurs. Ces deux méthodes n'ont besoin que d'une faible quantité de données du locuteur cible. Ainsi, dans l'étude (Chen *et al.*, 2019), elles génèrent de bons résultats à partir de dix secondes de parole, et de très bons à partir de dix minutes.

Dans cette étude, nous utilisons l'approche par encodage de locuteurs. En effet, malgré des résultats de qualité légèrement inférieure à l'approche par adaptation au locuteur (Arik *et al.*, 2018), elle ne nécessite qu'une seule phase d'entraînement et permet donc de généraliser les tests plus facilement à de nouveaux locuteurs. Les modèles choisis pour cette approche sont le modèle x-vecteurs (Snyder *et al.*, 2017) comme encodeur de locuteurs et Tacotron 2 comme modèle acoustique.

Le modèle x-vecteurs est très largement utilisé dans les travaux de vérification du locuteur, et par extension, de clonage de voix. Il est basé sur une architecture en trois blocs. Le premier est un ensemble de couches fonctionnant à l'échelle de la trame pour extraire une représentation de chaque trame et de son contexte. Le deuxième est une agrégation statistique permettant de condenser l'information apportée par chaque trame contextualisée à l'échelle du segment audio entier. Enfin, le dernier bloc est un ensemble de couches fonctionnant à l'échelle du segment et d'où est extrait le plongement de locuteur, appelé x-vecteur.

Le modèle Tacotron 2 est présenté dans la section 2.2.1. Néanmoins, le modèle utilisé ici est multilocuteurs : un plongement de locuteur est concaténé à la sortie de l'encodeur du Tacotron, avant d'être transmis à son décodeur.

2.2.3. *Conversion de voix*

L'objectif de la conversion de voix est de transformer un énoncé produit par un locuteur source, en conservant l'information linguistique, afin que celui-ci soit perçu comme ayant été prononcé par un locuteur cible.

Une étude récente reprend l'évolution des différentes techniques de conversion de voix (Sisman *et al.*, 2021). Historiquement, les premières approches se concentraient sur la modification des caractéristiques spectrales de la voix, par exemple le spectre et les formants, en utilisant des données parallèles, c'est-à-dire des données pour lesquelles les locuteurs source et cible ont prononcé le même contenu. Grâce à un alignement dynamique temporel entre séquences source et cible, il était alors possible

de calculer une fonction de transformation effectuant la correspondance entre les espaces acoustiques des locuteurs source et cible. Les premières approches proposées se fondaient sur la quantification vectorielle (Abe *et al.*, 1990), puis ont évolué vers des approches probabilistes utilisant des mélanges de lois gaussiennes (Toda *et al.*, 2007).

Les travaux de recherche dans le domaine se sont ensuite orientés vers l'apprentissage de fonctions de transformation en utilisant des données non parallèles (Erro *et al.*, 2009 ; Wang *et al.*, 2015). Plus récemment, l'introduction des PPG (*Phonetic PosteriorGrams*) constitue une nouvelle technique permettant de s'affranchir de données parallèles (Sun *et al.*, 2016). Les Phonetic PosteriorGrams (PPG) représentent l'évolution temporelle de la probabilité *a posteriori* des phonèmes (Hazen *et al.*, 2009). Ils capturent le contenu linguistique d'un énoncé tout en gardant l'information temporelle. Généralement, l'extraction des PPG est effectuée par un modèle acoustique multilocuteur permettant d'effacer les composantes liées au locuteur source. Un modèle de synthèse spécifique au locuteur cible peut ensuite être utilisé pour générer un signal de parole synthétique ayant les caractéristiques de ce locuteur. À l'heure actuelle, les techniques utilisant les PPG donnent les meilleurs résultats (Zheng *et al.*, 2020) au Voice Conversion Challenge (VCC). Notamment, lors de l'édition VCC 2020, quatre types de systèmes de conversion pouvaient être distingués : a) ceux combinant la reconnaissance automatique de la parole suivie de synthèse de parole, b) les méthodes basées sur les PPG (Tian *et al.*, 2018 ; Liu *et al.*, 2021), c) les approches de type auto-encodeur (Ho et Akagi, 2020), et d) les approches génératives antagonistes (GAN) (Tobing *et al.*, 2020).

Les méthodes qui combinent la reconnaissance et la synthèse de parole utilisent le texte comme pivot, ce qui permet également d'effacer les caractéristiques du locuteur source. Les performances d'un tel système sont, par nature, dépendantes de la qualité de la reconnaissance de parole.

Dans cet article, nous utilisons un système de conversion de l'état de l'art reposant sur les PPG (Zhao *et al.*, 2019). Le choix de cette approche est motivé par sa capacité à préserver les propriétés temporelles de l'audio source au travers du PPG. De plus, le système reprend l'architecture Tacotron 2 utilisée pour la synthèse vocale monolocuteur, ce qui permet d'avoir une base de comparaison entre les deux approches.

2.3. *Vocodeur*

La dernière étape du processus de synthèse monolocuteur est la transformation du mel-spectrogramme en un flux audio PCM (*Pulse Code Modulation*). Ce travail est opéré par un modèle distinct du modèle acoustique présenté en section 2.2 et entraîné de manière indépendante.

L'état de l'art offre un grand nombre de propositions pour cette tâche. Le modèle neuronal le plus connu est Wavenet (van den Oord *et al.*, 2018), mais des modèles comme SampleRNN (Mehri *et al.*, 2017) ou ParallelWaveGAN (Yamamoto *et al.*,

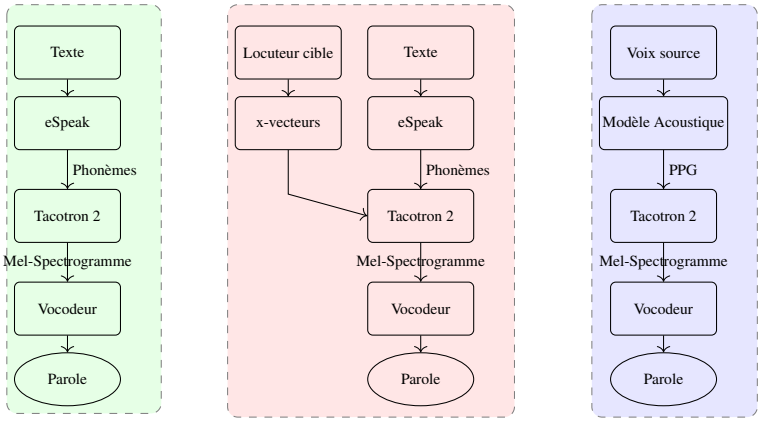

Figure 1. *Architecture des différentes techniques de synthèse vocale : monolocuteur (à gauche), le clonage (au milieu, encodeur identique au premier) et conversion de voix (à droite).*

2020) existent. Dans cette étude, nous utilisons WaveGlow (Prenger *et al.*, 2019). Notre choix s'est porté sur ce vocodeur car il n'est pas autorégressif.

2.4. *Défis propres à la synthèse vocale neuronale*

Les techniques de synthèse vocale neuronale, malgré leurs différences, sont confrontées à des problématiques semblables telles que la disponibilité des données en quantité suffisante et de haute qualité, et une évaluation adaptée à la tâche pour laquelle la technique a été mise en place.

2.4.1. *Données d'entraînement*

Malgré le progrès des techniques de synthèse vocale, la disponibilité des données en quantité suffisante et de haute qualité est souvent considérée comme un prérequis pour l'obtention d'un système de synthèse vocale de l'état de l'art. Ce défi est d'autant plus difficile à relever lorsqu'il s'agit de données issues de langues autres que l'anglais. La plupart des travaux de *benchmark* sur les différentes techniques de synthèse vocale se font ainsi sur des corpus en anglais : LJSpeech (Ito et Johnson, 2017), VCTK (Veaux *et al.*, 2017), LibriTTS (Zen *et al.*, 2019), ARCTIC (Kominek et Black, 2004). Les corpus dédiés à la synthèse vocale en langue française sont, par contraste, peu exploités et relativement peu volumineux. On peut citer FrenchSiwis (Honnet *et al.*, 2017) et SynPaFlex (Sini *et al.*, 2018), tous deux conçus pour la synthèse vocale et contenant des données d'une seule locutrice. On peut aussi évoquer Att-HACK (Le Moine et Obin, 2020) pour la conversion de voix avec plusieurs locuteurs de haute qualité, mais avec une quantité de données par locuteur relativement

limitée. BREF (Lamel *et al.*, 1991) est un autre corpus susceptible d'être utilisé pour l'apprentissage de synthèse vocale multilocuteur ou le clonage de voix, en particulier en raison d'une bonne qualité d'enregistrement. Tous ces corpus peuvent être utilisés pour avoir un système de synthèse vocale de qualité.

Cependant, les systèmes de synthèse vocale dits de bout en bout peuvent-ils utiliser toutes sortes de données, en particulier des données qui n'ont pas été collectées à des fins de synthèse, lorsqu'elles sont présentes en grande quantité ? Exposer les systèmes de synthèse vocale de l'état de l'art à des données amateurs en grande quantité sans préalablement poser de contraintes liées aux caractéristiques du locuteur ou au style semble pertinent afin d'explorer les limites des architectures actuelles sur une langue autre que l'anglais.

2.4.2. *Méthodologie d'évaluation de la parole synthétique*

L'objet de la synthèse vocale étant de produire un stimulus de parole destiné à l'humain, l'évaluation subjective reste incontournable et ce malgré ses défauts biens connus dans la communauté (Wagner *et al.*, 2019). Les raisons principales reposent sur l'absence d'une métrique restituant parfaitement l'opinion de l'humain et sur le faible nombre de travaux dans ce domaine.

Une évaluation adaptée à la tâche de synthèse est l'une des tâches les plus onéreuses du processus expérimental car elle dépend de nombreux facteurs : recrutement des testeurs et vérification de leur niveau linguistique, choix du test audio permettant de répondre à la question de recherche avec le moins de biais possible, considération des conditions d'évaluation, du matériel utilisé lors du test et du temps imparti pour éviter la fatigue des testeurs, optimisation de focalisation de l'effort cognitif des testeurs avec une interface ergonomique.

Dans les challenges VCC (Yi *et al.*, 2020 ; Lorenzo-Trueba *et al.*, 2018) et les éditions du challenge Blizzard (depuis 2005), dédiés respectivement aux techniques de conversion de voix et à la synthèse vocale à partir du texte, des tests de similarité au locuteur, de qualité et d'intelligibilité sont le plus souvent mis en place afin d'évaluer tous les participants. Les tests de qualité et d'intelligibilité sont en général fusionnés pour qualifier la qualité globale, car ils sont intrinsèquement liés.

Le principal outil utilisé pour l'évaluation reste le test MOS (*Mean Opinion Score*) qui agrège les notes attribuées en aveugle par des testeurs entre un minimum de 1 et un maximum de 5 (ITU-T, 1996). Sa déclinaison pour la dégradation par rapport à une référence (DMOS) est également fréquemment employée, tout comme les tests de préférence (A/B) et les évaluations de type MUSHRA. Cette dernière impose cependant une tâche plus astreignante au testeur, tous les stimuli évalués (en aveugle) lui étant présentés en simultané et une notation fine étant exigée pour chacun d'entre eux. En outre, MUSHRA impose l'ajout d'une ancre basse en plus de l'ancre haute (référence) généralement utilisée.

L'évaluation objective, quant à elle, repose sur des métriques utilisées dans les algorithmes pour quantifier la « qualité » du signal synthétique, à savoir : métriques

spectrales, MCD (*Mel-Cepstral Distortion*) et SNR (*Signal Noise Ratio*) ; fréquence fondamentale (F_0, rapport voisement/non-voisement, Likelihood-ratio (LL-Ratio)) ; BAP (*Band Aperiodicity Parameter*) ; durée syllabes/phonèmes. On peut même citer des initiatives reposant sur l'apprentissage profond comme MOSNet, qui visent à estimer des scores de tests subjectifs (Lo *et al.*, 2019), mais ces approches, manquant de maturité, restent marginales. Concernant l'intelligibilité, il est le plus souvent fait usage d'un système de reconnaissance automatique de la parole (Vích *et al.*, 2008). Pour la similarité au locuteur cible, un calcul de distance cosinus avec des plongements locuteurs est souvent employé. Ces méthodes d'évaluation objectives peuvent être de bons indicateurs lors de l'entraînement de systèmes de synthèse vocale, mais elles ne sont pas suffisantes. En outre, un processus d'alignement est parfois nécessaire car le calcul de ces métriques entraîne souvent le passage par des représentations intermédiaires pouvant engendrer des artefacts.

Ceci dit, les méthodes objectives reposent souvent sur une comparaison à une référence absolue. Ceci n'est cependant pas toujours adapté à la tâche d'évaluation. En effet, le fait qu'un échantillon soit différent d'un autre (même d'une référence) n'implique pas nécessairement que celui-ci lui est inférieur. Les évaluations objectives, qui s'appuient sur les systèmes de reconnaissance de la parole ou de vérification du locuteur pour le calcul de l'intelligibilité et de la similarité, ne sont pas meilleures car ces techniques comportent aussi des erreurs de prédiction.

3. Jeu de données

Pour cette étude, nous avons cherché à disposer de lectures enregistrées et accompagnées de leur texte, qui proposent du contenu expressif, et pour une multiplicité de locuteurs. La langue ciblée est le français, et une durée minimale d'environ dix heures est requise pour au moins quelques voix.

De nombreux livres audio enregistrés et partagés par des amateurs sont accessibles aujourd'hui. Ces données présentent des lectures diversifiées, par des donneurs de voix singuliers et relativement libres dans l'interprétation des œuvres de leur choix. C'est pourquoi nous avons choisi d'utiliser le corpus MUFASA (*MUltispeaker French Audiobooks corpus dedicated to expressive read Speech Analysis*) pour nos expériences.

3.1. *Le corpus MUFASA*

Le corpus MUFASA est une base évolutive d'enregistrements de livres audio réalisés par des particuliers, à partir de textes libres de droits, principalement en langue française. Les données sont issues de collectes au format MP3 128 Kbit/s pour l'audio (plus rarement 64 Kbit/s), et aux formats texte et PDF pour les transcriptions. Elles sont progressivement annotées et validées, et leur contenu en français au moment de l'étude est d'environ 600 heures de parole réparties entre vingt locuteurs, dix hommes et dix femmes.

Au cours de certaines lectures, quelques phrases ont été ajoutées au texte de référence pour présenter ou pour clore la section lue. Une partie des textes concernés ont été conformés à la parole. D'autres annotations manuelles ont relevé et documenté, pour certains fichiers audio, des dégradations de qualité ou la présence de sons exogènes à la parole. Les conditions non professionnelles des prises de son favorisent des variations d'intensité et de répartition des fréquences, et sont susceptibles d'enregistrer réverbérations et bruits de fond, ou artefacts liés à la qualité d'encodage ou à la captation. Sont signalés également, quand ils sont repérés, musique ou bruitages insérés intentionnellement par montage.

Les locuteurs ont des profils très divers. Ils sont fidèles aux textes mais assez libres dans les pauses, souvent décorrélées des ponctuations. Quatre d'entre eux ont un accent régional. Seule une voix fait quelques écarts de prononciation et a une prosodie légèrement hésitante. Trois tranches d'âge perçu sont représentées : adulte (10), senior (6) et jeune (4). Les textes sont issus de différents courants littéraires, les plus récents datent du milieu du vingtième siècle. Le genre narratif y est le plus représenté, avec une quantité importante de dialogues dans la plupart des œuvres.

Du point de vue de la prosodie, chaque lecteur a une posture de référence qui lui est propre, et que l'on retrouve dans la narration de façon générale. On peut les classer en trois groupes : onze lecteurs produisent un motif prosodique récurrent, à l'échelle de la phrase. Cinq lisent d'une parole proche du spontané, naturelle. Les quatre autres ont des stratégies plus amples, où les phrases se succèdent de façon contrastée, où le texte est plus incarné. Pour les passages au style direct, également, chaque locuteur organise sa lecture avec plus ou moins de variabilité. Le premier groupe ne marque pas nécessairement d'expressivité ou de personnification (l'abandon du motif narratif fait déjà rupture). Les lecteurs plus naturels dans les passages narratifs marquent une emphase expressive au style direct, et aussi des changements de timbre, subtils le plus souvent. Les lecteurs les plus stratégiques dans la narration sont aussi les plus théâtraux au style direct, avec une nette emphase expressive et des changements de timbre parfois radicaux.

3.2. *Annotation, sélection des données*

Pour cette étude, nous avons extrait du corpus MUFASA un sous-ensemble en langue française d'une durée globale de 222 heures, correspondant à 667 unités audio (généralement des chapitres) de durée très variable : de moins d'une minute à plus de deux heures. Les sous-corpus de chaque locuteur durent d'une centaine de minutes à près de quinze heures (dont huit font plus de dix heures). Des regroupements de chapitres d'une même œuvre ont été préférés à une pioche disparate. En moyenne, deux à trois livres différents sont représentés dans chaque sélection par locuteur, il peut y en avoir jusqu'à huit.

L'annotation automatique consiste en un alignement du texte et de la parole associée à un découpage, sur les silences, en énoncés courts. Pour notre étude, cette

fragmentation a visé à obtenir des unités audio d'une durée inférieure à 10 secondes, une partie présente donc toujours des pauses internes. Tous les textes ont été normalisés et phonétisés au format IPA avec le logiciel *eSpeak*[1]. Des règles ont été dérivées de la validation manuelle d'une partie des segments, puis appliquées aux autres. Cette opération a abouti principalement à exclure les unités correspondant aux débuts et aux fins de chapitre, les plus susceptibles de présenter de la musique. Un dernier traitement permet d'exclure automatiquement les paires signal-texte les moins vraisemblables dans le rapport entre durée audio et nombre de mots, et aussi les unités de plus de 10 secondes qui subsistent. Notre étude porte sur un ensemble correspondant à 161 heures de parole, découpées en énoncés d'une durée moyenne de 4 secondes.

Après nos expérimentations, une expertise acoustique a porté sur les segments audio utilisés, regroupés par enregistrement d'origine. Leur analyse, réalisée à l'aide des outils *open source Audacity* pour les spectres de fréquence et *FreeLCS* pour les intensités en unités LUFS (*Loudness Unit Full Scale*), montre globalement des caractéristiques constantes pour chaque locuteur. Pour six d'entre eux, le signal de parole est de qualité bonne à convenable, six autres présentent des artefacts d'acquisition (résonance médium ou ventilation). La qualité pour les locuteurs restants peut être considérée comme moyenne, elle est irrégulière pour deux d'entre eux. Quelques écarts d'intensité et de qualité sont néanmoins constatés entre les enregistrements et, pour au moins deux voix, la sélection comporte des segments où bruitages et musiques sont superposés à la parole.

3.3. *Sélection des locuteurs cibles*

Le critère principal de sélection des locuteurs cibles est l'existence de traits distinctifs saillants pour chacun d'eux. En outre, les données pour chaque locuteur doivent être disponibles en quantité suffisante, de l'ordre de dix heures de parole au minimum. Notre choix s'est ainsi porté sur *Nadine*, *Jean-Luc*, *René* et *Victoria*.

Les flux de parole de *Nadine* et de *René* sont assez constants, chacun à leur manière : ils présentent un motif rythmique et intonatif récurrent qui pour *Nadine* est emblématique de la narration, et pour *René* relève d'un style personnel très marqué, pittoresque, se déployant aussi au style direct. *Nadine* l'abandonne au style direct pour porter des expressivités plus naturelles, avec quelques changements de timbre. Les stratégies narratives de *Jean-Luc*, par effet d'énigme, et de *Victoria*, au style fantasque, sont plus sophistiquées, et au style direct ces deux lecteurs mettent en œuvre des expressivités exacerbées et des changements de timbre radicaux.

Après nos expériences, une description de l'audio utilisé pour ces locuteurs cibles a porté sur les regroupements des segments par enregistrement. Elle relève une bonne qualité du sous-corpus *Nadine*, quoique perfectible pour des écarts d'intensité (un quart des données sont en excès de 6 dB LUFS sur les autres), et la présence d'artefacts

1. https ://espeak.sourceforge.net/

sur 4 % de sa durée. La voix de *Jean-Luc* est aussi très bien enregistrée, mais 5 à 6 % de ses données sont corrompues par des musiques et bruitages forts, superposés à la parole. Musiques et bruitages apparaissent aussi superposés à la voix de *Victoria*, plus légers mais dans les mêmes proportions, et la qualité globale des enregistrements de cette locutrice est, elle, moyenne et irrégulière (5 % des durées audio sont accompagnées d'une onde parasite). La voix de *René*, quant à elle, est accompagnée d'une onde médium avec plus ou moins de résonance. Les deux dernières voix citées présentent par ailleurs des écarts d'intensité sur un cinquième de leurs données (respectivement - 10 et + 3 dB LUFS). Autres phénomènes, on note dans le sous-corpus de *Jean-Luc* la disparition du timbre pour un personnage qui s'exprime sur 3 % des segments, et l'apparition parcimonieuse d'effets dramatiques sur la voix (réverbération ou jeu d'éloignement spatial). Plus d'informations sur le corpus sont disponibles en ligne [2].

4. Entraînement

4.1. *Synthèse monolocuteur*

Dans cette étude, le modèle acoustique choisi pour la synthèse monolocuteur est Tacotron 2. Nous utilisons l'implémentation d'ESPNET (Hayashi *et al.*, 2020) qui reproduit l'architecture et les hyperparamètres[3] du Tacotron 2 introduits dans l'article originel (Shen *et al.*, 2018). La seule modification apportée à la recette d'ESPNET est le passage du facteur de réduction de 2 à 1, afin d'augmenter la précision des prédictions, au coût d'une convergence plus longue.

L'apprentissage de Tacotron 2 nécessite une grande quantité de données pour chacun des locuteurs cibles. Dans les données décrites section 3.3, nous utilisons la totalité des données disponibles pour les quatre locuteurs cibles. Pour chacun d'entre eux, 200 échantillons de parole sont conservés pour l'évaluation des systèmes, le reste est utilisé pour l'entraînement.

Les échantillons audio sont convertis en mel-spectrogrammes de dimension 80 avec une fenêtre glissante de taille 1024 trames et un décalage de 221 trames. Les silences en début et en fin des échantillons sont supprimés afin de faciliter la convergence du modèle d'attention du modèle acoustique.

Un modèle Tacotron 2 différent est appris pour chacun des locuteurs indépendamment. Cet apprentissage est effectué pendant 200 époques, avec 200 échantillons du jeu d'entraînement mis de côté pour former un jeu de validation. Un mécanisme d'arrêt prématuré est mis en place (*early stopping*) : si la fonction de coût cesse de diminuer sur le jeu de validation pendant 20 époques, l'apprentissage est interrompu pour éviter tout surapprentissage. En pratique, l'apprentissage s'est arrêté automatiquement autour de 70 époques.

