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.