2. `https://sites.google.com/view/machahu`
3. `https://shorturl.at/otK59`

4.2. *Clonage de voix*

Comme discuté en section 2.2.2, l'approche choisie pour le clonage de voix est l'encodage de locuteurs. Le système est donc composé de deux modèles : le modèle x-vecteurs en guise d'encodeur de locuteurs et Tacotron 2 pour le modèle acoustique. L'entraînement est effectué en deux étapes. Le modèle x-vecteurs est d'abord entraîné seul, puis utilisé pour l'entraînement du Tacotron 2. Chacune de ces deux étapes est effectuée sur un corpus différent.

L'implémentation pour le modèle de x-vecteurs est ici celle de l'outil Kaldi ASR[4]. Sa dimension d'entrée est de 23. La taille de ses couches intermédiaires, et donc la dimension des x-vecteurs produits, est de 512. Les autres hyperparamètres correspondent à ceux par défaut de la recette. Il nécessite un très grand nombre de locuteurs d'entraînement, mais est moins sensible à la qualité des données que le modèle acoustique. Nous utilisons ici comme corpus d'entraînement la version française du corpus CommonVoice (Mozilla, 2020 ; Ardila *et al.*, 2020), contenant 682 heures de parole (version de décembre 2020). Du fait de la diversité des moyens d'enregistrement et des environnements sonores, la qualité des échantillons est très variable. Comme nous utilisons la version du modèle x-vecteurs indépendante du texte, seuls les échantillons ont été fournis au modèle, sous forme de MFCC (*Mel Frequency Cepstral Coefficients*). L'entraînement a été réalisé en 420 étapes, sur le jeu d'entraînement par défaut, contenant 3605 locuteurs, avec une moyenne de 70 échantillons par locuteur.

Le modèle acoustique utilisé, Tacotron 2, correspond quant à lui à une implémentation d'ESPNET[5], avec les hyperparamètres par défaut de la recette. Il est entraîné sur le corpus MUFASA, présenté en section 3. Contrairement à la synthèse monolocuteur et à la conversion de voix, le principe du clonage implique que les locuteurs cibles soient absents du corpus d'entraînement. Il est donc entraîné sur les données disponibles pour tous les locuteurs, à l'exception de *Nadine*, *Jean-Luc*, *René* et *Victoria*. Le modèle converge après 30 époques d'entraînement.

4.3. *Conversion de voix*

Le système de conversion de voix utilisé ici repose sur (Zhao *et al.*, 2019) et consiste en trois principaux modèles : un modèle d'extraction de PPG, un modèle de conversion des PPG en mel-sprectrogramme (PPG-to-Mel) et un vocodeur. L'architecture globale du système de conversion de voix est présentée dans la figure 1.

Le modèle d'extraction de PPG repose sur le modèle acoustique d'un système de reconnaissance de parole. Dans cette étude, il s'agit d'un modèle TDNN-HMM (*Time Delay Neural Network - Hidden Markov Model*) (Peddinti *et al.*, 2015) préentraîné sur des données multilocuteurs[6].

4. `https://github.com/kaldi-asr/kaldi/tree/master/egs/sre16/v2`
5. `https://github.com/espnet/espnet/tree/master/egs/libritts`
6. `https://github.com/pguyot/zamia-speech`

Le modèle acoustique (*PPG-to-Mel*) est dérivé de Tacotron 2, et a pour objectif de prédire un spectrogramme à partir des PPG. Chaque locuteur cible nécessite un modèle PPG-to-Mel différent. Pendant la phase d'apprentissage, le modèle PPG-to-Mel est appris spécifiquement pour une voix cible. L'apprentissage est basé sur les PPG dérivés du signal audio de la voix cible et est indépendant de la voix source. Pendant la phase d'inférence, les PPG en entrée du modèle sont extraits du signal de parole source et un modèle PPG-to-Mel, spécifique à la voix cible, les convertit en signal audio. Pour le modèle acoustique, le jeu d'entraînement utilisé est identique à celui utilisé pour l'apprentissage du modèle acoustique de synthèse monolocuteur.

4.4. *Vocodeur*

Pour les trois systèmes de synthèse, le vocodeur WaveGLOW, reposant sur des réseaux à flux est utilisé. Sa génération de signal, à partir de mel-spectrogrammes, est rapide, tout en conservant une haute qualité. Dans cette étude, nous utilisons l'implémentation officielle (Nvidia, 2018).

Ainsi, le modèle WaveGLOW préentrainé publié sur le GitHub officiel[7] est adapté sur le corpus MUFASA. Il est entraîné sur le même corpus que le modèle acoustique de clonage, sur environ 40 époques (372 500 batches), sans les locuteurs cibles, pour éviter de corriger en partie la voix produite, et présenter un biais en faveur du clonage en termes de similarité au locuteur. Le modèle obtenu est utilisé pour toutes les expériences.

5. Protocole expérimental

Nous évaluons les performances des systèmes entraînés dans la section précédente selon deux critères : la qualité globale de la synthèse, et la capacité à reproduire fidèlement la voix des locuteurs cibles. Ces deux critères sont évalués à l'aide de mesures objectives automatiques (MCD et similarité cosinus) et de tests perceptifs (un MOS et un DMOS). Le but de ces évaluations est de commencer par identifier la qualité absolue du rendu pour chacune des trois techniques avant de qualifier leur capacité à reproduire l'identité vocale du locuteur souhaité, et donc par extension leur capacité à rester fidèle aux styles parfois très marqués des locuteurs de livres audio.

5.1. *Évaluation objective*

5.1.1. *Métriques*

Afin d'avoir une première estimation de la performance des différents systèmes, nous avons utilisé deux mesures objectives.

7. `https://github.com/NVIDIA/waveglow`

La première, la distortion mel-cepstrale (*Mel-Cepstral Distortion*, MCD) (Kominek *et al.*, 2008), permet d'estimer la qualité globale d'un système. Elle se calcule sur les coefficients mel-cepstraux généralisés (*Mel-Generalized Coefficients*, MGC) de deux signaux audio. Pour les trois paradigmes étudiés, nous appliquons une déformation temporelle dynamique (*Dynamic Time Warping*, DTW) entre le signal synthétique et le signal naturel de référence pour compenser les différences de durée. Les MGC sont extraits à l'aide des outils SPTK[8] et WORLD (Morise *et al.*, 2016).

La seconde mesure objective est la similarité cosinus entre plongements de locuteurs. On l'utilise ici pour évaluer la similarité entre locuteurs naturels et synthétiques. Elle est calculée à partir de plongements extraits d'un encodeur de locuteurs. Pour éviter de biaiser l'évaluation, nous utilisons un encodeur de locuteur différent de celui utilisé dans le modèle de clonage de voix. Dans cette étude, nous utilisons le modèle sur étagère *Resemblyzer*[9] qui implémente l'encodeur de locuteur présenté dans (Wan *et al.*, 2018). Les plongements de locuteurs à comparer sont calculés sur des signaux de parole synthétique et des signaux de parole naturelle correspondant à la même voix.

5.1.2. *Corpus de test*

Pour chacun des quatre locuteurs cibles définis dans la section 3.3, on sélectionne au moins 200 échantillons audio et leur transcription. Ces données n'ont pas été utilisées lors de l'entraînement des modèles. Les énoncés peuvent varier d'un locuteur à un autre. Ces échantillons sont utilisés pour évaluer objectivement la performance du vocodeur, ainsi que celle des systèmes proposés.

Pour évaluer le système de conversion de voix, il est nécessaire de définir la parole d'un locuteur source différent du locuteur cible. On sélectionne donc un second ensemble de couples texte/audio parmi les données n'ayant pas servi à l'entraînement des modèles acoustiques de conversion de voix. Ce choix est limité par les données parallèles à disposition dans le corpus MUFASA. Cette contrainte n'est pas nécessaire pour les évaluations subjectives qui suivront. Pour cette évaluation objective, *Nadine* et *Victoria*, sont locutrices sources l'une de l'autre. Les locuteurs sources de *Jean-Luc* et de *René* sont des femmes, *Victoria* et *Pomme* (autre locutrice de MUFASA) respectivement. *Pomme* est constante dans la narration et modérément expressive au style direct, alors que son timbre est âgé.

Pour le clonage de voix, les échantillons de référence utilisés pour extraire les x-vecteurs ne sont pas compris dans le corpus de test. Ils sont construits à partir d'échantillons du locuteur cible, sélectionnés au hasard dans le corpus d'entraînement des modèles de synthèse monolocuteur et de conversion de voix, présentés section 4. Ces échantillons sont concaténés jusqu'à obtenir un échantillon de dix minutes. Pour chacun des quatre locuteurs cibles, un seul échantillon de dix minutes est produit afin d'extraire un x-vecteur ; ce dernier est utilisé dans les tests objectifs et perceptifs.

8. `https://sp-tk.sourceforge.net/`
9. `https://github.com/resemble-ai/Resemblyzer`

5.2. *Évaluation subjective*

5.2.1. *Test de qualité (MOS)*

Le test MOS permet d'évaluer la qualité générale de la parole. La question posée au testeur est : « Merci d'écouter l'échantillon à évaluer. Comment jugez-vous la qualité générale de la parole dans cet échantillon ? ». Le testeur a comme échelle de notation : très mauvais - mauvais - moyen - bon - très bon, les notes associées allant de 1 à 5. Le test est composé de deux étapes d'introduction, non évaluées, permettant à l'évaluateur de se familiariser avec la parole produite et le processus de notation, suivies de 100 étapes de tests dont les résultats sont enregistrés. Tous les échantillons sont écoutés au casque ou aux écouteurs. Ce test a été réalisé par 19 testeurs, ce qui conduit à 1643 notes, hors étapes d'introduction.

5.2.2. *Test de similarité locuteur (DMOS)*

La question posée au testeur dans le cadre de ce test est « Merci d'écouter d'abord l'échantillon de référence, puis l'échantillon à évaluer. La voix dans l'échantillon à évaluer vous semble-t-elle proche de celle du locuteur de référence ? Il ne s'agit pas de noter la qualité de l'échantillon mais bien l'identité vocale. »Le testeur a comme échelle de notation : très éloigné - éloigné - moyennement proche - proche - très proche, les notes associées allant de 1 à 5. Les contenus textuels des échantillons de référence sont différents de ceux des échantillons à évaluer. Ce test possède le même nombre d'étapes que le précédent et a été réalisé par 13 testeurs (1204 notes, hors étapes d'introduction). Dans les deux tests d'écoute, les évaluateurs sont recrutés parmi des experts et des non-experts non rémunérés. Les conditions d'écoute ne sont pas contrôlées, mais il est demandé d'utiliser des écouteurs ou un casque audio.

5.2.3. *Corpus de test*

Le corpus est constitué pour chaque locuteur de 50 échantillons ne contenant pas de bruit ni de musique dans le naturel et dont la synthèse correspondante n'a pas échoué (synthèse vide, uniquement pour *René* et *Victoria* en clonage). Les échantillons ainsi sélectionnés sont utilisés dans le cadre des évaluations MOS et DMOS décrites précédemment comme échantillons de référence.

Au cours des deux tests, les échantillons évalués pour chaque locuteur sont des échantillons (1) vocodés à partir de mel-spectrogrammes naturels, (2) synthétisés par le modèle monolocuteur, (3) obtenus par clonage de voix en utilisant comme entrée du modèle x-vecteurs un échantillon de dix minutes de parole (voir section 5.1.2), ou (4) convertis à partir de parole du locuteur de même genre dans le jeu de test. On obtient un total de 800 stimuli (4 systèmes, 4 locuteurs, 50 échantillons). Pour le test DMOS de manière spécifique, nous évaluons aussi des faux positifs sous la forme d'échantillons auto-encodés du même genre que le locuteur dans la référence (200 stimuli supplémentaires).

6. Résultats et discussion

6.1. *Résultats*

Les résultats de l'évaluation objective (mesure de la MCD et de la similarité cosinus entre plongements de locuteurs) sont présentés dans les tableaux 1 et 2. Les résultats de l'évaluation subjective (mesure de la qualité générale et de la similarité entre locuteurs) sont présentés dans les tableaux 3 et 4. Des exemples d'échantillons synthétiques sont mis à disposition en ligne[10].

Locuteurs→ Systèmes ↓	**Nadine**	**Victoria**	**Jean-Luc**	**René**	**Moyenne**
Vocodeur	2,56 ± 0,04	2,52 ± 0,09	1,62 ± 0,08	1,67 ± 0,06	2,06 ± 0,04
Synthèse	4,69 ± 0,09	5,46 ± 0,12	4,78 ± 0,18	4,50 ± 0,13	4,79 ± 0,07
Clonage	3,86 ± 0,10	4,78 ± 0,16	4,81 ± 0,15	3,80 ± 0,14	4,20 ± 0,07
Conversion	4,74 ± 0,25	4,64 ± 0,25	3,26 ± 0,34	5,13 ± 0,22	4,53 ± 0,13
Moyenne	3,87 ± 0,08	4,33 ± 0,11	3,62 ± 0,14	3,74 ± 0,10	

Tableau 1. *Distortion mel-cepstrale (± demi-intervalle de confiance à 95 %), en dB.*

Locuteurs → Systèmes ↓	**Nadine**	**Victoria**	**Jean-Luc**	**René**	**Moyenne**
Vocodeur	0,98 ± 0,00	0,98 ± 0,00	0,98 ± 0,00	0,98 ± 0,00	0,98 ± 0,00
Synthèse	0,74 ± 0,01	0,75 ± 0,01	0,75 ± 0,01	0,78 ± 0,01	0,76 ± 0,00
Clonage	0,62 ± 0,01	0,59 ± 0,01	0,64 ± 0,01	0,59 ± 0,01	0,61 ± 0,01
Conversion	0,75 ± 0,02	0,76 ± 0,01	0,52 ± 0,01	0,48 ± 0,01	0,59 ± 0,01
Moyenne	0,77 ± 0,01	0,77 ± 0,01	0,72 ± 0,01	0,71 ± 0,01	

Tableau 2. *Similarité cosinus entre locuteurs (± demi-intervalle de confiance à 95 %).*

Locuteurs → Systèmes ↓	**Nadine**	**Victoria**	**Jean-Luc**	**René**	**Moyenne**
Vocodeur	4,7 ± 0,1	4,1 ± 0,2	4,4 ± 0,2	3,9 ± 0,2	4,2 ± 0,1
Synthèse	3,7 ± 0,2	2,7 ± 0,2	1,9 ± 0,2	2,5 ± 0,2	2,7 ± 0,1
Clonage	3,0 ± 0,2	2,6 ± 0,2	2,3 ± 0,2	1,7 ± 0,1	2,4 ± 0,1
Conversion	3,2 ± 0,2	2,9 ± 0,2	1,7 ± 0,1	2,6 ± 0,2	2,6 ± 0,1
Moyenne	3,6 ± 0,1	3,1 ± 0,1	2,6 ± 0,1	2,7 ± 0,1	

Tableau 3. *Scores MOS moyens en fonction des systèmes et des locuteurs (± demi-intervalle de confiance à 95 %)*

10. `https://sites.google.com/view/machahu`

Locuteurs $\rightarrow$ **Systèmes** $\downarrow$	**Nadine**	**Victoria**	**Jean-Luc**	**René**	**Moyenne**
Vocodeur	4,3 $\pm$ 0,3	4,3 $\pm$ 0,3	4,3 $\pm$ 0,3	4,9 $\pm$ 0,2	4,5 $\pm$ 0,1
Synthèse	3,9 $\pm$ 0,3	3,8 $\pm$ 0,3	3,0 $\pm$ 0,3	4,3 $\pm$ 0,3	3,7 $\pm$ 0,2
Clonage	3,1 $\pm$ 0,3	2,0 $\pm$ 0,2	2,0 $\pm$ 0,2	1,8 $\pm$ 0,3	2,2 $\pm$ 0,1
Conversion	3,6 $\pm$ 0,3	3,9 $\pm$ 0,3	2,4 $\pm$ 0,3	4,5 $\pm$ 0,2	3,6 $\pm$ 0,2
Moyenne	3,7 $\pm$ 0,2	3,5 $\pm$ 0,2	3,0 $\pm$ 0,2	3,9 $\pm$ 0,2	

Tableau 4. *Scores DMOS moyens en fonction des systèmes et des locuteurs ($\pm$ demi-intervalle de confiance à 95 %)*

Les mesures obtenues pour le vocodeur nous renseignent sur l'impact du vocodeur sur la qualité générale lorsqu'il sera intégré à la chaîne de traitement pour la synthèse, la conversion et le clonage. Ce système représente la meilleure qualité atteignable par le vocodeur et donc la borne haute de ce qu'il est possible de reproduire à l'aide du modèle acoustique. Le système obtient une MCD moyenne de 2,06 dB sur les quatre locuteurs, ce qui est comparable à l'état de l'art (Hsu et Lee, 2020). Lors du test d'écoute, son score MOS est 4,2, ce qui peut paraître légèrement bas en comparaison avec l'état de l'art. Ceci n'est pas surprenant puisqu'il s'agit ici d'un vocodeur multilocuteur entraîné sur des données issues de livres audio non professionnels. De plus, ce score varie énormément en fonction du locuteur. Le vocodeur obtient un score MOS de 4,7 pour *Nadine*, ce qui laisse présager de très bons résultats des systèmes synthétiques pour cette voix. Il obtient un score MOS plus bas pour *Jean-Luc* (4,4) et des scores décevants pour les voix de *René* et de *Victoria* (autour de 4).

Le vocodeur obtient une similarité cosinus de 0,98 pour l'ensemble des locuteurs et un score DMOS de 4,5. Cela suggère que notre vocodeur est capable de reproduire fidèlement la voix des locuteurs de test. De manière surprenante, le vocodeur obtient un score DMOS identique pour les voix de *Nadine*, de *Victoria* et de *Jean-Luc* mais *René* obtient une note significativement plus élevée. Cela peut s'expliquer par le fait que le caractère atypique de la voix de *René* et de son élocution est si accentué qu'il favorise grandement sa reconnaissance par les testeurs.

En moyenne, le système de synthèse monolocuteur voit une augmentation significative de la MCD et une diminution significative du MOS par rapport au vocodeur. Cela correspond à la diminution de la qualité globale de la synthèse, attendue en raison des erreurs de prédictions introduites par le modèle acoustique. Bien que la MCD de 4,79 dB ne soit pas surprenante, un score MOS de 2,7 est en deçà de l'état de l'art (Weiss *et al.*, 2021). Cependant, celui-ci varie d'un locuteur à l'autre. *Nadine* obtient un score MOS de 3,7 comparable à l'état de l'art pour des données similaires en anglais (Zen *et al.*, 2019). En revanche, *Victoria*, *Jean-Luc* et *René* obtiennent des scores MOS inférieurs à 3. Cela peut s'expliquer par le caractère dégradé de leurs données. L'expressivité n'étant pas modélisée explicitement, la qualité de l'apprentissage sur les voix atypiques de *René* et de *Victoria* est probablement entravée. Il est surprenant

que la synthèse monolocuteur obtienne les meilleurs scores MOS malgré une mauvaise MCD. Comme dans Weiss *et al.* (2021), les systèmes présentant les meilleurs scores MOS n'obtiennent pas toujours les meilleures MCD, probablement à cause de l'alignement DTW.

Le système de synthèse monolocuteur subit aussi une diminution significative de la similarité cosinus et du score DMOS par rapport au vocodeur. Cela traduit une diminution globale de la fidélité de la reproduction d'une voix due au modèle acoustique. La similarité cosinus moyenne est élevée (0,76) et varie peu d'un locuteur à l'autre (celle de *René* restant légèrement supérieure). Le système de synthèse obtient un score DMOS moyen de 3,7 avec une plus grande variabilité en fonction du locuteur. Il n'y a pas de différence significative entre *Nadine*, *Victoria* et *René*. *Jean-Luc* obtient cependant une note significativement plus basse. Il est intéressant de noter que cette voix synthétique a aussi obtenu le moins bon score MOS. Il y a probablement une corrélation entre le score MOS pour la qualité globale et le DMOS pour la similarité au locuteur. En effet, il est difficile de noter la similarité d'un échantillon à un locuteur cible lorsque la qualité globale de cet échantillon est mauvaise. Ces mesures suggèrent que la synthèse monolocuteur est capable de reproduire des voix, même expressives, tant que la quantité de données pour le locuteur est suffisante. Cependant, les résultats pour *Jean-Luc* suggèrent qu'avoir des données de qualité reste important.

Comme la synthèse monolocuteur, le clonage de voix montre une diminution de la qualité globale par rapport au vocodeur qui peut être observée par une dégradation de la MCD et du score MOS. En moyenne, le clonage est aussi significativement moins bon que la synthèse monolocuteur en termes de MOS. Ceci n'est pas surprenant puisque le clonage est une tâche plus complexe que la synthèse monolocuteur. Il est cependant intéressant de noter que la tendance est inversée dans le cas de *Jean-Luc*. Le clonage pourrait ainsi s'avérer intéressant dans les cas où les données du locuteur cible sont perturbées de façon ponctuelle.

On observe aussi une diminution significative des performances du système de clonage en ce qui concerne la fidélité de la reproduction des voix des locuteurs cibles en comparaison avec le vocodeur et la synthèse monolocuteur. Le faible score de similarité cosinus (0,6 en moyenne) est confirmé par les résultats du test d'écoute DMOS (2,2 en moyenne). À part pour *Nadine*, les testeurs ne semblent pas avoir été capables de reconnaître la voix des locuteurs cibles. Bien qu'en théorie, le clonage de voix permette de reproduire fidèlement la voix de locuteurs non vus lors de l'apprentissage, sa capacité de généralisation ne semble pas bonne sur des données de livres audio amateurs français. Trois causes seraient possibles : la qualité générale des données (enregistrement amateur), l'expressivité des locuteurs (lecture de contenu non neutre), le nombre limité de locuteurs disponibles dans cette première version du corpus.

Le système de conversion de voix subit lui aussi une dégradation de la qualité générale par rapport au vocodeur, mais ne présente pas de différence significative avec le système de synthèse monolocuteur en termes de MOS moyen. Ceci n'est pas surprenant puisque les modèles acoustiques de ces deux systèmes ont été entraînés sur les mêmes quantités de données des locuteurs cibles. Il n'y a pas non plus de différence

significative entre la conversion de voix et le clonage de voix en termes de MOS. Cela suggère que la conversion de voix n'apporte pas d'amélioration ou de dégradation significative par rapport aux deux autres paradigmes pour la qualité globale.

Enfin, la conversion de voix présente une dégradation globale par rapport au vocodeur en ce qui concerne la fidélité de la reproduction de la voix des locuteurs cibles. Relativement à la similarité cosinus moyenne, la conversion de voix est inférieure à la synthèse monolocuteur et comparable au clonage de voix. Cela dit, locuteur par locuteur, les voix de *Nadine* et de *Victoria* obtiennent des scores similaires à la synthèse monolocuteur, et les voix de *Jean-Luc* et de *René* obtiennent des performances significativement dégradées par rapport à tous les autres systèmes. En moyenne et locuteur par locuteur, le score DMOS de la conversion de voix n'est pas significativement distinguable de celui pour la synthèse monolocuteur. Bien que le score DMOS moyen de la conversion de voix soit significativement supérieur à celui du clonage de voix (3,6 et 3,4 respectivement), les différences ne sont significatives que pour la moitié des locuteurs. Ces observations suggèrent que la conversion de voix n'apporte pas non plus d'amélioration ou de dégradation significative pour la similarité au locuteur par rapport aux deux autres paradigmes.

6.2. *Discussion*

Locuteur	**Jitter**	**Shimmer**	F_0 (Hz)	F_0 **min**	F_0 **max**	**HNR**
Nadine	2,54	1,10	$187{,}27 \pm 41$	90,87	503,25	8,03
Victoria	2,39	1,07	$199{,}59 \pm 51$	89,39	507,16	7,09
René	3,68	1,44	$145{,}85 \pm 34$	44,90	299,91	3,92
Jean-Luc	4,37	1,39	$116{,}91 \pm 37$	43,28	311,53	3,18

Tableau 5. *Mesures de Jitter (%), Shimmer (dB), F_0 (moyenne avec écart-type, minimum et maximum) ainsi que le rapport signal à bruit (HNR, dB) pour les 4 locuteurs de test. Jitter, Shimmer et HNR sont calculés avec OpenSmile. Les attributs liés au F_0 sont calculés avec Praat.*

L'objectif de cette étude était d'entraîner des modèles de synthèse de parole sur des données du tout-venant issues de livres audio amateurs. L'état de l'art en matière de synthèse monolocuteur montre que cette technologie est capable de produire de la parole de qualité reproduisant fidèlement la voix de locuteurs. En revanche, ces études sont souvent réalisées sur des données enregistrées spécifiquement pour entraîner un système de synthèse de parole. Dans ces travaux, nous avons montré que l'approche monolocuteur est sensible aux données d'apprentissage. Pour la voix de *Nadine*, le système appris obtient des résultats similaires à ceux de l'état de l'art alors que les systèmes appris sur les trois autres voix obtiennent des résultats bien inférieurs. Il convient alors de se poser la question de l'origine de la variabilité au sein de nos résultats. La quantité de données n'explique pas cette variabilité, puisque chaque locuteur dispose d'un nombre d'heures de parole comparable. La qualité des données

en revanche semble être un facteur important. En effet, les échantillons de *René* et de *Jean-Luc* sont plus bruités (HNR, tableau 5) que ceux de *Nadine* et mènent à des systèmes moins performants. Cependant, *Victoria* donne des résultats significativement moins bons que ceux de *Nadine* malgré un HNR proche. Il reste donc un autre facteur impactant la qualité de l'apprentissage. Nous pensons qu'il s'agit de l'expressivité. Ainsi, la prosodie de *Victoria* est moins régulière et plus stratégique dans la narration que celle de *Nadine*, elle est aussi plus expressive au style direct. Malheureusement, ces aspects de la parole sont difficilement quantifiables et mesurables.

Ces remarques à propos de l'impact de la qualité des données s'appliquent aussi au clonage et à la conversion de voix. Ainsi, les méthodes présentes dans l'état de l'art ne sont pas actuellement applicables à tous les types de données. Pour pallier ce défaut, deux pistes sont possibles. La première consiste à améliorer les modèles acoustiques pour les rendre plus robustes aux bruits et à l'expressivité présente dans les données d'apprentissage. Par exemple, la modélisation à l'aide d'un auto-encodeur variationel pourrait être une option prometteuse. La seconde piste consiste à travailler sur les données, en mettant en place des procédures de sélection automatique en fonction de leur qualité.

Le clonage de voix est soumis à un impératif supplémentaire en termes de données. Les travaux de la communauté montrent qu'un système de clonage est capable de reproduire n'importe quelle voix à partir de courts échantillons, lorsqu'ils sont entraînés sur une (ou plusieurs) centaine(s) de locuteurs. Le français ne dispose malheureusement pas, à notre connaissance, d'un jeu de données libre de droit contenant de la parole de qualité provenant d'autant de locuteurs et les résultats obtenus dans cet article montrent que la vingtaine de locuteurs présents dans le corpus MUFASA n'est pas un nombre suffisant pour généraliser à n'importe quelle voix de locuteur inconnu. Une solution naïve consisterait à ajouter plus de locuteurs au corpus MUFASA. Cependant, dans le cas de données du tout-venant, bien que de nouvelles données soient faciles à trouver, leur qualité reste un problème crucial et non trivial comme discuté précédemment. Une autre solution serait de travailler sur des modèles multilingues pour tirer profit des larges jeux de données disponibles en anglais.

Enfin, il est intéressant de noter que la conversion de voix obtient des résultats similaires à la synthèse monolocuteur malgré son cas d'usage différent. Cette observation n'est pas surprenante si l'on s'en tient au fait que les deux paradigmes ont été entraînés sur les mêmes données (mêmes locuteurs, même quantité). Cela indique cependant que, dans notre cas, l'impact de la qualité des données est plus important que celui des modalités d'entrée. Nous n'avons pas mesuré de différences significatives entre l'utilisation de séquences phonétiques issues du texte d'une part et l'utilisation de PPG issus de l'audio d'autre part, en tant qu'entrée d'un modèle acoustique. Bien que les PPG encodent des informations prosodiques en plus des informations phonétiques, ces informations supplémentaires n'ont pas eu d'impact, positif ou négatif, sur les performances du système.

7. Conclusion

La synthèse de la parole est en règle générale effectuée à partir de corpus de données de qualité construits spécifiquement pour la tâche de synthèse. La mitigation de cette contrainte sur la qualité et sur l'adéquation des données d'entraînement utilisées en synthèse vocale déverrouillerait un grand potentiel. Ceci dit, il convient avant tout de faire un état des lieux des performances actuelles avec de telles données. Dans cette étude, nous nous sommes focalisés sur le cas de la langue française.

Nous avons évalué la capacité de trois paradigmes de synthèse de parole à produire des livres audio à partir de données du même type. La synthèse monolocuteur est capable de reproduire fidèlement la voix du locuteur d'entraînement mais la qualité globale de la synthèse est grandement impactée par le niveau d'expressivité de ce même locuteur. Le clonage de voix propose une approche intéressante pour baisser le coût de la production de la synthèse vocale. Cependant, ce paradigme souffre du même défaut sur la qualité globale et présente une forte complexité pour produire un modèle acoustique capable de restituer fidèlement la voix du locuteur cible. Enfin, la conversion de voix semble une piste prometteuse car offrant des performances proches de la synthèse monolocuteur.

De futurs travaux sont à mener pour améliorer la qualité des systèmes. Tout d'abord, l'amélioration du procédé de sélection des données au sein des corpus de synthèse reste une étape importante pour une meilleure maîtrise des propriétés de la voix synthétique. On peut également noter l'intérêt de raffiner les procédures automatiques naissantes dans la communauté pour sélectionner les meilleures données parmi une large quantité disponible. Un autre axe d'étude est la modélisation et un contrôle explicite de l'expressivité. Cela peut se faire par des contraintes explicites lors de la construction du modèle acoustique. L'intégration de méthodes génératives capables de tirer parti des spécificités des données à plusieurs échelles est une solution potentielle.

Remerciements

Ces travaux ont bénéficié d'un accès aux moyens de calcul de l'IDRIS au travers de l'allocation de ressources 2023-AD011011870R2 attribuée par GENCI.

8. Bibliographie

Abe M., Nakamura S., Shikano K., Kuwabara H., « Voice conversion through vector quantization », *Journal of the Acoustical Society of Japan (E)*, 1990.

Ardila R., Branson M., Davis K., Kohler M., Meyer J., Henretty M., Morais R., Saunders L., Tyers F., Weber G., « Common Voice : A Massively-Multilingual Speech Corpus », *LREC*, 2020.

Arik S., Chen J., Peng K., Ping W., Zhou Y., « Neural voice cloning with a few samples », *Advances in Neural Information Processing Systems : Annual Conf. on Neural Information Processing Systems*, 2018.

Arık S. Ö., Chrzanowski M., Coates A., Diamos G., Gibiansky A., Kang Y., Li X., Miller J., Ng A., Raiman J. *et al.*, « Deep voice : Real-time neural text-to-speech », *Int. Conf. on Machine Learning*, 2017.

Chen Y., Assael Y., Shillingford B., Budden D., Reed S., Zen H., Wang Q., Cobo L. C., Trask A., Laurie B., Gulcehre C., van den Oord A., Vinyals O., de Freitas N., « Sample Efficient Adaptive Text-to-Speech », *Int. Conf. on Learning Representations*, 2019.

Elias I., Zen H., Shen J., Zhang Y., Jia Y., Weiss R. J., Wu Y., « Parallel tacotron : Non-autoregressive and controllable tts », *ICASSP*, 2021.

Erro D., Moreno A., Bonafonte A., « INCA algorithm for training voice conversion systems from nonparallel corpora », *IEEE Tr. on Audio, Speech, and Language Processing*, 2009.

Hayashi T., Yamamoto R., Inoue K., Yoshimura T., Watanabe S., Toda T., Takeda K., Zhang Y., Tan X., « Espnet-TTS : Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit », *ICASSP*, 2020.

Hazen T. J., Shen W., White C., « Query-by-example spoken term detection using phonetic posteriorgram templates », *IEEE Workshop on Automatic Speech Recognition Understanding*, 2009.

Hinterleitner F., Manolaina C., Möller S., « Influence of a voice on the quality of synthesized speech », *2014 Sixth International Workshop on Quality of Multimedia Experience*, 2014.

Ho T. V., Akagi M., « Non-parallel Voice Conversion based on Hierarchical Latent Embedding Vector Quantized Variational Autoencoder », *Blizzard Challenge Workshop*, 2020.

Honnet P.-E., Lazaridis A., Garner P. N., Yamagishi J., The siwis french speech synthesis database ? design and recording of a high quality french database for speech synthesis, Technical report, Idiap, 2017.

Hsu P.-c., Lee H.-y., « WG-WaveNet : Real-Time High-Fidelity Speech Synthesis Without GPU », *Interspeech*, 2020.

Ito K., Johnson L., « The LJ Speech Dataset », , https ://keithito.com/LJ-Speech-Dataset/, 2017.

ITU-T, ITU-T Recommendation P.800, Technical report, International Telecommunication Union, 1996.

Jia Y., Zhang Y., Weiss R., Wang Q., Shen J., Ren F., Chen Z., Nguyen P., Pang R., Moreno I., Wu Y., « Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis », *Neural Information Processing Systems Conf.*, 2018.

Kominek J., Black A. W., « The CMU Arctic speech databases », *SSW 5*, 2004.

Kominek J., Schultz T., Black A. W., « Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion. », *SLTU*, 2008.

Lamel L. F., Gauvain J.-L., Eskénazi M., « Bref, a large vocabulary spoken corpus for french », *Eurospeech*, 1991.

Le Moine C., Obin N., « Att-HACK : An Expressive Speech Database with Social Attitudes », *Speech Prosody*, 2020.

Ling Z.-H., Zhou X., King S., « The Blizzard Challenge 2021 », *Blizzard Challenge Workshop*, 2021.

Liu S., Cao Y., Wang D., Wu X., Liu X., Meng H., « Any-to-Many Voice Conversion With Location-Relative Sequence-to-Sequence Modeling », *IEEE/ACM Tr. on Audio, Speech and Language Processing*, 2021.

Lo C.-C., Fu S.-W., Huang W.-C., Wang X., Yamagishi J., Tsao Y., Wang H.-M., « MOSNet : Deep Learning based Objective Assessment for Voice Conversion », *Interspeech*, 2019.

Lorenzo-Trueba J., Yamagishi J., Toda T., Saito D., Villavicencio F., Kinnunen T., Ling Z., « The Voice Conversion Challenge 2018 : Promoting Development of Parallel and Nonparallel Methods », *Speaker Odyssey 2018*, ISCA, p. 195-202, June, 2018.

Mehri S., Kumar K., Gulrajani I., Kumar R., Jain S., Sotelo J., Courville A. C., Bengio Y., « SampleRNN : An Unconditional End-to-End Neural Audio Generation Model », *5th International Conference on Learning Representations, ICLR 2017, Toulon, France*, 2017.

Morise M., Yokomori F., Ozawa K., « WORLD : a vocoder-based high-quality speech synthesis system for real-time applications », *IEICE Tr. on Information and Systems*, 2016.

Mozilla, « CommonVoice », , `https://commonvoice.mozilla.org/`, 2020.

Nvidia, « Waveglow Github repository », , `https://github.com/NVIDIA/waveglow/`, 2018.

Panayotov V., Chen G., Povey D., Khudanpur S., « Librispeech : An ASR corpus based on public domain audio books », *ICASSP*, 2015.

Peddinti V., Povey D., Khudanpur S., « A time delay neural network architecture for efficient modeling of long temporal contexts », *Interspeech*, 2015.

Ping W., Peng K., Chen J., « ClariNet : Parallel Wave Generation in End-to-End Text-to-Speech », *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, OpenReview.net, 2019.

Ping W., Peng K., Gibiansky A., Arik S. O., Kannan A., Narang S., Raiman J., Miller J., « Deep Voice 3 : 2000-Speaker Neural Text-to-Speech », *Int. Conf. on Learning Representations*, 2018.

Prenger R., Valle R., Catanzaro B., « Waveglow : A flow-based generative network for speech synthesis », *ICASSP*, 2019.

Ren Y., Hu C., Tan X., Qin T., Zhao S., Zhao Z., Liu T.-Y., « FastSpeech 2 : Fast and High-Quality End-to-End Text to Speech », *Int. Conf. on Learning Representations*, 2020.

Shen J., Pang R., Weiss R. J., Schuster M., Jaitly N., Yang Z., Chen Z., Zhang Y., Wang Y., Skerrv-Ryan R., Saurous R. A., Agiomvrgiannakis Y., Wu Y., « Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions », *ICASSP*, 2018.

Sini A., Lolive D., Vidal G., Tahon M., Delais-Roussarie É., « SynPaFlex-Corpus : An Expressive French Audiobooks Corpus dedicated to expressive speech synthesis. », *LREC*, 2018.

Sisman B., Yamagishi J., King S., Li H., « An Overview of Voice Conversion and Its Challenges : From Statistical Modeling to Deep Learning », *IEEE/ACM Tr. on Audio, Speech, and Language Processing*, 2021.

Snyder D., Garcia-Romero D., Povey D., Khudanpur S., « Deep Neural Network Embeddings for Text-Independent Speaker Verification », *Interspeech*, 2017.

Sun L., Li K., Wang H., Kang S., Meng H., « Phonetic posteriorgrams for many-to-one voice conversion without parallel data training », *IEEE Int. Conf. on Multimedia and Expo*, 2016.

Tan X., Qin T., Soong F., Liu T.-Y., « A Survey on Neural Speech Synthesis », *arXiv preprint arXiv :2106.15561v3*, 2021.

Tian X., Wang J., Xu H., Chng E.-S., Li H., « Average Modeling Approach to Voice Conversion with Non-Parallel Data », *Speaker and Language Recognition Workshop (Odyssey)*, 2018.

Tobing P. L., Wu Y.-C., Toda T., « Baseline System of Voice Conversion Challenge 2020 with Cyclic Variational Autoencoder and Parallel WaveGAN », *Blizzard Challenge Workshop*, 2020.

Toda T., Black A. W., Tokuda K., « Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory », *IEEE Tr. on Audio, Speech, and Language Processing*, 2007.

van den Oord A., Li Y., Babuschkin I., Simonyan K., Vinyals O., Kavukcuoglu K., Driessche G., Lockhart E., Cobo L., Stimberg F. *et al.*, « Parallel wavenet : Fast high-fidelity speech synthesis », *Int. Conf. on Machine Learning*, 2018.

Veaux C., Yamagishi J., MacDonald K., CSTR VCTK corpus : English multi-speaker corpus for CSTR voice cloning toolkit, Technical report, University of Edinburgh. CSTR, 2017.

Vích R., Nouza J., Vondra M., « Automatic speech recognition used for intelligibility assessment of text-to-speech systems », *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction*, Springer, p. 136-148, 2008.

Wagner P., Beskow J., Betz S., Edlund J., Gustafson J., Eje Henter G., Le Maguer S., Malisz Z., Székely E., Tånnander C., Voße J., « Speech Synthesis Evaluation — State-of-the-Art Assessment and Suggestion for a Novel Research Program », *SSW 10*, 2019.

Wan L., Wang Q., Papir A., Moreno I. L., « Generalized end-to-end loss for speaker verification », *ICASSP*, 2018.

Wang H., Soong F., Meng H., « Aa spectral space warping approach to cross-lingual voice transformation in hmm-based tts », *ICASSP*, 2015.

Weiss R. J., Skerry-Ryan R., Battenberg E., Mariooryad S., Kingma D. P., « Wave-tacotron : Spectrogram-free end-to-end text-to-speech synthesis », *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, p. 5679-5683, 2021.

Yamamoto R., Song E., Kim J.-M., « Parallel WaveGAN : A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram », *ICASSP*, 2020.

Yi Z., Huang W.-C., Tian X., Yamagishi J., Das R. K., Kinnunen T., Ling Z.-H., Toda T., « Voice Conversion Challenge 2020 –- Intra-lingual semi-parallel and cross-lingual voice conversion –- », *Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, ISCA, oct, 2020.

Zen H., Dang V., Clark R., Zhang Y., Weiss R. J., Jia Y., Chen Z., Wu Y., « LibriTTS : A Corpus Derived from LibriSpeech for Text-to-Speech », *Interspeech*, 2019.

Zhao G., Ding S., Gutierrez-Osuna R., « Foreign Accent Conversion by Synthesizing Speech from Phonetic Posteriorgrams », *Interspeech*, 2019.

Zheng L., Tao J., Wen Z., Zhong R., « CASIA Voice Conversion System for the Voice Conversion Challenge 2020 », *Blizzard Challenge Workshop*, 2020.

Avatar signeur – Synthèse de la langue des signes française à partir de texte

Sylvie Gibet

IRISA, Université Bretagne Sud
Campus de Tohannic, rue Yves Mainguy
F-56017 Vannes cedex
sylvie.gibet@univ-ubs.fr

RÉSUMÉ. *Nous présentons dans cet article un système de synthèse de mouvement multimodal qui produit de la langue des signes française (LSF) à partir d'énoncés textuels au moyen d'un personnage virtuel 3D appelé avatar signeur. Notre système SignCom s'appuie sur un principe de composition multicanale, chaque canal d'information étant associé à une information linguistique (depuis le niveau phonologique vers les niveaux lexical, syntaxique ou sémantique) ou articulatoire. La composition permet un agencement spatio-temporel de ces éléments qui s'exécutent en parallèle, et donne la possibilité d'éditer et de générer des phrases en LSF. Les nouveaux modules de synthèse qui enrichissent le système initial sont décrits. Ils incluent la synthèse de mouvement corporel et des mains ainsi que la synthèse faciale, et mettent en œuvre des dynamiques grammaticales propres à la LSF, en s'appuyant sur les concepts fondamentaux de spatialité et d'iconicité. Enfin, nous présentons les principaux défis technologiques qui restent à relever avant de conclure.*

MOTS-CLÉS : *langues des signes, synthèse texte-vers-LS, avatar signeur*

ABSTRACT. *In this paper, we present a multimodal synthesis system that translates text-to-LSF (French sign language) by means of a 3D virtual character, also called virtual signer. Our SignCom system is based on a multichannel composition mechanism, each channel being associated to linguistic information (from the phonological level to the lexical, syntactic or semantic levels), or to articulatory information. The composition is based on a spatio-temporal arrangement of these elements that are parallelized, and is able to edit and generate utterances in LSF. The new synthesis modules that enrich the initial system are described, including body movement synthesis, facial and hand movement synthesis. They implement grammatical dynamics specific to sign language, based on the fundamental concepts of spatiality and iconicity. Finally, we present the main technological challenges that remain before concluding.*

KEYWORDS: *Sign languages, Text-to-SL synthesis, Signing avatar*

1. Introduction

Les langues des signes repoussent les frontières habituelles des théories linguistiques associées aux langues vocales. Ceci est principalement dû au fait qu'elles utilisent l'information visuelle et gestuelle, contrairement aux langues vocales qui utilisent le canal audio-oral. Ainsi, les personnes sourdes développent avec la pratique de cette langue une dextérité dans leur gestuelle et dans leur perception visuelle, une acuité de représentation de l'espace et une capacité d'expression qui se manifestent à travers les différentes modalités[1] de communication propres aux langues gestuelles, incluant les mouvements des mains, les mouvements corporels non manuels, les mimiques faciales, la direction du regard et la labialisation éventuellement associée au son. C'est pourquoi ces langues peuvent être qualifiées de multimodales.

Cette spécificité de la gestualité s'accompagne de mécanismes iconiques et spatiaux omniprésents dans les langues des signes. L'iconicité met en jeu des processus par lesquels le locuteur décrit l'expérience vécue, imaginée ou exécutée en s'inspirant de représentations imagées ou mimétiques (Cuxac, 2000). Elle est caractérisée par le lien de ressemblance plus ou moins étroit entre les entités du monde réel, le référent et le signe qui s'y rapporte. Cuxac propose ainsi une théorie de l'iconicité dans laquelle plusieurs structures linguistiques se combinent lors d'activités discursives : les structures dites de grande iconicité à visée illustrative et les structures dites standard (dans leur forme de citation) sans visée illustrative, comprenant des unités lexicales, de pointage ou des unités dactylologiques (Sallandre et Garcia, 2020). Millet propose une grammaire descriptive de la langue des signes française (LSF) qui s'appuie également sur la spatialité et l'iconicité structurant à tous les niveaux (phonologique, lexical, syntaxico-sémantique) cette langue (Millet, 2019).

En nous appuyant sur ces théories linguistiques de la LSF, nous nous intéressons aux outils numériques à destination des personnes sourdes signantes qui permettent de produire automatiquement des contenus en langue des signes (LS). À l'heure actuelle, la plupart des applications disponibles reposent sur de la vidéo. Or, si la vidéo est le média le plus partagé par les sourds, elle ne permet pas de garantir l'anonymat et impose des contraintes fortes au niveau du stockage et du transport d'information. En contrepartie, la production automatique de contenus en LS et la visualisation au moyen d'avatars signeurs 3D constituent une réponse alternative appropriée et permettent à la fois la réduction des informations stockées, l'anonymisation ainsi que la manipulation des informations pour éditer, visualiser et produire à moindre coût de nouveaux énoncés.

Nous nous focalisons ici sur les systèmes de génération, de synthèse et de traduction texte-vers-LS (que nous regrouperons sous le sigle TSL[2]) et qui incorporent des avatars signeurs. Avec les avancées significatives du traitement automatique des langues parlées, ces systèmes TSL connaissent un regain d'intérêt ces dernières années. Dans les LS, la traduction peut être réalisée en deux étapes, comme illustré dans la figure 1. La première transforme le français écrit en un langage pivot intermédiaire

1. Terme employé dans le domaine des agents conversationnels animés.
2. *Text to Sign Language.*

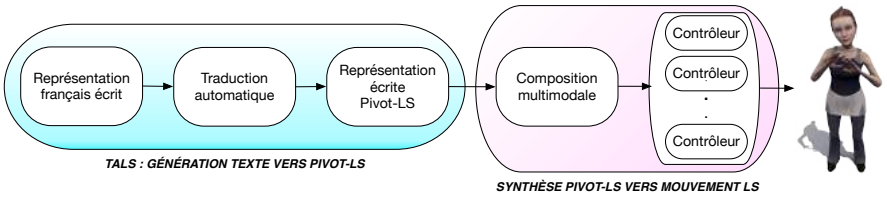

FIGURE 1. *Traduction texte-vers-LS en deux étapes : 1) Génération texte vers Pivot-LS, 2) Synthèse Pivot-LS vers mouvement en LS*

(appelé *Pivot-LS*) qui tient compte des spécificités des LS (partie *TALS* : traitement automatique des langue des signes). La seconde permet de passer de la spécification symbolique exprimée dans ce langage pivot à la production d'un flux continu de postures de l'avatar, au moyen d'un système de synthèse comprenant un procédé de composition multimodale et un ensemble de contrôleurs de mouvement.

Bien que les travaux linguistiques sur les LS aient permis de mieux appréhender les mécanismes grammaticaux en jeu (Millet, 2019), la conception de système TSL se révèle être une tâche complexe, qui soulève deux problèmes principaux : *i)* la génération TALS est loin de couvrir toute la variabilité linguistique des LS et n'est pas encore entièrement automatisée ; *ii)* la synthèse est confrontée à la nature même des signes, par essence multimodale, et est contrainte par des structures linguistiques sous-jacentes.

Nous nous concentrons dans cet article sur ces deux types de problèmes indissociables, en mettant davantage l'accent sur le second qui considère la synthèse automatique d'une description textuelle aux mouvements LS. Plus précisément, nous présentons notre système *SignCom* qui est une version étendue du système précédemment développé (Gibet *et al.*, 2011). Ce système s'appuie sur un principe de composition phonologique pour construire les signes, et d'édition d'énoncés en LSF, en combinant des modèles de synthèse basée données et de synthèse procédurale. Après avoir décrit les principaux travaux de l'état de l'art, nous présentons notre système *SignCom*, en soulignant les principales avancées technologiques réalisées ces dernières années, puis nous décrivons les principaux défis qui restent à relever pour traduire un texte en LS. Quelques exemples en LSF illustrent notre propos.

2. État de l'art

Les études sur les LS sont relativement récentes, avec des approches linguistiques très diversifiées, des modèles d'animation de personnages virtuels et des réalisations informatiques qui dépendent des avancées technologiques et des données disponibles. Dans cette section, nous explorons les principales représentations linguistiques des LS utilisées pour les systèmes TSL avec avatars signeurs ; pour une revue exhaustive sur le sujet, se référer à (Naert *et al.*, 2020) et (Núñez-Marcos *et al.*, 2023).

Représentations linguistiques des langues des signes. Les premiers travaux sur la phonologie des langues des signes ont donné lieu à différents types de représentations. Parmi celles-ci, les travaux de Stokoe (Stokoe, 1972) ont abouti à la description phonologique de l'ASL (*American Sign Language*) sous la forme d'une combinaison de paramètres constituant les signes : l'emplacement du signe, la forme de la main et son mouvement. Cette représentation paramétrique repose sur la possibilité de distinguer le sens des signes à partir de la modification d'un de ses paramètres constitutifs (notion de paire minimale). Un dictionnaire de l'ASL a été créé à partir de cette représentation. Poursuivant les travaux de Stokoe, d'autres paramètres qui participent à la formation et à la distinction des signes ont été identifiés. Ils comprennent l'orientation de la main ainsi que les éléments non manuels tels que les expressions faciales, la direction du regard, les orientations du buste et certains gestes corporels (Battison, 1978). Le système de notation HamNoSys (Prillwitz et Zentrum, 1989) reprend les paramètres précédents et transcrit de manière linéaire les signes en utilisant les symboles informatiques *Unicode*. Les travaux linguistiques remarquables sur l'ASL menés par Liddell et Johnson ont abouti à la définition d'un modèle phonétique qui s'appuie sur le schéma *Posture-Detention-Transition-Shift* (PDTS), en distinguant sur chaque composante phonologique – configuration de la main (HC), orientation (FA), placement (PL), caractéristiques non manuelles (NM) – des éléments statiques et des éléments dynamiques transitionnels.

En linguistique computationnelle, les gestes des LS ont été décrits au moyen de formalismes allant de scripts à des langages informatiques dédiés. Ainsi, le langage SiGML (Elliott *et al.*, 2008) basé sur HamNoSys a été développé pour générer les animations d'avatars 3D. Ce langage a ensuite été étendu en incorporant le modèle PDTS de Johnson et Liddell (Glauert et Elliott, 2011). Une grammaire générative de signes a également été développée à partir d'un système de composition phonologique parallélisé s'appuyant sur des cibles et des mouvements articulés entre cibles (Gibet *et al.*, 2001). Le langage formel AZee, quant à lui, intègre un formalisme symbolique qui s'appuie sur la géométrie et a pour ambition de représenter un ensemble de procédés grammaticaux de manière parallélisée et non linéaire (Filhol *et al.*, 2017).

Ces langages de script ou de spécification permettent de décrire des signes ou des énoncés de manière très analytique et précise. Cependant, la spécification de nouveaux signes peut être très fastidieuse. De plus, la plupart de ces langages intègrent au sein de leur formalisme des éléments temporels explicites, par exemple SIGML, EMBRScript (Héloir et Kipp, 2010) et AZee, où les postures clés de l'avatar sont spécifiées à des instants prédéterminés. Par contre, le modèle *de partition/constitution (P/C)* (Huenerfauth, 2006) propose un schéma de synchronisation implicite qui repose sur une représentation 2D d'un graphe syntaxique 3D. Ce modèle facilite la visualisation et la coordination d'éléments linguistiques sur des axes temporel et spatial. Peu de représentations linguistiques se sont intéressées aux flexions grammaticales des signes. Parmi celles-ci, le projet ATLAS (Lombardo *et al.*, 2010) en LS italienne incorpore des processus flexionnels impliquant l'emplacement, la configuration et le mouvement, ainsi que des spécificateurs de forme et de taille (SFT). De son côté, le système AZee-Paula (Filhol et McDonald, 2018) permet la génération d'un large panel de mécanismes flexionnels en ASL.

Systèmes d'animation d'avatars signeurs. Parmi les méthodes et technologies disponibles pour animer des avatars signeurs, on distingue trois approches principales.

La première consiste, à partir de données restreintes, à animer l'avatar en utilisant des méthodes dites de *synthèse pure*. Nous regroupons dans cette catégorie les approches à base de postures clés, qu'elles soient déterminées manuellement ou automatiquement, associées à des processus d'interpolation, et les approches dites procédurales qui automatisent le processus de génération de mouvement. Avec de telles méthodes, il est possible de synthétiser des signes isolés identifiés par une glose[3] et de construire des séquences signées à partir de procédés de concaténation et de mélange de mouvements. Cela nécessite de contrôler toute la chaîne de production dans ses moindres détails, depuis la spécification des éléments linguistiques de base, leur agencement au moyen d'un langage de description ou de spécification, jusqu'à la synthèse du mouvement. Si ces systèmes d'animation, qui couplent un langage *symbolique* à un moteur d'animation, permettent d'atteindre des objectifs de précision et de contrôle fin des postures statiques, ils aboutissent généralement à des mouvements peu naturels, voire robotisés. De plus, le processus de spécification, long et fastidieux, conduit à la création d'un nombre limité de signes, avec peu ou pas de possibilités de flexions grammaticales. Enfin, la gestion du temps reste complexe à mettre en œuvre, tant au niveau des signes (gestion de la synchronisation entre les composantes des signes) que des transitions entre signes (gestion de la coarticulation). Deux systèmes relatifs à cette approche ont été développés. Avec EMBR (Kipp *et al.*, 2011), les signes dans leur forme de citation sont générés à partir de séquences de poses spécifiées au moyen du langage EMBRScript. JASigning intègre quant à lui le moteur d'animation *AnimGen* qui permet la création de signes spécifiés à partir du langage SiGML (Kennaway *et al.*, 2007 ; Elliott *et al.*, 2008). Ces systèmes d'animation ont été utilisés pour différentes LS (Ebling *et al.*, 2016 ; Efthimiou *et al.*, 2010 ; Roelofsen *et al.*, 2021). Dans les deux cas, la flexion des signes n'est possible qu'au niveau lexical. Plus récemment, le système *Paula* développé à DePaul University génère des animations à partir de la représentation formelle AZee, en combinant des techniques à base de postures clés et d'algorithmes procéduraux qui améliorent la fluidité du mouvement et facilitent la synthèse multimodale (McDonald *et al.*, 2016 ; McDonald et Filhol, 2021).

La seconde approche consiste à développer des méthodes de *synthèse basée données*. Dans ce cas, les mouvements du signeur virtuel sont capturés par des techniques de capture de mouvement qui enregistrent simultanément les mouvements manuels, corporels, ainsi que les expressions faciales et la direction du regard. Par exemple, les projets *SignCom* (Gibet *et al.*, 2011), *Sign3D* (Gibet *et al.*, 2016) ou *Rosetta* (Dauriac *et al.*, 2022) ont permis de développer un système d'animation d'avatars signeurs en LSF à partir de mouvements capturés haute résolution. Les approches basées données facilitent la production d'animations fluides et crédibles d'avatars 3D. Elles permettent de rejouer des séquences relativement longues de LS, mais aussi de modifier des phrases préenregistrées pour créer de nouveaux énoncés. Cependant, la manipulation et l'adaptation des mouvements à de nouveaux contextes nécessitent la prise

3. Une glose est la représentation lexicale en français écrit du signe.

en compte de processus complexes afin de conserver la cohérence du contenu en LS produit, tant au niveau des animations que de l'intelligibilité des phrases en LS.

Les deux types de méthodes – synthèse basée données et synthèse pure – peuvent être combinées pour conduire à des méthodes de synthèse dite *hybride*. Il est en effet possible de remplacer des segments de mouvement par des mouvements synthétisés, ou de combiner des méthodes procédurales avec des données en s'appuyant éventuellement sur des techniques d'apprentissage automatique. Cette synthèse hybride apporte une certaine flexibilité et la possibilité d'enrichir les bases de données en générant, à partir de séquences existantes, des séquences signées avec variations synthétisées. De telles approches ont été développées, notamment en combinant mouvements capturés et méthodes procédurales pour l'étude des verbes directionnels (Huenerfauth *et al.*, 2015) ou pour générer des énoncés avec flexion spatiale (modification de la spatialisation ou du pointage) ou syntaxique (modification de l'agent, de l'objet ou du bénéficiaire) (Naert *et al.*, 2021). Le projet *Rosetta* se place également dans cette approche, en combinant dans un processus d'édition parallélisée une synthèse basée données avec des algorithmes procéduraux (Dauriac *et al.*, 2022).

Il est à noter l'émergence de technologies TSL, notamment celles développées par Keia[4] pour la LSF, et *Hand Talk App*[5] pour la LS brésilienne (Libras) et l'ASL.

L'annotation des données est au cœur des systèmes d'analyse et de synthèse des LS. En effet, c'est lors du processus de segmentation et d'étiquetage des contenus LS que l'information linguistique est incorporée au niveau des pistes du système d'annotation. Ces pistes permettent d'encoder des informations textuelles (gloses) de nature phonétique, phonologique, lexicale, syntaxique ou sémantique. Elles permettent également d'informer sur le type de mouvement (s'agit-il d'un mouvement signifiant ou d'une transition, etc.), ou bien de structurer hiérarchiquement des groupes d'articulations. L'annotation s'appuie par conséquent sur un formalisme linguistique et sur une représentation structurelle du mouvement qui conditionnent la reconnaissance ou la synthèse des LS. Les premiers systèmes d'annotation on été réalisés de manière manuelle à l'aide d'outils informatiques dédiés (Chételat-Pelé et Braffort, 2008). Plus récemment, des modèles d'annotation par apprentissage automatique ont vu le jour. Ils sont appliqués sur des données de capture de mouvement (Naert *et al.*, 2018) ou sur des données vidéo (Chaaban *et al.*, 2021).

Systèmes de traduction automatique texte-vers-LS. Avec l'avènement de l'apprentissage profond, des modèles récents à base de réseaux de neurones (NN) ont été développés avec succès pour la traduction automatique en LS, en s'inspirant du principe de la traduction du texte vers la parole ou d'une langue parlée vers une autre. Quelques états de l'art permettent de recenser ces systèmes TSL (Kahlon et Singh, 2021) ou plus largement parole/texte-vers-LS et LS-vers-texte (Farooq *et al.*, 2021 ; Núñez-Marcos *et al.*, 2023). Dans ce contexte, des NN ont été proposés pour la traduction en LS arabique (Brour et Benabbou, 2019), ou pour la traduction

4. https://www.keia.io/

5. https://www.handtalk.me/en/

de l'anglais parlé en ASL dans le cadre applicatif de la diffusion de bulletins météo. Des réseaux générateurs de type GAN, associés à des graphes de mouvement, ont également été proposés pour produire des vidéos de LS à partir de phrases en langue parlée (Stoll *et al.*, 2020). Les résultats sont prometteurs mais encore insuffisants.

3. Caractéristiques des langues des signes

Nous partons de l'hypothèse que la formation des signes et des énoncés en LS est déterminée par la réalisation simultanée de formes de main, d'orientations, d'emplacements, de mouvements, qui constituent les unités minimales, dites phonologiques des signes (Stokoe, 1972 ; Battison, 1978 ; Liddell et Johnson, 1989). La composition parallèle de ces unités phonologiques en nombre restreint permet de construire un ensemble structuré et codifié qui définit les bases du lexique, avec une économie de représentation qui est à rapprocher de la structure phonétique des langues vocales. De plus, les LS s'appuient sur deux dynamiques essentielles de l'expression gestuelle, à savoir la spatialité et l'iconicité. L'ensemble des règles qui sont structurées relativement à ces dynamiques gestuelles fonde la grammaire de la LSF. Une autre spécificité des LS concerne la difficulté de séparer les différents niveaux linguistiques – phonologique, lexical, syntaxique et discursif – qui sont propres aux langues vocales. Dans cette section, nous évoquons de manière non exhaustive quelques mécanismes linguistiques de la formation des signes isolés et des énoncés en LSF. Nous nous focalisons plus particulièrement sur les procédés incorporés à notre système de synthèse. La terminologie est empruntée au modèle linguistique de Millet (Millet, 2019).

3.1. *Formation des signes isolés*

La formation des signes isolés dans leur forme de citation nécessite la combinaison spatio-temporelle des composantes phonologiques décrites précédemment. Ces signes sont toujours exécutés de la même manière à un emplacement spécifique de l'espace du signeur, qui peut être soit une zone neutre de l'espace qui l'entoure, soit un emplacement sur son corps. Ces signes n'étant pas soumis aux processus de flexion grammaticale, leurs variations proviennent essentiellement de la façon dont les mouvements sont exécutés (occupation de l'espace, cinématique des mouvements). Leur réalisation requiert toutefois une grande précision, la modification d'une composante phonologique engendrant un sens différent, comme le signe [CHOCOLAT] qui devient le signe [BRICOLER][6] en modifiant le mouvement de la main dominante (vitesse et amplitude), l'emplacement et les configurations manuelles étant inchangés. Du point de vue temporel, les règles de synchronisation entre les éléments composant le signe doivent être respectées.

6. `http://www.sematos.eu/lsf.html`

3.2. *Formation des énoncés à visée illustrative*

Nous explorons ci-après quelques mécanismes grammaticaux en LSF, en organisant notre propos suivant les concepts de spatialité et d'iconicité, les éléments de spatialité étant implicitement reliés aux dynamiques iconiques. Il s'agit d'une description succincte, schématisée à des fins de modélisation pour la synthèse présentée dans la section 4. La flexion grammaticale, qui se traduit par la modification d'une ou de plusieurs composantes phonologiques, conduit à modifier le sens de l'énoncé.

Espace de signation, Locus. L'espace de signation est défini comme étant l'espace dans lequel le discours du locuteur va se déployer. Le signeur utilise cet espace en positionnant des entités, animées ou non, présentes dans son discours. Pour ce faire, il définit des zones symboliques discrètes, parfois présémantisées (par exemple pour représenter la 1^{re} ou la 3^{e} personne). Les emplacements abstraits (*Locus*) associés à ces zones de l'espace de signation, permettent d'identifier et de rappeler ces entités, apportant ainsi une cohérence sémantique à la phrase.

Ancrage et spatialisation. Au niveau lexical, les signes sont réalisés à un emplacement neutre (juste devant le signeur) ou sur son corps. Certains signes à ancrage neutre peuvent être repositionnés à d'autres emplacements de l'espace de signation (spatialisation), ce qui modifie ainsi leur rôle syntaxique/sémantique dans l'énoncé. Par exemple, pour décrire le placement d'un livre sur une étagère, le signe [LIVRE] est d'abord signé dans son ancrage lexical, puis la forme de la main en *proforme* [LIVRE] est positionnée à un endroit cible sur l'étagère.

Pointage. Le pointage a soulevé de nombreuses questions dans la communauté internationale (Garcia *et al.*, 2011 ; Blondel *et al.*, 2004). Il peut prendre différentes formes, soit en étant porteur d'une signification propre (valeur de pronom par exemple), soit en ayant l'objectif de montrer une entité. Dans ce dernier cas, l'entité pointée constitue l'information signifiante, le mouvement de pointage réalisant le déplacement de la main vers cet élément pointé. La désignation d'un emplacement (ou locus) se fait souvent par pointage de l'index ou autre configuration manuelle.

Formes de main. La configuration manuelle (HC) comporte une forte dimension iconique. Du point de vue lexical, elle est l'une des composantes de formation des signes. Par exemple le signe [ESCARGOT] devient [LIMACE] en modifiant la HC (Y devient H), les mouvements étant inchangés (Naert *et al.*, 2021).

Spécificateurs de forme et de taille (SFT). Les SFT sont des procédés qui permettent de décrire la forme ou la taille des entités signées. Si l'on prend l'exemple des signes lexicalisés [BOL] et [VERRE] en LSF, ils peuvent être fléchis de manière à leur adjoindre une valeur adjectivale, au niveau de la forme pour devenir [VERRE-A-EAU] ou [VERRE-DE-CHAMPAGNE], ou bien au niveau de la taille ([GRAND-BOL], [PETIT-BOL]) (Gibet *et al.*, 2011).

Verbes directionnels ou à trajectoire. Ces verbes s'exécutent au moyen d'un mouvement allant d'un locus à l'autre et mettent en jeu plusieurs actants (agent, objet, bénéficiaire). Ainsi, il est possible de fléchir le verbe [DONNER] suivant différentes

trajectoires allant d'un locus agent à un locus bénéficiaire, ces actants pouvant être des pronoms positionnés dans l'espace de signation. Par exemple, la phrase en français « Je te donne » peut être traduite en LSF par un mouvement de la main d'une zone neutre près du buste (personne 1) vers une zone devant le signeur (personne 2), alors que la phrase « Tu me donnes » est traduite par une trajectoire inversée. Outre leur flexion suivant la trajectoire, certains verbes directionnels peuvent être fléchis suivant l'actant objet. Dans ce cas, la forme de la main représentant l'objet est modifiée. Ainsi, dans les exemples en français « Je te donne un verre » ou « Je te donne un livre », les objets sont représentés par des HC différentes, représentant soit un verre, soit un livre.

Proformes manuelles statiques. Les proformes statiques sont représentées par des HC qui référencient, lorsqu'elles sont maintenues, des éléments lexicaux. Elles peuvent représenter des entités animées (personnes, véhicules). Elles assurent également une fonction pronominale (substitution à une entité lexicale), ou anaphorique. Ce mécanisme favorise ainsi le rappel d'une entité dans le discours. Par exemple, la proforme [PERSONNE] (index pointé vers le haut) peut être rapidement positionnée dans l'espace de signation. De plus, la personne peut être représentée dans différentes positions (debout, assise ou allongée), associées à des HC (et orientations) différentes. Par extension, on peut représenter facilement plusieurs personnes dans un espace (autour d'une table par exemple) ou dans une salle de conférence. Les proformes permettent également de positionner des entités les unes par rapport aux autres. Par exemple, la phrase « Le stylo est dans le verre » peut se traduire en LSF par la forme de main en proforme [STYLO] qui vient se positionner dans la proforme représentant le signe [VERRE]. Les proformes constituent ainsi des procédés iconiques efficaces qui assurent la cohérence syntaxique du discours.

Proformes manuelles dynamiques. Certains comportements ou démarches peuvent être représentés par des proformes dynamiques. C'est le cas par exemple lorsque l'on veut décrire la démarche d'un humain ou d'un animal (un oiseau, un ours, un lion, etc.). La forme de la main représente la forme de la patte de l'animal, et le mouvement des mains indique quant à lui la qualité de la démarche (souplesse/raideur, légèreté/lourdeur) (Naert *et al.*, 2021).

Expressions faciales. Les expressions faciales (FE), par essence iconiques, sont fondamentales dans les LS car elles véhiculent trois types d'informations d'ordre lexico-syntaxique (Millet, 2019) : des informations relatives à la modalité de la phrase, des informations expressives liées à la nature émotionnelle du discours, ou des informations de nature adverbiale ou adjectivale.

– *Modalité de la phrase.* Trois modalités principales ayant un rôle syntaxique s'expriment à travers des FE appropriées : la modalité assertive, correspondant à une phrase affirmative, s'exprime le plus souvent par un visage neutre ; la modalité interrogative se traduit par une FE exprimant l'interrogation ; la modalité impérative est employée lorsque le locuteur donne un ordre. La négation quant à elle peut s'exprimer par le signe [NON], par une mimique faciale, ou les deux combinés. Enfin, une mimique faciale, associée à un mouvement du buste et de la tête peut aussi exprimer la condition (« S'il pleut, je reste à la maison »).

– *Affect.* Au-delà des FE primaires identifiées par (Ekman et Friesen, 1978) (joie, colère, peur, etc.), il existe de très nombreuses FE en LSF (inquiétude, admiration, réflexion, etc.) (Cuxac, 2000). D'autres FE expriment les états affectifs caractérisant l'attitude du locuteur vis-à-vis du contenu de l'énoncé. Les principales concernent l'expression exclamative (surprise) ou dubitative (doute) (Millet, 2019).

– *Valeur adverbiale ou adjectivale.* Enfin certaines mimiques faciales, dans des situations de dialogue ou de récit, ont des fonctions adverbiales (gonflement des joues qui accompagne la phrase « Le vent souffle fort »), ou adjectivales (joues rentrées ou rebondies dans « un homme mince ou gros »).

Ces FE varient en fonction du style du locuteur, de son état affectif et de la morphologie de son visage. De plus, les FE peuvent être combinées entre elles en fonction du contexte du discours.

Autres mouvements – tête, buste, épaules. Les mouvements impliquant d'autres parties du corps (tête, buste, épaules) sont aussi porteurs d'information. Ainsi, en LSF, le torse légèrement penché en avant indique une action réalisée dans le futur. L'orientation du buste est également utilisée pour passer d'un personnage à l'autre dans une narration. Enfin, pour des formes 2D ou 3D, il est possible de décrire les différents niveaux de détail de l'objet en inclinant en avant le buste.

De nombreux autres procédés grammaticaux existent en LSF (proformes non manuelles, prises de rôle, exploitation du regard, etc.). Ils sont décrits précisément dans l'ouvrage de Millet (Millet, 2019).

Dans la suite de cet article, nous nous intéressons à la modélisation, la spécification et l'implémentation de ces procédés linguistiques, en mettant l'accent sur les avancées techniques du système TSL *SignCom*.

4. Notre système de synthèse *SignCom* texte-vers-LSF

Les langues des signes requièrent une grande précision et un haut niveau de réalisme dans l'exécution des mouvements corporels, manuels et faciaux, afin qu'ils soient compris et acceptés par les sourds. La capture du mouvement (MoCap) associée à l'animation d'un avatar 3D permet de répondre à ces exigences. Cependant, le coût de production de ces données MoCap reste très élevé, et il est pertinent d'enrichir les bases de données existantes au moyen de processus d'édition. C'est cette motivation qui nous a conduits à développer un système d'édition et de synthèse multimodale produisant simultanément les mouvements corporels, manuels, parallèlement aux expressions faciales et à la direction du regard. Ainsi, dans le système *SignCom* (Gibet *et al.*, 2011), l'édition a permis de construire de nouvelles phrases, en coupant/collant/transformant/mixant les segments de mouvement sélectionnés, et de produire automatiquement l'animation d'un avatar signeur en 3D. Différents processus d'édition on été explorés, i) par remplacement de signes ou de groupes de signes, ii) par instanciation de schémas syntaxiques prédéfinis, ou, iii) par remplacement de composantes phonologiques des signes – configurations des mains, mouvements ma-

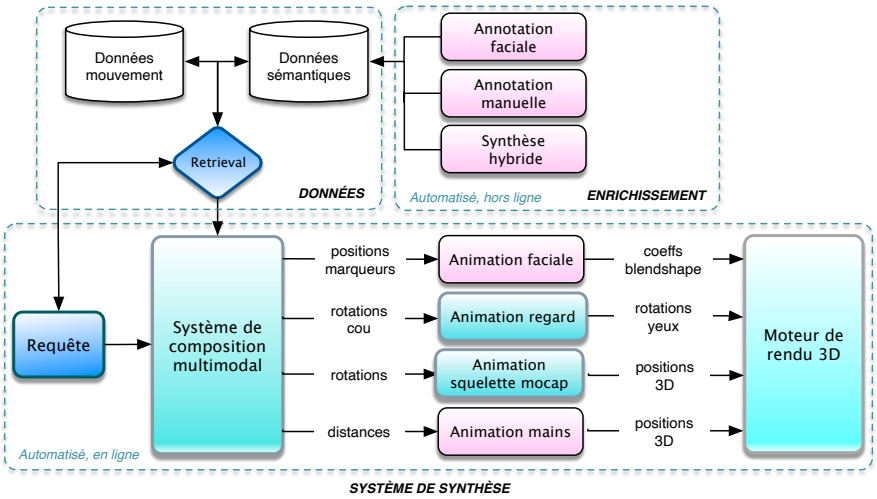

FIGURE 2. *Système de synthèse texte-vers-LSF (SignCom) avec avatar signeur ; en bleu : système de base (Gibet* et al.*, 2011) ; en rose : avancées technologiques*

nuels, du torse et de la tête, et expressions faciales (Gibet, 2018). Cependant, si le système *SignCom* dans sa première version était capable de manipuler des contenus LS aux niveaux phonologique, lexical et discursif, en s'appuyant sur des processus d'annotation manuelle, il ne permettait pas de prendre en compte des mécanismes grammaticaux plus complexes tels que ceux évoqués dans la section 3.

Nous avons développé une extension du système *SignCom* qui intègre certains procédés flexionnels de la grammaire de la LSF se rapportant à la spatialité et à l'iconicité propres à cette langue. Ainsi, la possibilité de fléchir des composantes des signes – notamment les composantes manuelles ou faciales –, permet de produire des variations grammaticales des phrases du corpus initial. Ces processus flexionnels s'expriment à partir des nouveaux modules de synthèse présentés dans cette section. La figure 2 illustre le système *SignCom* dans son ensemble, en intégrant les modules fonctionnels de base et ceux constituant les avancées technologiques. Après avoir rappelé le principe de la synthèse concaténative (section 4.1), nous décrivons ci-après les modules d'annotation automatique (section 4.2) qui visent à réduire le temps d'annotation manuelle, puis nous présentons les modules de synthèse hybride corporelle (section 4.3), de synthèse faciale (section 4.4) et de synthèse du mouvement des mains (section 4.5) qui enrichissent notre système initial.

4.1. *Principe de la synthèse concaténative*

Notre système de synthèse concaténative repose sur des données de mouvement préalablement capturées. Ces données sont caractérisées par des séquences de postures du squelette du signeur. Elles sont enregistrées au moyen d'un système de capture de mouvement à base de marqueurs passifs et de caméras infrarouges, qui permettent de déterminer de manière très précise la position des marqueurs placés sur le corps, les mains et le visage de l'acteur (fréquence d'acquisition de 200 Hz). Il en résulte la constitution d'un ensemble de séquences de mouvement représentant les énoncés du corpus en LSF. Ces données sont ensuite annotées en suivant des schémas multicanaux pour lesquels nous distinguons i) les canaux linguistiques qui respectent la description phonologique des signes (Johnson et Liddell, 2011) et leur classe grammaticale (Johnston, 1998), et ii) les canaux physiques qui correspondent à des groupes d'articulations (main, bras, etc.).

Le système de synthèse est divisé en deux parties : un processus hors ligne de stockage des données et un processus en ligne d'extraction des données et de contrôle de l'animation. La nécessité d'encoder dans les énoncés signés, d'une part les informations linguistiques traduisant la structure multilinéaire des LS et d'autre part les flux de mouvement, nécessite de construire au préalable deux bases de données hétérogènes couplées, l'une contenant les données sémantiques issues de l'annotation, l'autre les données de mouvement (positions ou angles aux articulations du squelette). La base de données sémantique réalise le couplage entre les données symboliques suivant notre schéma multicanal d'annotation et une liste de mouvements indexés au moyen d'un nom, de marqueurs temporels, et d'une séquence d'articulations (au sens biomécanique) impliquées dans le mouvement. Il s'agit d'un couplage *un-vers-plusieurs* pour tenir compte de l'existence de plusieurs instances d'un même signe ou partie de signe dans le corpus. Un langage de requêtes multiconditions permet, à partir de la spécification d'éléments propres à cette annotation, d'extraire automatiquement un flux continu de postures du mouvement qui lui correspondent.

Le système d'animation à proprement parler s'appuie sur un processus de composition multicanale du mouvement qui reçoit en entrée des flux de données associées à des segments corporels qui sont, soit des groupes d'articulations du squelette appelés *effecteurs* (corps, bas du corps, torse, colonne vertébrale, bras gauche/droit, main gauche/droite), soit des données propres aux expressions faciales ou à la direction du regard. La composition est réalisée à la fois spatialement, avec des niveaux de priorité appliqués aux effecteurs, en suivant l'organisation hiérarchique structurelle du squelette, et temporellement en déclenchant au moment approprié le contrôleur de synthèse spécifique à l'effecteur considéré. Pour chaque élément du squelette, un contrôleur paramétré permet de rejouer le mouvement préenregistré avec la possibilité de lui appliquer des traitements spécifiques tels que la répétition, l'inversion, etc. La synthèse du mouvement du regard est réalisée par un modèle de cinématique inverse guidé par des positions pointées en 3D couplées aux mouvements de la tête. Puis une technique de mélange de mouvement est appliquée hiérarchiquement aux données de sortie des contrôleurs, afin de fluidifier le mouvement produit. Ce système permet,

par agencement de mouvements préenregistrés, de construire de nouveaux énoncés en substituant des signes ou des portions de phrase par d'autres, ou en modifiant des éléments linguistiques sur un ou plusieurs canaux phonologiques. Dans l'exemple de la phrase en français « J'aime les jus de fruits », transformée en « Je n'aime pas le jus d'orange », le mouvement du buste ainsi que celui du bas du corps et du bras gauche sont conservés. Par contre, les mouvements de la tête et du bras droit, ainsi que l'expression faciale sont modifiés, de façon à préserver la cohérence sémantique de la phrase (Duarte, 2012).

4.2. *Annotation par apprentissage automatique*

Nous avons développé un système d'annotation couplé au système de synthèse concaténative, qui s'appuie sur des algorithmes d'apprentissage automatique (Naert *et al.*, 2018). En effet, la précision de l'annotation, à la fois spatiale (nature et structure des pistes d'information) et temporelle (marqueurs temporels associés aux segments) conditionne la finesse d'édition du mouvement, la cohérence grammaticale du résultat et la qualité de l'animation produite. Cependant, annoter manuellement un corpus LSF constitue une tâche chronophage qui nous a conduits à opter pour un processus d'annotation automatique pour les HC et FE. L'annotation des HC est réalisée à travers une chaîne de traitements qui permet d'extraire les principaux descripteurs manuels, de segmenter les phrases à partir d'une détection de seuil sur des distances moyennes variationnelles, puis d'étiqueter ces phrases au moyen de méthodes d'apprentissage automatique (ML). Une évaluation quantitative a été réalisée sur la base d'un corpus contenant 32 classes de configurations manuelles. Un sous-ensemble de 29 distances a permis de classifier de manière optimale les HC. Les différentes méthodes utilisées ont donné des scores de bonne classification de 87 % pour la régression logistique, 89 % pour les k plus proches voisins (kNN) avec k = 3, et de 93 % pour la méthode *Support Vector Machine* (SVM). L'annotation des FE a été réalisée de manière analogue. Le processus de segmentation s'appuie sur la détection des *maxima* des dérivées première et seconde des descripteurs de *blendshape* (section 4.4). Les résultats de ML donnent une précision de 91 % pour les forêts aléatoires, de 86 % pour SVM et de 72 % pour kNN (k = 3). À la fois pour les HC et FE, l'étiquetage automatique est ensuite effectué sur la base de la classe prédominante sur chaque segment.

4.3. *Enrichissement par synthèse hybride du mouvement corporel*

La synthèse corporelle dite *hybride* a pour objectif d'enrichir la base de données initiale en adjoignant à ces données des données synthétisées avec variations flexionnelles, facilitant ainsi la création de nouveaux énoncés en LSF. En effet, la capture d'un ensemble volumineux de données de mouvements MoCap s'avère longue et fastidieuse, et il existe encore peu de corpus couvrant la grande variabilité des LS. Nous décrivons ci-après quelques techniques de synthèse qui permettent d'augmenter les données enregistrées tout en respectant les mécanismes de spatialité et d'iconicité

de la LSF. Certaines de ces techniques ont été implémentées et évaluées (pointage, modifications lexicales et syntaxiques à partir de la manipulation des HC) (Naert *et al.*, 2021) ; d'autres procédés sont spécifiés (spatialisation, verbes directionnels, SFT) en vue d'une intégration dans notre système *SignCom*.

Pointages dans l'espace de signation. Nous avons vu que l'espace de signation définit des zones spatiales discrétisées qui peuvent être exploitées dans le discours en LSF. En particulier, les gestes de pointage définissent des procédés syntaxiques de type locus/pointage qui permettent d'assurer la fonction pronominale et de référencer des entités du discours. Cependant, ces gestes, présents dans les données capturées, correspondent à un ensemble limité de zones pointées. Or, par synthèse, il est possible de produire un nombre illimité de gestes de pointage qui couvrent l'espace de signation. À cette fin, nous avons défini un modèle d'inversion cinématique (IK) qui, à partir de la seule spécification d'un ensemble de positions cibles référencées (locus), génère le mouvement de pointage vers ces cibles.

Spatialisation. Ce même procédé peut être utilisé pour répondre à la problématique de spatialisation des signes. En effet, disposant des signes réalisés dans la zone de leur ancrage lexical, la technique de l'IK permet de déplacer la main vers la zone souhaitée et de signer l'entité lexicale à cette position spécifique.

Verbes directionnels. Les verbes directionnels sont caractérisés par une trajectoire qui définit la trace du mouvement d'un locus à l'autre et permet ainsi de distribuer les rôles actanciels agent/bénéficiaire/objet dans la phrase. Notre système permet de produire de telles phrases en spécifiant les locus correspondant aux pronoms souhaités, puis par IK en synthétisant le mouvement de la main. Par exemple, la phrase en français « Je te donne un livre » peut être modélisée par l'expression paramétrée [DONNER]([PRO-1],[PRO-2],[LIVRE]) dans laquelle le verbe [DONNER] s'exécute par un mouvement de la main, depuis la localisation du pronom « je » [PRO-1] vers celle du pronom « tu » [PRO-2] avec une configuration manuelle qui est celle du signe [LIVRE]. Pour générer la phrase « Tu me donnes un livre », il suffit d'inverser le mouvement de [PRO-2] à [PRO-1]. Pour générer la phrase « Je lui donne un livre », nous spécifions le nouveau locus correspondant à la 3^e personne [PRO-3], et synthétisons par IK le mouvement de [PRO-1] à [PRO-3] tout en conservant la configuration manuelle. Pour les verbes directionnels comportant un actant objet, la phrase est signée avec une configuration manuelle qui correspond à l'objet. Ce procédé de synthèse peut être généralisé à la plupart des phrases comportant des verbes directionnels.

Spécificateurs de forme et de taille (SFT). Nous distinguons les SFT qui agissent sur les mouvements, configurations manuelles ou orientations des mains, ou sur une combinaison de ces composantes des signes, et nous proposons ci-dessous une liste non exhaustive de flexions des signes rencontrées en LSF. Ces SFT ont été modélisés et spécifiés. Ils impliquent des procédés de synthèse très différents. Seules les trois premières situations ont été implémentées.

– La taille de la trajectoire du mouvement peut être modifiée en spécifiant une trajectoire rectiligne dans le plan transversal du signeur et en synthétisant le mouvement

par une technique d'IK ou d'interpolation. C'est le cas du signe [TABLE], dans lequel les mains plates s'écartent plus ou moins en fonction de la taille de la table.

– Il est relativement simple de spécifier et de remplacer la configuration manuelle statique sur tout le signe, comme dans les signes [VERRE-FIN], [GROS-VERRE], où seule la forme de la main change (*C* plus ou moins ouvert).

– Modifier dynamiquement la configuration manuelle au cours d'un signe nécessite d'employer un modèle de cinématique directe (FK) entre deux ou plusieurs formes clés de la main. Ainsi, à partir du signe [VERRE], il est possible de générer le signe [COUPE-DE-CHAMPAGNE], dans lequel la forme de la main s'évase vers le haut.

– D'autres SFT impliquent de modifier simultanément la trajectoire et l'orientation de la main, la configuration manuelle étant statique. C'est le cas du signe [BOL] dont la taille peut varier. De la même manière, on peut modifier les trajectoires et configurations manuelles simultanément, l'orientation étant statique, comme dans le signe [BANANE]. Les techniques de synthèse supposent de générer des trajectoires plus ou moins complexes et de synthétiser par IK le mouvement des articulations des bras, et de manière simultanée de synthétiser par FK les configurations ou orientations manuelles dynamiques.

Évaluations du système global. Deux évaluations perceptuelles ont été réalisées, dans lesquelles nous avons comparé les résultats de synthèse avec les données de rejeu MoCap qui constituent la vérité terrain (Naert *et al.*, 2021). La première évaluation concerne l'étude de la spatialisation et du pointage. 57 participants de bon niveau en LSF et au-delà (très bon, natif et interprète) ont répondu à des questionnaires sur le web avec des consignes vidéo, les réponses étant fournies au moyen de textes (menus déroulants), d'images ou de vidéos. Pour la tâche consistant à reconnaître l'emplacement du signe [BOL] à travers la visualisation de 8 vidéos (5 vidéos de synthèse et 3 de rejeu), le taux de reconnaissance est de 86 % pour les signes synthétisés contre 63 % pour les signes rejoués, la différence s'expliquant par une plus grande variabilité des signes réels. L'évaluation du réalisme de ces mêmes signes a donné un score moyen de 3,6 pour les énoncés de synthèse contre 3,8 pour les énoncés de rejeu, sur une échelle de Likert de 5 points, montrant qu'il n'y a pas de différence significative entre les séquences synthétisées et les séquences de rejeu (*p-value* de 0,031). La seconde étude a permis d'évaluer le réalisme des gestes de pointage. 9 vidéos ont été présentées aux mêmes participants (6 vidéos de synthèse et 3 de rejeu). Les scores sont respectivement de 3,15 pour les gestes de synthèse et de 3,45 pour les gestes de rejeu. Là également, nous avons conclu qu'il n'y a pas de différence marquée entre les gestes réels et de synthèse (*p-value* de 0,081).

La seconde évaluation concerne les processus relatifs à la manipulation de configurations manuelles (HC), statiques ou dynamiques. Pour cette étude, 39 participants ont visualisé 20 vidéos présentées dans un ordre aléatoire (13 de synthèse et 7 de rejeu), représentant des signes issus de la dactylologie tels que [LSF] ou [OK], ou différentes démarches d'animaux (par exemple celles du coq et du chat). Pour la tâche consistant à évaluer le remplacement des HC par d'autres (5 vidéos), les résultats ont montré qu'il n'y avait pas de différence significative entre les énoncés de rejeu ou de synthèse, ces

derniers étant même jugés plus réalistes. Pour la tâche qui consiste à reconnaître les signes issus de la dactylologie (15 vidéos), les taux de reconnaissance obtenus pour la synthèse sont de 95 % pour les signes synthétisés et de 91 % pour les signes de rejeu. De plus, aucune différence notable n'a été observée en ce qui concerne le réalisme des signes synthétisés (score de 3,08/5) et rejoués (score de 3,03/5).

4.4. *Synthèse faciale*

Motivation et approche. La synthèse de mimiques faciales associées aux trois fonctions de la LSF – expression de la modalité de la phrase, expression adverbiale ou expression affective – nécessite de couvrir un large spectre de mimiques expressives, avec toute la richesse et les nuances spécifiques aux LS. Nous nous sommes intéressés plus spécifiquement aux mimiques faciales affectives dans le cadre des émotions de base (Ekman et Friesen, 1978), ainsi qu'aux mimiques exprimant les modalités négative, interrogative, exclamative ou injonctive de la phrase. Ces expressions faciales (FE) peuvent se combiner entre elles avec différents degrés d'intensité. De plus, afin de préserver la cohérence grammaticale des phrases produites, nous avons opté pour une technique de synthèse faciale basée données capturées, ce qui permet d'atteindre une grande précision, tant spatiale que temporelle (fréquence d'acquisition > 200 Hz) ainsi que la synchronie avec les mouvements corporels et manuels. Par ailleurs, nous avons adopté une représentation paramétrique unifiée qui s'appuie sur des formes clés (*blendshapes*) associées à des coefficients de pondération (dits coefficients de *blendshape*). Ces coefficients sont calculés automatiquement à partir des données de MoCap grâce à une chaîne de synthèse que nous détaillons ci-après.

Synthèse à base de formes clés (*blendshapes*). L'animation basée *blendshapes* permet de construire une expression faciale $\boldsymbol{v}$ (vecteur des N positions du maillage de l'avatar) à partir de la combinaison linéaire de N formes de base $\boldsymbol{b}_i$, chacune étant pondérée par un coefficient c_i, la forme $\boldsymbol{b}_0$ correspondant à l'expression neutre :

$$\boldsymbol{v} = \boldsymbol{b}_0 + \sum_{i=1}^{i=N} c_i \boldsymbol{b}_i \qquad [1]$$

Les formes de base représentent des expressions unitaires impliquant une petite partie du visage (sourcil, bouche, menton, etc.). La figure 3 (partie droite) illustre le principe de la synthèse par blenshapes. L'équation 1 peut se réécrire : $\boldsymbol{v} = B.\boldsymbol{c}$, où B représente la matrice des formes de base et $\boldsymbol{c}$ le vecteur des coefficients de *blendshapes*. Ce type de modèle présente plusieurs avantages. D'une part, il fournit un niveau d'abstraction permettant le transfert d'une animation d'un modèle d'avatar 3D à un autre. D'autre part, il s'agit d'un modèle linéaire très simple, ce qui conduit à des temps de calcul autorisant les applications temps réel. Enfin, cette représentation stable et régulière permet la segmentation et l'étiquetage des séquences d'expressions faciales. Une étude préalable (Reverdy *et al.*, 2015) a permis de confirmer l'hypothèse selon laquelle l'utilisation de *blendshapes* pour animer le visage permet non seulement

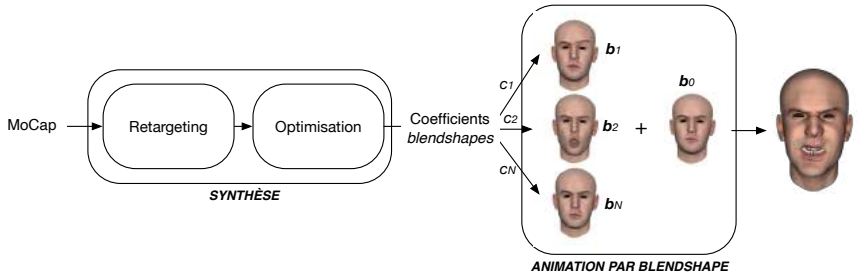

FIGURE 3. *Chaîne de synthèse faciale à partir des données MoCap*

de réduire les coûts de calcul, mais aussi de produire une animation faciale convaincante. Par la suite, afin de synthétiser des expressions précises et subtiles, nous avons choisi un grand nombre de formes de base (soit 51). Cette étude a également permis de comparer et d'évaluer plusieurs jeux de marqueurs (nombre et positionnement) grâce à une méthode de *clustering*, conduisant ainsi à un jeu optimal de 41 marqueurs.

Chaîne de synthèse. Notre méthode LSTS de synthèse faciale permet de transformer automatiquement des données 3D de MoCap faciale en coefficients de *blendshape* utilisés ensuite pour l'animation d'un avatar virtuel (Reverdy, 2019). Cette méthode comporte deux étapes principales (figure 3). La première, dite de *retargeting*, a pour objectif d'adapter morphologiquement les trajectoires enregistrées sur l'acteur afin qu'elles correspondent à la morphologie de l'avatar ciblé. Après un post-traitement qui consiste à supprimer les transformations rigides (translation et rotation) des données 3D et à aligner les données dynamiques de l'acteur sur celles de l'avatar, une méthode de régression de type *RBF* (*Radial Basis Function*) permet de résoudre ce problème d'adaptation. La deuxième étape exploite une méthode d'optimisation pour synthétiser automatiquement les coefficients de *blendshape* de l'avatar à partir des positions de marqueurs. Le problème revient à minimiser l'erreur quadratique entre les positions variationnelles des P marqueurs MoCap $\delta\boldsymbol{m} = \hat{\boldsymbol{m}} - \hat{\boldsymbol{m}}_{\mathbf{0}}$ obtenues après *retargeting* et celles des P points du maillage de l'avatar correspondant, calculées à partir du modèle de *blendshapes* $\delta\boldsymbol{a} = B_P.\boldsymbol{c}$, où B_P correspond à la projection de la matrice B sur les P points du maillage alignés sur les marqueurs MoCap :

$$\hat{\boldsymbol{c}} = \arg\min_{\boldsymbol{c}} ||\delta\boldsymbol{m} - \delta\boldsymbol{a}|| + \alpha_b.E_b + \alpha_t.E_t \qquad [2]$$

E_b et E_t étant deux énergies de régularisation qui pénalisent l'espace des solutions de ce problème d'optimisation. L'énergie de normalisation E_b permet d'éviter que les coefficients de *blendshape* ne sortent de l'intervalle $[0, 1]$. L'énergie Laplacienne de déformation E_t permet de minimiser la déformation de la structure du maillage par rapport au maillage original dans son expression neutre. Les paramètres de pondération α_b et α_t sont tels que $\alpha_b + \alpha_t$ = 1

Résultats. Nous avons travaillé principalement sur les expressions affectives en choisissant les six classes d'émotions primaires – la colère (C), le dégoût (D), la peur (P), la joie (J), la tristesse (T), la surprise (S) – auxquelles nous avons ajouté le neutre (N), et sur les modalités syntaxiques assertive, interrogative, exclamative et négative (Reverdy, 2019). Un corpus a permis de valider notre méthode de synthèse *LSTS* et d'animation. Il a été enregistré pour un seul signeur de niveau B2 au moyen de deux dispositifs de capture employés simultanément : le système MoCap précédent et une caméra RGB-D. Ce corpus est constitué de séquences alternant les expressions neutres et expressives pour les 6 émotions, en tenant compte de 3 degrés d'intensité différente (N – C1 – N – C2 – N – C3 – N – J1 – N – J2 – ... avec 1 : faible ; 2 : marqué ; 3 : exagéré) et de phrases expressives avec différentes modalités syntaxiques, soit au total environ 30 minutes d'enregistrement. Une première étude perceptuelle a été réalisée sur la base de ce corpus. 27 personnes (17 hommes et 10 femmes, âgés de 17 à 31 ans) ont participé en répondant à des questionnaires en ligne. 54 vidéos de synthèse réalisées en variant les facteurs de pondération des deux énergies de régularisation leur ont été présentées dans un ordre aléatoire. Tout d'abord, le choix d'un paramétrage optimal de la méthode *LSTS* a été établi en analysant subjectivement les facteurs de reconnaissance de l'émotion, d'identification de l'intensité perçue et de réalisme des animations. Les résultats ont donné pour le meilleur modèle de synthèse un taux de reconnaissance moyen de 62,5 %, les émotions les mieux reconnues étant la joie, la surprise et la colère (88 %), un score de reconnaissance de l'intensité de 5,09 (sur une échelle de Likert de 1 à 7) et un score de réalisme de 5,2/7.

Une seconde étude perceptuelle a permis de valider le modèle *LSTS* par comparaison avec le modèle *FS* proposé par *faceshift* [7] qui est une référence au niveau de l'état de l'art, en utilisant les mêmes facteurs d'évaluation auxquels on a rajouté le critère de fidélité par rapport aux vidéos originales. 41 personnes séparées en 2 groupes ont participé à l'étude. Les méthodes *FS* et *LSTS* ont donné des résultats similaires, avec des taux de reconnaissance respectifs de 62,4 pour *LSTS* et 62,6 pour *FS*, et un réalisme de 5,1/7 pour *LSTS* et de 4,9/7 pour *FS*. La fidélité des vidéos de synthèse par rapport aux vidéos réelles a donné des scores de 5,1/7 pour *LSTS* et de 4,9/7 pour *FS*.

4.5. *Synthèse du mouvement des mains*

En raison de la rapidité et de la précision des gestes de la LSF, la reconstruction des données manuelles est particulièrement longue et fastidieuse. Parmi les différentes méthodes exploitables (captation des positions des marqueurs, par gants équipés d'accéléromètres/gyroscopes, captation basée vision), l'utilisation de la MoCap à base de marqueurs présente les résultats les plus aboutis et en adéquation avec la qualité attendue pour les LS, grâce à la précision spatiale et temporelle de ces marqueurs. Elle souffre néanmoins d'une phase coûteuse de post-processing pour corriger les problèmes d'occultation, très nombreux en LS.

7. *faceshift* : `http://www.faceshift.com/product/,2012`

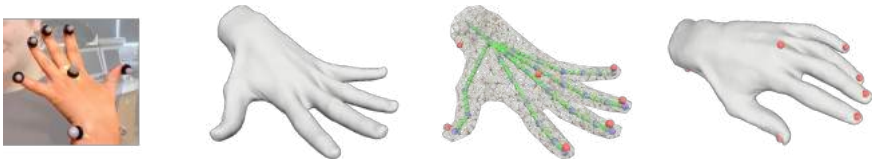

FIGURE 4. *Synthèse du mouvement des mains en 4 étapes : 1. Mains avec marqueurs réfléchissants ; 2. Maillage de référence ; 3. Maillage volumétrique intégrant le squelette et les positions de marqueurs ; 4. Posture générée par notre système*

Nous proposons une chaîne d'animation complète pour la synthèse du mouvement des mains à partir d'un jeu de marqueurs simplifié (figure 4.1). Notre méthode se fonde sur une technique de déformation Laplacienne qui intègre dans une structure de contrôle volumétrique le maillage haute résolution, le squelette, ainsi que les emplacements des marqueurs pertinents (figure 4.3). Une méthode d'optimisation itérative, qui préserve à la fois les caractéristiques géométriques, les longueurs des segments et les butées articulaires, est appliquée à cette structure. En suivant cette approche présentée dans Le Naour *et al.* (2019), nous avons montré la capacité du modèle d'optimisation à animer et à déformer interactivement des modèles de main en haute définition à partir d'un faible nombre de contraintes de position, tout en conservant tous les détails du mouvement. Ce modèle a été appliqué avec succès aux mouvements des mains et des doigts en LSF, avec la possibilité d'éditer et d'adapter les données morphologiques de façon à ce que les contraintes de contact et de non-interpénétration des doigts soient bien respectées. Une évaluation quantitative a été réalisée sur la base des erreurs *root mean square* entre la séquence synthétisée et la vérité terrain (erreur < 5 %).

5. Problèmes non résolus

Ces problèmes sont multiples. Nous présentons ci-après de manière non exhaustive ceux qui nous paraissent les plus importants au regard des TSL.

Modélisation des mécanismes flexionnels des LS. Les systèmes actuels TSL sont loin d'être complètement automatisés et d'intégrer l'ensemble des mécanismes flexionnels des LS. Cette constatation concerne à la fois les représentations linguistiques des LS et les techniques de synthèse employées. Cela peut s'expliquer par le fait que les LS sont des langues à modalité gestuelle, dont le fonctionnement et les structures linguistiques sont encore peu formalisés au sens des langages formels en informatique. Quelques procédés linguistiques ont été modélisés et traduits dans des langages informatiques dédiés (section 2). Cependant, ces langages sont encore loin de couvrir la multitude de phénomènes répertoriés, de manière générique et paramétrisée. De plus, il est nécessaire d'aller vers l'automatisation des processus, depuis la traduction des phrases en langage Pivot-LS vers la spécification de l'animation (figure 1).

Conjointement, les modèles d'animation classiques d'avatars ne peuvent s'appliquer directement aux LS car ils ne s'adaptent pas aux variations flexionnelles mentionnées. Ces deux objectifs doivent être envisagés simultanément. Cela implique de maîtriser à la fois le traitement automatique des langues signées (TALS) et les méthodes d'animation par ordinateur.

Coordination multi-effecteur et synchronisation temporelle. Du point de vue linguistique *computationnelle*, l'un des enjeux majeurs concerne la prise en compte du temps dans le formalisme de synchronisation entre les canaux d'information. Le type de langage utilisé détermine la façon dont cette synchronisation est gérée. En particulier, les langages de script ou impératifs gèrent le temps de manière explicite, ce qui peut s'avérer fastidieux pour traduire les phrases dans le formalisme choisi. Des langages réactifs permettraient d'intégrer ces éléments de synchronisation, mais ils sont en général plus difficiles à manipuler. Du point de vue de la synthèse du mouvement, l'un des défis principaux réside dans la coordination multi-effecteur et la synchronisation des mouvements générés sur chacune des pistes – soit par extraction dans la base de données, soit par synthèse –, de façon à respecter les schémas spatio-temporels des signes. En effet, pour que les signes générés paraissent plausibles, il est nécessaire que les mouvements relatifs à des groupes d'articulations respectent des schémas de synchronisation propres au contrôle moteur : par exemple la configuration manuelle doit toujours être atteinte avant que la main ait atteint son objectif cible (Duarte, 2012). De plus, si l'on mélange plusieurs styles de mouvements, la recherche d'une cohérence de style peut imposer de compresser ou de dilater des portions de mouvement de façon à ce que les règles de synchronisation soient vérifiées (Héloir et Gibet, 2007). Cette synchronisation se répercute également aux mouvements secondaires apparaissant dans certains signes. Enfin, il est primordial de gérer la coarticulation, et ceci aux niveaux intrasigne et intersigne de façon à tenir compte du contexte passé et futur de chaque signe dans la séquence générée.

Adaptation morphologique et prise en compte des contacts. La plupart des avatars signeurs utilisent les données signées d'un seul locuteur. L'adaptation morphologique à d'autres avatars signeurs passe par des processus d'adaptation morphologique (*retargeting*) permettant par exemple de transférer les animations vers d'autres personnages (homme, femme, enfant, voire animal). De plus, les contraintes des LS sont liées au contenu des signes et des énoncés. En effet, de nombreux signes impliquent des contacts entre les deux mains, ou entre chacune des mains et une partie du corps. Ces contraintes spatiales doivent être spécifiées précisément dans l'espace du signeur (par exemple « les mains doivent rester au-dessus de la table ») ou exprimées de manière qualitative (par exemple « l'index doit toucher la paume de la main »). Les algorithmes proposés pour la synthèse des mouvements manuels (section 4.5) sont capables de traiter un ensemble de ces contraintes numériques par optimisation, par exemple en ajoutant une connaissance liée à l'environnement ou en permettant le relâchement de certains degrés de liberté. Il serait intéressant d'intégrer ces contraintes dans le système de synthèse grâce à un langage utilisateur de haut niveau.

Modélisation physique. La physique des mouvements (au sens des forces mises en œuvre) fait partie intégrante des dynamiques gestuelles impliquées dans les LS. Ainsi, la façon dont les contacts sont exécutés (de manière effleurée ou frappée) modifie le sens des signes. Dans un futur proche, il paraît inévitable que les systèmes de synthèse d'avatars signeurs soient modélisés et simulés physiquement.

Rareté des données. Les données sont au cœur des technologies et méthodes employées pour les avatars signeurs. Plusieurs types de données sont disponibles : la vidéo, la capture de mouvement, les images et les textes (français écrit, LSF-glosée, annotations, etc.). Plusieurs questions se posent pour la définition du corpus. La première concerne le compromis entre étendue et profondeur du corpus. Si l'objectif est de disposer d'un lexique qui couvre un large domaine, comprenant plusieurs thématiques, un corpus étendu sera privilégié. Si, au contraire, l'objectif est d'avoir un vocabulaire limité et de le réutiliser dans différents énoncés avec flexions grammaticales, alors on choisira l'approche en profondeur. Dans ce cas, de nombreuses instances des mêmes signes avec variations doivent être considérées dans le vocabulaire prédéfini. La deuxième question concerne la nature des variations elles-mêmes qui doivent être incluses dans le corpus pour l'édition et la synthèse. Pour pallier ces difficultés, on peut à court terme remplacer les données de MoCap par des données vidéo, plus faciles à acquérir. Cela permet de disposer de gros volumes de données et d'envisager la traduction TSL en exploitant des méthodes performantes d'apprentissage profond développées dans le domaine de la vision et de l'animation par ordinateur. Enfin, une préoccupation essentielle relative à la construction du corpus est la qualité actée ou spontanée des mouvements produits par les locuteurs signants.

Traduction automatique texte-vers-LS et LS-vers-texte. Les systèmes actuels de traduction automatique d'une langue vocale vers une autre laissent entrevoir la possibilité de traduire automatiquement une langue parlée/écrite vers une LS. Les approches de *deep learning* (DL) devraient faciliter cette étape. Cependant, l'absence de système d'écriture communément accepté pour les LS ne permet pas de disposer de suffisamment de données mettant en correspondance un texte dans une langue vocale et sa transcription écrite en LS. Avec le peu de corpus parallèles disponibles, seuls des systèmes de traduction à base de règles, ou ceux portant sur un vocabulaire restreint sont actuellement en cours de développement. De nouvelles méthodes d'apprentissage frugal reposant sur des connaissances préalables devraient être capables d'aider les modèles DL à intégrer de nouveaux concepts à partir de peu d'exemples.

Par ailleurs, les approches TSL neuronales basées vidéos s'appuient sur la transformation entre données vidéo 2D et séquences de postures en 3D. Cependant, si les architectures de réseaux neuronaux conduisent à des séquences de postures relativement précises spatialement, elles n'intègrent toujours pas les configurations manuelles, ou du moins pas de manière précise. De plus, parmi les approches développées, très peu s'intéressent aujourd'hui à la qualité du mouvement produit. Or, la précision des configurations manuelles en LS, et la manière dont les séquences de LS se déroulent dans le temps, constituent des enjeux majeurs pour le développement des systèmes de traduction automatique TSL.

Notons que la reconstruction de séquences de postures squelettiques 3D à partir de vidéo 2D est un moyen de constituer des bases de données conséquentes associant vidéo et MoCap. Celles-ci peuvent être exploitées pour la reconnaissance de signes à partir de vidéos ou pour la synthèse de mouvements à partir de texte. Il serait certainement intéressant de s'affranchir du passage par le squelette 3D, et de produire directement du texte en français à partir de vidéos LS (pour la reconnaissance), ou des animations LS à partir de texte en français (pour la synthèse).

Enfin, il subsiste la question de l'alignement vidéo/texte qui n'est pas résolue. Il est nécessaire de l'aborder dans le langage pivot ou à défaut dans la langue vocale.

6. Conclusion et perspectives

Dans cet article, nous avons abordé les questions principales qui se posent pour la traduction et la synthèse en LS à partir d'énoncés textuels, et décrit les avancées principales de notre système TSL *SignCom*. Ce système intègre au niveau de sa conception les bases permettant de produire des contenus en LSF en respectant les procédés grammaticaux répertoriés. En particulier, il s'appuie sur la composition multimodale de segments de mouvements attachés d'une part à des composantes linguistiques de la LSF, et d'autre part à des contrôleurs de synthèse spécifiques. La possibilité d'éditer les énoncés, depuis le niveau phonologique jusqu'aux niveaux lexical, syntaxique et discursif permet ainsi de construire de nouveaux énoncés en LSF. Plusieurs modules de synthèse ont été intégrés au système initial. Tout d'abord, il devient possible de synthétiser des mécanismes flexionnels de la grammaire de la LSF qui s'appuient sur la spatialité et les dynamiques iconiques de cette langue. Notamment notre système permet de générer automatiquement des processus de type (locus/pointage/spatialisation) des signes, facilitant ainsi l'agencement des référents dans l'espace de signation. Il permet également de générer de manière flexible certains procédés propres à l'iconicité, comme les proformes manuelles lexicales ou syntaxiques. Un module de synthèse des expressions faciales a également été implémenté et évalué. Ce module génère automatiquement, à partir de données capturées, des données d'animation faciale qui peuvent être utilisées dans le système d'édition de manière similaire aux données de mouvement. Compte tenu du corpus enregistré, il est capable de synthétiser les qualités affectives et modales des mimiques faciales. De plus, un module spécifique de synthèse du mouvement des mains a été développé pour s'adapter aux exigences de précision et de rapidité des LS. La prise en compte d'un maillage volumétrique des mains, incorporant un modèle de squelette permet d'atteindre un niveau de précision jusqu'à présent inégalé, tout en évitant les interpénétrations des mains entre elles et avec les autres parties du corps. La synthèse du mouvement des mains ainsi que l'animation faciale ont donné des résultats très satisfaisants. Enfin, un procédé d'annotation automatique, appliqué aux configurations manuelles et expressions faciales permet d'accélérer la constitution des bases de données annotées pour la synthèse concaténative.

Si les avancée des systèmes TSL avec avatars signeurs sont très prometteuses, de nombreuses questions de recherche restent ouvertes. En particulier, l'un des enjeux majeurs consiste à mieux intégrer les travaux sur les formalismes de représentation des LS et les systèmes d'animation. De plus, la grande variabilité de ces langues visuo-gestuelles et la complexité des mécanismes de flexion qu'elles sous-tendent, nécessitent la mise en œuvre de processus de modélisation dédiés, d'un point de vue linguistique et animation, ce qui ouvre des voies de recherche encore peu explorées. Dans un futur proche, la possibilité de capturer de grands volumes de données et le développement des méthodes d'apprentissage automatique profond vont conduire à des systèmes automatiques de synthèse texte-vers-LS ou LS-vers-texte, dans la mesure où ils intègrent une connaissance linguistique des LS. Plus largement, cela ouvre des perspectives vers des systèmes de traduction automatique des langues vocales vers les LS ou vice-versa, ou d'une LS vers une autre.

7. Bibliographie

Battison R., *Lexical borrowing in American sign language*, ERIC, 1978.

Blondel M., Tuller L., Lecourt I., « Les pointés et l'acquisition de la morphosyntaxe en LSF », *La linguistique de la LSF : recherches actuelles. Silexicales 4*, In Berthonneau, A-M. and DAL, G. (eds), Univ. Lille 3, p. 17-32, 2004.

Brour M., Benabbou A., « ATLASLang MTS 1 : Arabic Text Language into Arabic Sign Language Machine Translation System », *Proc. Computer Science*, vol. 148, p. 236-245, 2019.

Chaaban H., le Gouiffès M., Braffort A., « Automatic Annotation and Segmentation of Sign Language Videos : Base-level Features and Lexical Signs Classification », *Int. Conf. VISIGRAPP 2021, Vol. 5*, p. 484-491, 2021.

Chételat-Pelé E., Braffort B., « Sign Language Corpus Annotation : toward a new Methodology », *LREC 2008, Marrakech, Morocco*, ELRA, 2008.

Cuxac C., *La langue des signes française (LSF) : les voies de l'iconocité (French) [French Sign Language : the iconicity ways]*, Faits de langues, Ophrys, 2000.

Dauriac B., Braffort A., Bertin-Lemée E., « Example-based Multilinear Sign Language Generation from a Hierarchical Representation », *Sign Language Translation and Avatar Technology ; Junction of the Visual and the Textual ; Challenges and Perspectives*, p. 21-28, 2022.

Duarte K., Motion Capture and avatars as Portals for Analyzing the Linguistic Structure of Sign Languages, PhD thesis, Université Bretagne Sud, 2012.

Ebling S., Glauert J. R., Kennaway J., Marshall I., Safar E., « Building a Swiss German Sign Language avatar with JASigning and Evaluating it among the Deaf community », *Universal Access in the Information Society*, vol. 15, p. 577-587, 2016.

Efthimiou E., Fontinea S., Hanke T., Glauert J., Bowden R., Braffort A., Collet C., Maragos P., Goudenove F., « Dicta-sign–sign language recognition, generation and modelling : a research effort with applications in deaf communication », *Workshop on the Representation and Processing of Sign Languages, LREC 2010*, p. 80-83, 2010.

Ekman P., Friesen W., *Facial Action Coding System : A Technique for the Measurement of Facial Movement.*, Consulting Psychologists Press, 1978.

Elliott R., Glauert J. R., Kennaway J., Marshall I., Safar E., « Linguistic modelling and language-processing technologies for Avatar-based sign language presentation », *Universal Access in the Information Society*, vol. 6, n° 4, p. 375-391, 2008.

Farooq U., Rahim M., Sabir N., Hussain A., Abid A., « Advances in machine translation for sign language : approaches, limitations, and challenges », *Neural Computing and Applications*, 11, 2021.

Filhol M., McDonald J., « Extending the AZee-Paula shortcuts to enable natural proform synthesis », *Workshop on the Representation and Processing of Sign Languages, LREC 2018*, Japan, 2018.

Filhol M., McDonald J., Wolfe R., « Synthesizing Sign Language by connecting linguistically structured descriptions to a multi-track animation system », *Int. Conf. on Universal Access in Human-Computer Interaction*, Springer, p. 27-40, 2017.

Garcia B., Sallandre M.-A., Schoder C., L'Huillier M.-T., « Typologie des pointages en Langue des Signes Française (LSF) et problématiques de leur annotation », *in* Boutora, L. et Braffort, A (eds), *TALN 2011*, Montpellier, France, p. 107-119, 2011.

Gibet S., « Building French Sign Language Motion Capture Corpora for Signing Avatars », *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, 2018.

Gibet S., Courty N., Duarte K., Le Naour T., « The SignCom System for Data-driven Animation of Interactive Virtual Signers : Methodology and Evaluation », *ACM Transactions on Interactive Intelligent Systems*, vol. 1, 2011.

Gibet S., Lebourque T., Marteau P., « High level Specification and Animation of Communicative Gestures », *Journal of Visual Languages and Computing*, vol. 12, p. 657-687, 2001.

Gibet S., Lefebvre-Albaret F., Hamon L., Brun R., Turki A., « Interactive editing in French Sign Language dedicated to virtual signers : requirements and challenges », *Universal Access in the Information Society*, vol. 15, n° 4, p. 525-539, 2016.

Glauert J., Elliott R., « Extending the SiGML notation : a progress report », *Int. Workshop on Sign Language Translation and Avatar Technology (SLTAT)*, vol. 23, 2011.

Héloir A., Gibet S., « A Qualitative and Quantitative Characterization of Style in Sign Language Gestures », *Gesture Workshop*, 2007.

Héloir A., Kipp M., « Real-time animation of interactive agents : Specification and realization », *Applied Artificial Intelligence*, vol. 24, n° 6, p. 510-529, 2010.

Huenerfauth M., Generating American Sign Language classifier predicates for English-to-ASL machine translation, PhD thesis, University of Pennsylvania, 2006.

Huenerfauth M., Lu P., Kacorri H., « Synthesizing and Evaluating Animations of American Sign Language Verbs Modeled from Motion-Capture Data », *SLPAT@Interspeech*, 2015.

Johnson R., Liddell S., « A segmental framework for representing signs phonetically », *Sign Language Studies*, vol. 11, n° 3, p. 408-463, 2011.

Johnston T., « The Lexical Database of AUSLAN (Australian Sign Language) », *Proceedings of the First Intersign Workshop : Lexical Databases*, Hamburg, 1998.

Kahlon N., Singh W., « Machine translation from text to sign language : a systematic review », *Universal Access in the Information Society*, 07, 2021.

Kennaway R., Glauert J. R., Zwitserlood I., « Providing signed content on the Internet by synthesized animation », *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 14, n° 3, p. 15, 2007.

Kipp M., Héloir A., Nguyen Q., « Sign Language Avatars : Animation and Comprehensibility », *Proceedings of the 10th International Conference on Intelligent Virtual Agents*, 2011.

Liddell S., Johnson R., « American Sign Language : The phonological base », *Sign Language Studies*, vol. 64, n° 6, p. 195-278, 1989.

Lombardo V., Nunnari F., Damiano R., « A virtual interpreter for the Italian sign language », *International Conference on Intelligent Virtual Agents*, Springer, p. 201-207, 2010.

McDonald J., Filhol M., « Natural synthesis of productive forms from structured descriptions of sign language », *Machine Translation*, vol. 35, n° 3, p. 363-386, 2021.

McDonald J., Wolfe R., Schnepp J., Hochgesang J., Jamrozik D., Stumbo M., Berke L., Bialek M., Thomas F., « An automated technique for real-time production of lifelike animations of American Sign Language », *Universal Access in the Information Society*, vol. 15, n° 4, p. 551-566, 2016.

Millet A., *Grammaire descriptive de la langue des signes française : dynamiques iconiques et linguistique générale*, UGA Editions, 2019.

Naert L., Larboulette C., Gibet S., « A survey on the animation of signing avatars : From sign representation to utterance synthesis », *Comput. Graph.*, vol. 92, p. 76-98, 2020.

Naert L., Larboulette C., Gibet S., « Motion synthesis and editing for the generation of new sign language content », *Machine Translation*, vol. 35, n° 3, p. 405-430, 2021.

Naert L., Reverdy C., Larboulette C., Gibet S., « Per channel automatic annotation of sign language motion capture data », *Workshop on the Representation and Processing of Sign Languages : Involving the Language Community, LREC 2018*, 2018.

Núñez-Marcos A., de Viñaspre O. P., Labaka G., « A survey on Sign Language machine translation », *Expert Systems with Applications*, vol. 213, p. 118993, 2023.

Prillwitz S., Zentrum H., *HamNoSys : Version 2.0 ; Hamburg Notation System for Sign Languages ; An Introductory Guide*, Signum-Verlag, 1989.

Reverdy C., Data-driven annotation and synthesis of facial expressions in French sign language, PhD thesis, Université Bretagne Sud, 2019.

Reverdy C., Gibet S., Larboulette C., « Optimal marker set for motion capture of dynamical facial expressions », *Proceedings of the 8th ACM SIGGRAPH Conference on Motion in Games*, ACM, p. 31-36, 2015.

Roelofsen F., Esselink L., Mende-Gillings S., Smeijers A., « Sign Language Translation in a Healthcare Setting », *In Translation and Interpreting Technology Online (TRITON)*, p. 110-124, 2021.

Sallandre M.-A., Garcia B., « Semiological Approach to Sign Languages and "gloss-based notations" : Issues related to SL sub-units annotation », *Hesperia : Anuario de Filología Hispánica*, vol. 22, p. 57-79, 2020.

Stokoe W. C., *Semiotics and Human Sign Language*, Walter de Gruyter Inc., 1972.

Stoll S., Camgoz N. C., Hadfield S., Bowden R., « Text2Sign : towards sign language production using neural machine translation and generative adversarial networks », *International Journal of Computer Vision*, vol. 128, p. 891-908, 2020.

Notes de lecture

Rubrique préparée par Denis Maurel

Université de Tours, LIFAT (Laboratoire d'informatique fondamentale et appliquée)

Marcel CORI. Le traitement automatique des langues en question. *Éditions Cassini*. 2021. 248 pages. ISBN : 978-2-84225-255-7.

Lu par **Thierry POIBEAU**

Laboratoire LATTICE

Le livre de Marcel Cori présente à la fois un panorama du traitement automatique des langues et une réflexion sur le domaine. Marcel Cori a longtemps été en poste à l'Université Denis Diderot, puis à Nanterre. Il a contribué à former de nombreux étudiants et l'auteur de ces lignes se souvient encore de ses cours d'algorithmique, à la fois bienveillants et exigeants, qui assuraient une base solide pour poursuivre dans le domaine.

L'ouvrage se compose de sept chapitres principaux. Le premier chapitre essaie de définir ce qu'est le Tal. Il est suivi de deux chapitres qui se font écho, sur le Tal théorique (chapitre 2) d'une part, et sur le Tal robuste (chapitre 3) de l'autre. Schématiquement, pour l'auteur, le Tal théorique s'appuie sur des théories linguistiques et s'oppose ainsi au Tal robuste fondé sur des approches statistiques, sans base linguistique explicite. Le chapitre se termine par une réflexion sur la notion de compréhension : les systèmes robustes reposent sur des stratégies d'optimisation statistique. Ils peuvent faire illusion, mais n'incluent pas de réels mécanismes de compréhension, au sens où ils n'ont pas une représentation formelle et non ambiguë du sens.

Le chapitre suivant (chapitre 4) présente un rapide historique. L'auteur montre que le domaine a toujours été tiraillé entre, d'une part, le besoin de montrer des outils et des applications pratiques justifiant la recherche aux yeux des financeurs et du grand public et, d'autre part, la nécessité de s'appuyer sur une assise théorique ferme, moins immédiatement valorisable, mais permettant les avancées pratiques à moyen et long terme. Pour l'auteur, ces deux tendances se sont matérialisées à travers deux courants : la linguistique formelle *versus* la linguistique de corpus, qui sont détaillées dans les chapitres suivants.

Le chapitre 5 aborde ainsi les liens entre Tal et linguistique formelle. L'auteur défend vivement ces liens : la formalisation est pour Marcel Cori l'assurance d'une approche réellement scientifique. L'auteur le souligne en trois mots : « réfutabilité, prédictivité, objectivité », chacun de ces termes étant ensuite justifié. Il va de soi que cette description s'oppose au Tal dit robuste, qui offre des techniques applicables à

une grande variété de problèmes, avec des performances solides, mais sans formalisation *a priori*.

Le chapitre 6 examine certains des grands formalismes syntaxiques passés, des ATN jusqu'à HPSG. Le chapitre se termine sur la notion de théorie linguistique : qu'est-ce qu'une théorie, qu'est-ce qu'un modèle, dans quelle mesure un formalisme doit-il être contraint ? Il est dommage que l'auteur n'aborde pas certaines questions, qui sourdent pourtant dans la tête du lecteur à l'issue du chapitre : dans quelle mesure peut-on parler d'objectivité d'un modèle, s'il y a autant de formalisations possibles d'un même phénomène ? L'auteur ne s'interroge pas vraiment sur la prolifération des modèles, sur la raison de cet état de fait et sur la compatibilité des formalisations entre elles. Peut-être plus important encore : ces formalisations semblent très éloignées des approches du Tal robuste. Une bonne formalisation ne devrait-elle pas être robuste ? Pourquoi cet écart entre Tal théorique et Tal robuste ? Il y a certainement dans d'autres domaines aussi des écarts entre théorie et pratique, mais peut-être pas autant qu'entre linguistique et Tal.

Le chapitre 7 revient de manière critique sur la linguistique de corpus. L'auteur souligne l'intérêt des corpus pour partir de données observables, affiner les jugements d'acceptabilité, etc. mais note que l'absence d'une forme en corpus « n'est en rien une preuve d'impossibilité ». On ne peut qu'être d'accord avec cela et la plupart des linguistes de corpus (si tant est qu'une telle catégorie existe) seraient d'accord avec cette observation. Les données de corpus sont à utiliser avec précaution, et permettent surtout l'observation et le comptage (la prévalence d'un sens, d'une expression). Les linguistes de corpus ne sont pas tellement intéressés par les impossibilités en langue (et plus généralement par des jugements binaires, qui restent peut-être plus prégnants en linguistique formelle, surtout dans les approches syntaxiques), mais davantage par la diversité de la production effective des locuteurs étudiés.

Le livre se termine évidemment avec une conclusion. L'auteur y souligne les limites de la théorie et surtout les mystères de la langue, qui résistent à la formalisation et qui obligent à rester modeste. Enfin, Marcel Cori souligne le plaisir à travailler dans un domaine si riche et aussi si humain (le langage étant, en un sens, le propre de l'homme, tout autant que le rire).

Pour élargir la discussion, on peut dire que l'ouvrage de Marcel Cori pose des questions intéressantes sur la place de la linguistique dans les approches automatisées depuis 50 ans.

Avant de discuter de cela, il faut souligner certains points qui constituent, à nos yeux, les limites de cet ouvrage, sans remettre en cause l'intérêt des questions soulevées. Le premier point est cette opposition entre linguistique formelle et linguistique de corpus. La linguistique de corpus a certes vécu de beaux jours dans les années 1990 en France (marqués par exemple par la parution du livre *Les linguistiques de corpus*, par A. Salem, B. Habert et A. Nazarenko en 1998), à un moment où des données massives commençaient à devenir disponibles et permettaient, entre autres choses, des jugements plus nuancés que les jugements binaires souvent pratiqués en linguistique jusque-là. Mais il semble que l'intérêt pour la linguistique de corpus soit largement retombé depuis (même si les corpus restent essentiels dans certains

domaines, comme en énonciation, en acquisition du langage, etc.). Surtout, le Tal ne se réclame plus depuis longtemps de la linguistique de corpus. Ou, plus exactement, la quasi-totalité du Tal est aujourd'hui passé à des approches par apprentissage sur d'énormes quantités de textes, que l'on peut appeler corpus, mais cela n'a plus grand chose à voir avec la linguistique de corpus telle que la pratiquent les linguistes de cette obédience. Les Talistes ont changé de chapelle depuis longtemps.

La deuxième limite de l'ouvrage est liée à la précédente : le paysage est bouleversé depuis une dizaine d'années maintenant par les approches neuronales, et depuis 2018 par les grands modèles de langage (avec l'apparition de BERT). Cette évolution peut être vue comme la simple suite de ce qui précède (c'est-à-dire comme le dernier avatar du Tal robuste), mais il semble quand même que l'ampleur des évolutions ces dernières années change un peu la donne. Ou peut-être pas, cependant il est un peu frustrant que ces développements récents et moins récents ne soient pas évoqués dans le livre de M. Cori. Il est vrai que l'ouvrage est paru fin 2020, il y a déjà plus de deux ans et que ce compte-rendu est de ce point de vue un peu tardif.

Mais, malgré cela, les réflexions de Marcel Cori restent d'actualité. La question de la formalisation reste en effet primordiale : on a tous été frappé par les performances des générateurs de textes (famille de modèles GPT, plus récemment ChatGPT, etc.). Ces modèles peuvent produire des paragraphes cohérents, répondre à des questions, mais ils ont aussi été qualifiés de « perroquets stochastiques » par E. Bender et ses collègues, dans la mesure où ils produisent du texte sur une base statistique, sans modèle du monde (c'est-à-dire, en particulier, sans notion de vérité) sinon celui offert par la redondance des informations sur Internet (en gros, « plus un fait est répété, plus il a de chances d'être vrai », ce qui n'est pas tout à fait satisfaisant, on en conviendra). Le couplage de ces modèles avec des bases de connaissances vérifiées humainement pourrait toutefois intervenir prochainement et changer partiellement la donne.

Une autre question intéressante, mais peu abordée dans l'ouvrage de Marcel Cori, est celle de la nature des informations à formaliser. Marcel Cori semble défendre une approche très syntaxique, alors que le succès des approches récentes est en grande partie lié aux techniques d'analyse distributionnelle à large échelle, permettant de dresser des profils sémantiques des unités lexicales (les fameux *word embeddings*). Or, sauf exception, la description du sens lexical n'a jamais été au centre des approches formelles. C'est d'ailleurs là, à notre avis, la raison principale de l'échec relatif de ces approches formelles : elles ont permis de mettre en avant les règles de grammaire et les autres règles, mais ce qui est formalisé dans la langue ne constitue qu'une infime fraction des connaissances nécessaires. Le sens est contextuel et aucune théorie n'a jamais fourni de représentation adéquate du contexte, qui reste une notion subjective et malléable. À l'inverse, les approches automatiques en fournissent une représentation très fine, calculée à partir de milliards d'exemples et encodée dans les milliards de paramètres des réseaux de neurones actuels, donnent autant d'infimes précisions sur le comportement des mots (puis des unités supérieures) en contexte. C'est cette finesse qui permet une relative précision, par exemple dans les applications de traduction automatique, malgré la diversité des traductions possibles en fonction de l'infinie diversité des contextes possibles.

En conclusion, on peut dire que le livre de Marcel Cori présente une réflexion solide et intéressante sur les rapports entre Tal et linguistique (et, plus particulièrement, linguistique formelle). Si les développements les plus récents du Tal ne sont pas évoqués dans l'ouvrage, le panorama historique qu'il offre reste tout à fait valable et, surtout, l'ouvrage ouvre des pistes de réflexion précieuses et tout à fait actuelles.

Mathilde JANIER, Patrick SAINT-DIZIER. Argument Mining : Linguistic foundations. *ISTE Editions*. 2019. 177 pages. ISBN : 978-1-786-30303-5.

Lu par **Léa GUIZOL**

ICT - Consultants, Bruxelles

Cet ouvrage introduit les bases du domaine, en commençant par l'argumentation tout court. Il présente les défis actuels à relever et finit sur un cas pratique. Il sera particulièrement utile à un novice ayant envie de découvrir la fouille d'arguments et l'argumentation. L'ouvrage est divisé en deux grandes parties. La première définit le domaine de l'argumentation en général, ainsi que dans le contexte de son analyse et de sa fouille. La seconde est plus technique et présente un cas concret.

Première partie

L'ouvrage commence par une bonne introduction à l'argumentation en général. Qu'est-ce ? À quoi cela sert ? Dans quels contextes ? Il y a un peu d'histoire de l'argumentation, qui remonte à nos racines grecques, de la philosophie et des définitions. Tout le vocabulaire est soigneusement défini.

Comment se construit une argumentation ? Comment tester la solidité d'un argument ? On y décrit les discours vus, d'une part, comme des graphes (les arguments et hypothèses sont des nœuds, reliés par des arêtes représentant les liens de défense et d'attaque entre eux) ; et, d'autre part, les discours sont analysés à travers le prisme de schémas d'argumentation (schémas se basant sur l'analogie, la causalité, les opinions communément admises…). Pour chaque schéma, des questions testant la validité du discours sont présentées.

Le chapitre 3 introduit les notions de base de la rhétorique, qui utilise à la fois des arguments et des éléments non verbaux pour convaincre une audience. On a le *logos* (logique du discours), l'*ethos* (crédibilité, position de l'orateur…) et le *pathos* (émotions induites chez l'auditoire). Les auteurs entrent plus profondément dans la structure du discours et des arguments, donc dans le *logos*. Ils s'intéressent aussi au vocabulaire indiquant la présence et l'intensité des arguments. Ils soulignent que les arguments ne sont pas toujours explicitement reliés entre eux.

Les défis et les besoins de la fouille d'arguments sont abordés. Un argument peut à la fois défendre et attaquer une hypothèse (*être dans un quartier animé apporte des sources de distraction… et du bruit*), ce qui nécessite :

– des préférences pour comparer l'attaque et le soutien à ladite hypothèse ;

– un besoin de culture et de connaissances générales sur le monde (représentation des connaissances) ;

– et de faire les liens entre les concepts pour comprendre les relations entre hypothèse et arguments qui s'y rapportent.

La fouille d'arguments pourrait permettre de faire des synthèses automatiques des textes d'opinion qui contiennent beaucoup d'arguments qui se recoupent et qui sont peu digestes.

Transition

Le chapitre 5 fait la transition entre l'argumentation (science du discours) et la fouille d'arguments (traitement automatique de la langue). Que peut-on attendre de la fouille d'arguments ? Il y a une introduction à la méthodologie générale de la fouille d'arguments.

Les auteurs insistent sur certaines de ses limites actuelles :

– l'importance de l'annotation et du temps humain qu'elle requiert ;

– l'ambiguïté du langage naturel et ses difficultés, même pour des experts humains ;

– les discours passés à l'écrit perdent la forme, l'intonation, les silences et le non-verbal (*pathos*), qui ne sont donc pas pris en compte et qui sont pourtant très importants.

Seconde partie

Comment fait-on des annotations pour analyser l'argumentation ? Cette étape est nécessaire et très consommatrice de temps humain. Le pourquoi et les bonnes pratiques sont expliqués. Les auteurs soulignent un paradoxe : il se peut que des arguments, même très importants, soient complètement implicites. Il y a de la recherche sur ce sujet. Les personnes qui veulent s'y lancer trouveront des références d'outils d'annotation.

Quels sont les domaines d'application et les usages de la fouille d'arguments ? Les principes de fonctionnement des systèmes de fouille d'arguments sont décrits et divisés en deux grandes familles. Les auteurs soulignent une forte intersection avec le Tal et l'apprentissage automatique et font un état de l'art comprenant des exemples de travaux et des pointeurs vers des conférences pertinentes sur le sujet.

Le chapitre 8 présente une approche concrète.

Le chapitre 9 donne des ouvertures et des perspectives sur les aspects non verbaux de l'argumentation, qui peuvent retourner le sens perçu d'un argument : images, émoticônes, aspects visuels ou sonores, et musique.

Conclusion

J'ai beaucoup apprécié le retour aux sources de l'argumentation et les définitions du domaine, ce qui est également intéressant pour les personnes s'intéressant seulement à la philosophie ou à l'art du discours. Cependant, un lexique reprenant les nombreuses définitions de l'ouvrage aurait été pratique.

Ce livre introduit au domaine de la fouille d'arguments pour toutes les personnes intéressées. Il est très clair, et les différentes notions sont introduites de façon progressive. Tout au long de l'ouvrage, on voit apparaître également les limites et les défis de la fouille d'arguments, ainsi que les liens avec d'autres disciplines de Tal et d'IA, ce qui permet d'avoir une vision large.

Masato HAGIWARA. Real-World Natural Language Processing. Practical applications with deep learning. *Manning Publications*. 2021. 336 pages. ISBN : 978-1-617-29642-0.

Lu par **Yannis HARALAMBOUS**

IMT Atlantique, UMR CNRS 6285 Lab-STICC

L'avènement[1] *de ChatGPT ne fait que confirmer que nous vivons dans l'ère du* deep learning*, une ère où les « miracles » sont quotidiens et où les étudiants néophytes sont persuadés qu'il suffit de choisir le bon BERT pour résoudre n'importe quel problème qui peut se poser en traitement automatique de la langue. Le* deep learning *a accaparé le monde du Tal au point qu'il a aujourd'hui – comme le titre de cet ouvrage l'indique – acquis le statut de* real-world*, de seule réalité dure comme fer.*

Introduction

L'apprentissage du *deep learning* comporte certains écueils structuraux. Tout d'abord, il y a trois manières de décrire les architectures de réseaux de neurones profonds :

(a) par des formules mathématiques, mais la théorie mathématique sous-jacente est assez complexe, ou du moins nécessite des formalismes d'apparence très complexe pour être expliquée ;

(b) par des diagrammes illustratifs qui montrent bien les flux d'information, mais ils deviennent vite très complexes, eux aussi, surtout qu'ils sont, le plus souvent, statiques alors que le « miracle » des neurones se produit lorsque les informations circulent, de haut en bas, de droite à gauche, constamment renouvelées, se rapprochant inexorablement de la valeur souhaitée ;

(c) par le code, souvent en Python, qui peut être assez complexe mais qui a l'immense avantage qu'on peut le « décortiquer », exécuter les lignes de code une par une en visualisant les entrées et les sorties.

Des ouvrages suivent la première approche : par des théorèmes et des démonstrations, on acquiert le calcul aux dérivées partielles, la théorie de la mesure, les équations différentielles et tout le bagage mathématique nécessaire pour comprendre le *deep learning*. Sans la moindre application. D'autres tombent dans l'autre extrême et sont remplis d'illustrations et de code, sans la moindre formule (à croire que leurs lecteurs ont loupé tous les cours de maths du collège...). Mais ils

[1] Nous avons volontairement choisi un terme à connotation religieuse.

fourmillent d'applications. Et, en parlant d'applications, on est passé au stade où les lecteurs découvrent le Tal en même temps que le *deep learning*. Cela permet de motiver l'apprentissage de l'un par celui de l'autre, et pour ceux qui possèdent des connaissances dans l'un des deux domaines, de se sentir en terrain connu pour mieux apprivoiser l'autre côté. L'ouvrage de Hagiwara est de ce type : il suppose, au départ, que le lecteur ne connaît ni le Tal, ni les réseaux de neurones profonds, et il le guide dans les deux directions.

Première partie de l'ouvrage

Le premier chapitre est une introduction au Tal. Il pose la question habituelle « Qu'est-ce que le Tal ? » mais aussi la question bien plus surprenante « Qu'est-ce qui n'est pas Tal ? ». On s'attend à une discussion sur les limites du Tal (multimodalité, langage intérieur, etc.) mais il n'en est rien. En réalité, il ne s'agit pas de contenu mais de méthode : le message que l'auteur veut faire passer est qu'on ne peut traiter des données linguistiques par un simple arbre de décision.

Après la description de quelques tâches de base (analyses morphologique et syntaxique, génération de texte) et de l'architecture-type d'une application de Tal, l'auteur passe, au deuxième chapitre, à la description détaillée et accompagnée du code Python d'une première application : une fouille de sentiments basée sur le corpus *Stanford Sentiment Treebank*. Il s'agit d'un corpus d'arbres syntaxiques à constituants, avec des annotations (entre 0 et 4) tant au niveau des mots, que des syntagmes et des phrases (même les signes de ponctuation ont été annotés !). L'exemple est bien choisi parce que le but recherché est facile à comprendre, les données sont disponibles en accès libre, et le principe de cette annotation est assez inhabituel pour alimenter des discussions (est-ce que cela a un sens d'annoter positivement/négativement tous les mots, y compris les mots grammaticaux ? Et pourquoi utiliser une seule dimension, alors que *SentiNet* en utilise deux ?). L'auteur en profite pour introduire les notions de plongement lexical et de réseau neuronal profond.

La bibliothèque de *deep learning* choisie par l'auteur pour tous les exemples de l'ouvrage est AllenNLP. Ce choix s'avère pertinent parce que AllenNLP est un bon compromis entre facilité d'utilisation et performance.

Les chapitres 3, 4 et 5 constituent une progression dans la complexité des réseaux neuronaux utilisés et des tâches envisagées. Dans le chapitre 3, l'auteur explique les deux types classiques de plongement lexical (CBOW et *skip-gram*), ainsi que l'approche *fastText* qui considère une fenêtre glissante sur les caractères des mots. En guise d'illustration, l'auteur prend un corpus de phrases anglaises (du très intéressant projet Tatoeba, une base de données multilingue de phrases) et cherche les phrases les plus proches d’une phrase donnée.

Dans le chapitre 4, on passe des réseaux qui opèrent sur les sacs de mots, à des réseaux qui gèrent des phrases. Le problème de la disparition du gradient est évoqué et les réseaux de type LSTM sont introduits. L'exemple choisi est celui de la détection de langue, en utilisant de nouveau Tatoeba comme corpus. Ce dernier a l'avantage de proposer un ensemble assez homogène de phrases dans un grand nombre de langues,

ce sont donc des conditions optimales pour que la détection de langue se passe comme une lettre à la poste.

Le chapitre 5 va plus loin en s'intéressant à l'annotation des mots d'une phrase, à travers les réseaux de type Seq2Seq, multicouches et bidirectionnels. Comme applications, l'auteur mentionne l'identification des parties du discours ainsi que la recherche d'entités nommées et de leurs types. Et c'est cette dernière qu'il illustre par un exemple. Le corpus utilisé provient de Kaggle. Mais le chapitre ne s'arrête pas là : l'auteur évoque le problème de la génération de texte et propose une solution à travers un modèle probabiliste de langue. Bien sûr, avec les moyens du bord et les faibles ressources utilisées, le résultat ressemble plutôt à un cadavre exquis qu'à du texte cohérent – ce que l'auteur ne manque pas de souligner.

La première partie de l'ouvrage est très pédagogique : on se pose des problèmes de plus en plus difficiles et on explique l'évolution des réseaux neuronaux pour faire face à cette difficulté croissante.

Deuxième et troisième partie de l'ouvrage

Les chapitres de la deuxième partie présentent des problématiques plus avancées et sont un peu plus décorrélés les uns des autres.

Le chapitre 6 présente les systèmes encodeur-décodeur, où l'information provenant d'une phrase est d'abord recueillie et traitée jusqu'à être condensée en un seul vecteur de nombres, et c'est ce vecteur qui est ensuite utilisé pour produire une autre phrase, par exemple une traduction de la première. L'auteur construit un traducteur élémentaire en prenant soin de bien souligner, à chaque étape, les faiblesses de l'approche. Le corpus utilisé est, encore une fois, Tatoeba. L'auteur profite des très faibles performances de ce traducteur simpliste pour parler des problèmes de traduction et de l'évaluation des systèmes de traduction. Il conclut le chapitre par une autre application des réseaux encodeur-décodeur : un système question-réponse, basé sur un corpus de dialogues.

Le chapitre 7 présente un type de réseau de neurones qui est principalement connu pour ses performances en reconnaissance de formes : les réseaux convolutifs. Ce type de réseaux fonctionne en appliquant une fenêtre glissante sur les données, en extrayant de chaque position de la fenêtre une nouvelle valeur. On obtient ainsi des couches successives de plus en plus petites, mais avec de l'information de plus en plus synthétique. On peut se demander à quel type de problème Tal correspond ce type de réseau de neurones. L'exemple que donne l'auteur est celui de la classification de textes et on conçoit bien que, comme une image, un texte peut être analysé sur plusieurs niveaux, ce qui fait que cette approche est pertinente.

Les chapitres 8 et 9 présentent la grosse artillerie des réseaux de neurones appliqués au Tal : le mécanisme d'attention, les transformeurs et l'apprentissage par transfert. Le mécanisme d'attention est très bien expliqué. Grâce aux illustrations très soignées, on comprend aisément l'utilité de chaque partie du mécanisme : pour chaque mot de la phrase du décodeur, on va apprendre au modèle à focaliser son attention sur tel ou tel mot de la phrase de l'encodeur. Quand on applique l'attention à une seule et même séquence de mots, on parle d'auto-attention (*self-attention*). Et comme on

n'arrête pas le progrès, on passe d'un à plusieurs niveaux d'auto-attention -et on obtient ainsi un transformeur. Les exemples donnés par l'auteur pour illustrer les transformeurs sont : un nouveau système de traduction anglais-espagnol, avec des performances bien meilleures, et un correcteur orthographique, entraîné sur un corpus produit par l'auteur. Dans la dernière section du chapitre 8, l'auteur laisse libre cours à son imagination et présente des idées de plus en plus exubérantes, idées qu'il met immédiatement en œuvre. Pour n'en citer qu'une, il utilise le réseau Seq2Seq de traduction pour « corrompre » un corpus de textes grammaticalement corrects, afin de se servir ensuite du résultat comme corpus d'entraînement pour son correcteur orthographique...

Dans le chapitre 9, il présente BERT (et ses camarades, ELMo, XLNet, RoBERTa, DistilBERT, ALBERT) et l'apprentissage par transfert. Les exemples donnés sont : la fouille de sentiments et la détection de relations rhétoriques entre phases (qu'il appelle « inférences »).

La troisième partie, de loin la plus courte, est une collection d'informations plus ou moins pratiques sur les réseaux de neurones et de conseils sur la mise en œuvre et l'optimisation de systèmes de Tal.

Conclusion

Si on n'a pas besoin des aspects mathématiques des réseaux de neurones, ce livre est une très bonne solution pour comprendre les tenants et les aboutissants des différents types de réseaux de neurones et de leurs applications au Tal. Les figures sont extrêmement soignées, et cela contribue à la bonne compréhension des notions complexes. Et le texte de Hagiwara est très agréable à lire, il est informel, mais ne manque pas de rigueur.

L'ouvrage est donc très adapté à l'auto-apprentissage des applications des réseaux de neurones au Tal mais aussi à l'enseignement. Attention toutefois, pour les avoir testés, les exemples donnés sont corrects, mais leur temps d'exécution dépend fortement du matériel utilisé. Un des derniers exercices a demandé plus d'une journée de calcul sur un Mac d'avant-dernière génération (processeur Intel), difficile donc de le proposer en TP. Mais, en fin de compte, ce problème est inhérent aux réseaux de neurones profonds : il faut trouver un compromis entre les petits corpus, qui donnent des temps d'exécution raisonnables, mais de très mauvais résultats, et les gros corpus, qui ne peuvent pas être traités par les machines de monsieur Tout-le-Monde.

Le seul risque qu'on prend en investissant du temps et de l'énergie dans cet ouvrage est la vitesse à laquelle avance le domaine en question…

Résumés de thèses et HDR

Rubrique préparée par Sylvain Pogodalla

Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
sylvain.pogodalla@inria.fr

Gabriele CHIGNOLI : gabrielechignoli@icloud.com

Titre : Les composantes de la parole dans la caractérisation phonétique du locuteur : étude sur la complémentarité et la redondance véhiculées des informations

Mots-clés : locuteur, caractéristiques, composantes, CNN, spontané, clustering, informativité, comparaison.

Title: *Speech Components in Phonetic Characterisation of Speakers: A Study on Complementarity and Redundancy of Conveyed Information*

Keywords: *speaker, characteristics, components, CNN, spontaneous, clustering, informedness, comparison.*

Thèse de doctorat en sciences du langage, Laboratoire de Phonétique et Phonologie, UMR 7018, Université Sorbonne Nouvelle, sous la direction de M. Cédric Gendrot (Pr, Université Sorbonne Nouvelle). Thèse soutenue le 15/09/2022.

Jury : M. Cédric Gendrot (Pr, Université Sorbonne Nouvelle, directeur), M. Damien Lolive (Pr, Université de Rennes, IRISA, rapporteur), Mme Ioana Vasilescu (CR HDR, CNRS, LISN, rapporteuse), Mme Cécile Fougeron (DR, CNRS, Laboratoire de Phonétique et Phonologie, UMR 7018, examinatrice), M. Jean-François Bonastre (Pr, Université d'Avignon, examinateur), Mme Christine Meunier (CR HDR, CNRS, présidente).

Résumé : *The decomposition of the speech signal into phonetically meaningful units allows the analysis of between- and within- speaker variations. These are components associated with characteristics whose nature relates to the physical, psychological and social aspects of a speaker. In this thesis, we compare perceptual characterisation results with a phonetic analysis and advanced modelling techniques through*

Convolutional Neural Networks (CNN). Two French corpora of read and spontaneous speech are the object of our studies from which some results have emerged that should be considered important for the speaker characterisation domain.

The characteristics allowing the description of variation of speech components occurring in different stimuli by a single speaker appear consistent with the phonetic measurements that play an important role in the separation of stimuli by different speakers. This allows creating multiple groups of speakers inside the studied population that are characterised by similar distributions of speech components. In this sense, we observe that source and filter characteristics are more important in the description of female speakers' variation, while voice quality characteristics such as breathiness and hoarseness have a greater impact on male speakers. This suggests that the characterisation of female speakers relies more on linguistic and articulation aspects, as well as the role of interlocutors in the conversation, i.e., the example of formant dispersion, whereas paralinguistic aspects, such as the level of confidentiality between speakers that changes in breathiness may convey, are retained for the characterisation of male speakers.

The perceptual responses confirm these tendencies, with the human-based clustering showing consistency with CNN- and phonetic-based results. In particular, the clustering analysis further highlights consistency of CNN results with the statistical analysis of speech components, supporting further application of these methods for phonetic studies.

The last highlight of this thesis concerns the role of Mel Frequency Cepstral Coefficients (MFCC) in comparison to classical phonetic measurements. These show a great adaptation to speakers' characteristics, relating to different aspects for female and male speakers, and for the multiple groups of speakers present in our population. Rather than being representative of a specific trait, MFCC are mainly linked to intensity and fundamental frequency for female speakers characterisation, while to the distributions of energy and low-level spectral shape for male speakers.

URL où le mémoire peut être téléchargé :
https://hal.science/tel-03911819

Ghazi FELHI : ghazi.felhi@gmail.com

Titre : Représentations de phrases interprétables avec autoencodeurs variationnels et attention

Mots-clés : autoencodeurs variationnels, transformeurs, interpretabilité, désenchevêtrement, apprentissage semi-supervisé, apprentissage non supervisé, modèle de langue, syntaxe.

Title: *Interpretable Sentence Representation with Variational Autoencoders and Attention*

Keywords: *variational autoencoders, transformers, interpretability, disentanglement, semi-supervised learning, unsupervised learning, language modeling, syntax.*

Thèse de doctorat en informatique, Laboratoire d'Informatique de Paris-Nord, Université Sorbonne Paris-Nord, sous la direction de Mme Adeline Nazarenko (Pr, Université Sorbonne Paris-Nord), M. Joseph Le Roux (MC, Université Sorbonne Paris-Nord) et M. Djamé Seddah (MC, Université Paris Sorbonne, Inria). Thèse soutenue le 26/02/2023.

Jury : Mme Adeline Nazarenko (Pr, Université Sorbonne Paris-Nord, codirectrice), M. Joseph Le Roux (MC, Université Sorbonne Paris-Nord, codirecteur), M. Djamé Seddah (MC, Université Paris Sorbonne, Inria, codirecteur), M. Benjamin Piwowarski (CR, CNRS, Institut des Systèmes Intelligents et de Robotique, rapporteur), M. François Yvon (DR, CNRS, LISN, rapporteur), M. Laurent Besacier (Pr, Université Grenoble Alpes, président).

Résumé : *Dans cette thèse, nous développons des méthodes pour améliorer l'interprétabilité de techniques récentes d'apprentissage de représentation en traitement automatique de langues (TAL) en prenant en compte la difficulté d'obtention de données annotées. Nous utilisons des autoencodeurs variationnels (VAE) afin d'apprendre avec peu de données des représentations interprétables. Pour notre première contribution, nous identifions et supprimons des composants inutiles du fonctionnement des VAE semi-supervisés, améliorant ainsi leur vitesse de calcul et facilitant leur conception. Notre deuxième et principale contribution consiste à utiliser des VAE et des transformeurs pour construire deux modèles qui permettent de séparer l'information dans les représentations latentes en concepts interprétables sans données annotées. Le premier modèle, ADVAE, est capable de représenter et de contrôler séparément des informations sur les rôles syntaxiques dans les phrases. Le second modèle, QKVAE, utilise des variables latentes séparées pour former des clés et des valeurs pour son décodeur transformeur et est capable de séparer les informations syntaxiques et sémantiques dans ses représentations neuronales. Dans des expériences de transfert, QKVAE a une performance compétitive par rapport aux modèles supervisés et une performance équivalente à un modèle supervisé utilisant 50 000 échantillons annotés. De plus, QKVAE montre une capacité améliorée de désenchevêtrement des rôles syntaxiques par rapport à ADVAE. De manière générale, notre travail montre qu'il est*

possible d'améliorer l'interprétabilité des architectures de pointe utilisées pour les modèles de langage avec des données non annotées.

URL où le mémoire peut être téléchargé :
https://arxiv.org/abs/2305.02810

Laura NORESKAL : laura.noreskal@outlook.fr

Titre : Erreurs dans les phrases coordonnées au sein des rédactions universitaires : typologie et détection

Mots-clés : erreurs, rédactions universitaires, détection automatique de l'erreur, phrase coordonnée, linguistique de corpus, TAL, classification supervisée.

Title: *Errors in Coordinated Sentences in Academic Writing: Typology and Detection*

Keywords: *errors, students writings, error detection, coordinated sentence, corpus linguistics, NLP, supervised classification.*

Thèse de doctorat en sciences du langage, MoDyCo, UMR 7114, UFR Phillia, Université Paris Nanterre, sous la direction de Mme Iris Eshkol-Taravella (Pr, Université Paris Nanterre) et de Mme Marianne Desmets (MC, Université Paris Nanterre). Thèse soutenue le 14/12/2022.

Jury : Mme Iris Eshkol-Taravella (Pr, Université Paris Nanterre, codirectrice), Mme Marianne Desmets (MC, Université Paris Nanterre, codirectrice), Mme Anne Abeillé (Pr, Université Paris Cité, présidente), Mme Frédérique Sitri (Pr, Université Paris-Est Créteil, rapporteuse), M. Olivier Kraif (Pr, Université Grenoble Alpes, rapporteur), Mme Sarah de Vogüé (MC, Université Paris Nanterre, examinatrice), Mme Silvia Adler (Pr, Université Bar-Ilan, Israël, examinatrice), M. Éric Villemonte de la Clergerie (CR, Inria, examinateur).

Résumé : *Face aux difficultés rédactionnelles rencontrées par les étudiants à leur entrée dans l'enseignement supérieur, une quinzaine d'universités françaises ont décidé de se réunir pour proposer des solutions de remédiation dans le cadre d'un projet nommé écri+ (ANR 17-NCUN-0015). Le projet écri+ a pour but de permettre aux étudiants francophones d'améliorer leurs compétences langagières en leur proposant des outils d'évaluation, de formation et de certification pour l'expression et la compréhension écrite du français. Parmi les difficultés observées, on retrouve les constructions syntaxiques complexes et les séquences phrastiques longues, avec des coordinations ou des juxtapositions. Ainsi, afin que les étudiants puissent s'autoformer sur la reconnaissance de ce type d'erreurs dans leurs textes, écri+ propose de mettre à leur disposition un outil de détection automatique d'erreurs dans les phrases coordonnées. En mêlant TAL, didactique et linguistique de corpus, cette recherche porte sur l'étude et la détection automatique des erreurs dans les constructions coordonnées issues des rédactions des étudiants. Après avoir constitué le corpus de rédactions composé de*

mémoires, rapports de stage, exercices et devoirs maison, nous avons procédé à l'analyse manuelle des données afin d'élaborer une typologie des erreurs réalisées dans les phrases coordonnées. La recherche réalisée a montré que les erreurs sont les plus présentes dans les productions non préparées telles que les exercices, et sont conditionnées par la taille des phrases, mais également par le nombre de coordonnants présents dans la phrase. Ensuite, le corpus a été annoté selon une typologie proposée et a été exploité pour le développement de l'outil de la détection automatique de ces erreurs. Dans un premier temps, l'outil développé prédit la classe, correcte ou erronée, pour une phrase coordonnée donnée. Dans un second temps, l'outil catégorise l'erreur reconnue, c'est-à-dire qu'il classe l'erreur parmi les 11 types proposés.

URL où le mémoire peut être téléchargé :
https://www.theses.fr/s241290

Mathilde REGNAULT : regnaultm@icloud.com

Titre : Annotation et analyse syntaxique de corpus hétérogènes : le cas du français médiéval

Mots-clés : annotation syntaxique, métagrammaire, français médiéval, ancien français, corpus hétérogène, grammaire d'arbres adjoints, parsing.

Title: *Syntactic Analysis and Parsing of Heterogeneous Corpora: The Case of Medieval French*

Keywords: *syntactic annotation, metagrammar, Medieval French, Old French, heterogeneous corpus, tree-adjoining grammar, parsing.*

Thèse de doctorat en sciences du langage, Lattice, UMR 8094, UFR Littérature, Linguistique, Didactique, Université Sorbonne Nouvelle, sous la direction de Mme Sophie Prévost (DR, CNRS, Lattice) et de M. Éric Villemonte de la Clergerie (CR, Inria). Thèse soutenue le 16/06/2022.

Jury : Mme Sophie Prévost (DR, CNRS, Lattice, codirectrice), M. Éric Villemonte de la Clergerie (CR, Inria, codirecteur), M. Sylvain Kahane (Pr, Université Paris Nanterre, rapporteur), Mme Laura Kallmeyer (Pr, Heinrich Heine Universität Düsseldorf, Allemagne, rapporteuse), Mme Béatrice Daille (Pr, Université de Nantes, examinatrice), Mme Annie Forêt (MC, Université de Rennes 1, examinatrice), M. Achim Stein (Pr, Universität Stuttgart, Allemagne, examinateur).

Résumé : *Le français médiéval couvre les états de langue d'ancien français (IX^{e} – XIII^{e} siècle) et de moyen français (XIV^{e} – XV^{e} siècle). Nous disposons de données annotées pour ces états de langue, dont le SRCMF, un corpus arboré d'ancien français. Il est cependant difficile d'obtenir plus de données annotées syntaxiquement, car les spécialistes sont peu nombreux et il n'existe pas encore d'outil dédié pour l'ensemble de la période. Développer ce genre d'outil permet d'obtenir des annotations plus facilement et d'en contrôler la qualité. Cependant, ce n'est pas une tâche simple*

parce que les différents états de langue sont soumis à la variation, due à plusieurs facteurs, notamment l'absence de norme graphique, la variation dialectale, la souplesse de l'ordre des mots, l'évolution de la morphologie et de la syntaxe (sur sept siècles), qui fait passer le français d'une langue SOV à une langue SVO. La nature des écrits se diversifie aussi à mesure que la littérature évolue et que le latin est délaissé au bénéfice du français comme langue administrative et juridique. Les données à analyser sont donc hétérogènes, ce qui rend difficile le traitement automatique, comme l'ont précédemment montré des expériences d'annotation morphosyntaxique sur le SRCMF.

Pour obtenir un parseur du français médiéval, nous proposons d'adapter la métagrammaire du français contemporain FRMG. Bien que les différents états de langue présentent des différences manifestes, les points communs sont suffisants pour rendre possible la modification d'un système existant pour obtenir un outil dédié. Les changements concernent essentiellement l'ordre des mots (constituants majeurs, modifieurs du nom, position des pronoms conjoints). Pour utiliser cet outil sur corpus, il est nécessaire d'enrichir le lexique d'ancien français OFrLex, d'une part pour obtenir une couverture lexicale satisfaisante sur les textes, et, d'autre part, pour y intégrer des informations syntaxiques et sémantiques nécessaires à l'analyse syntaxique.

Le développement d'un parseur symbolique vient, d'une part, de la volonté de justifier linguistiquement des analyses, ce qui permet de confronter notre compréhension de la syntaxe du français médiéval aux sorties du parseur, et, ainsi, de l'affiner. D'autre part, nous souhaitons nous servir de la fouille d'erreurs pour améliorer les divers composants de la chaîne (segmenteur, lexique, grammaire). Ce système est encore en développement, mais il nous permet déjà d'annoter des données de toute la période du français médiéval et de comparer ces analyses à celles de parseurs neuronaux, ce qui participe à orienter le travail de relecture vers des exemples difficiles à traiter.

URL où le mémoire peut être téléchargé :
https://theses.fr/s195294