Traitement automatique des langues

Dialogue et systèmes de dialogue
Dialogue and Dialogue Systems

sous la direction de
Liesbeth Degand
Philippe Muller

Vol. 61 - n°3 / 2020

ATALA

Dialogue et systèmes de dialogue
Dialogue and Dialogue Systems

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

TAL
Vol.
61

n°3
2020

Dialogue et systèmes de dialogue
Dialogue and Dialogue Systems

ATALA

Traitement automatique des langues

Revue publiée depuis 1960 par l'Association pour le Traitement Automatique des LAngues (ATALA), avec le concours du CNRS, de l'Université Paris VII et de l'Université de Provence

ISSN 1965-0906

https ://www.atala.org/revuetal

Traitement automatique des langues

Secrétaire
Peggy Cellier - IRISA, INSA Rennes

Traitement automatique des langues

Volume 61 – n°3 / 2020

DIALOGUE ET SYSTÈMES DE DIALOGUE
DIALOGUE AND DIALOGUE SYSTEMS

Table des matières

Introduction to the Special Issue on Dialogue and Dialogue Systems

Liesbeth Degand* — **Philippe Muller****

* *Institute for Language and Communication, University of Louvain*
** *IRIT, University of Toulouse*

1. An expanding field with new questions

Recent progress in the field of NLP impacts all of its subfields and extends its domain of applications. Central to these developments is automated dialogue, either with chatbots (scripted, conversational, cognitive) or personal assistants, ubiquitous and widely distributed as services via smartphones or commercial websites.

At the same time, new written media continue to grow and most of them involve some sort of interaction: chats, forums, emails, microblogging, or collaborative instant messaging services. This progress and the generalization of natural language for interaction also make way for novel approaches taking into account multiple modalities (images, video) and the situation in which dialogues take place. In this context, all aspects of conversation analysis and dialogue system development are concerned, whether communication is oral or written, task-oriented or open.

The prevalence of neural network based approaches in NLP further influences approaches to dialogue, bringing into focus different matters, such as automatic generation of diverse and natural responses, although these approaches sometimes minimize the role of comprehension, and make it quite challenging to integrate linguistic and extra-linguistic context. New approaches also bring new problems that pervade machine-learning in general: black box models are not easy to explain, and blurring the lines between humans and machines may generate ethical quandaries—such as when conversational agents reproduce the biases and prejudices present in their training data—issues that generate a lot of concern and dedicated workshops.[1] Paucity

1. Such as the *Safety for Conversational AI Workshop*, `safetyforconvai.splashthat.com`.

of annotated data for supervised methods is another issue for machine learning approaches which motivates the collection of auxiliary data, semi-automatically annotated data, or artificial data far removed from real-world use cases. This can generate a gap between theory and practice that needs to be addressed.

For this special issue, the *TAL journal* invited contributions on all aspects of research related to the analysis of written or transcribed conversations, to the development and evaluation of dialogue systems, to data collection for all interaction modalities, and to ethical and social issues pertaining to dialogue and its applications. While not all of these aspects have been addressed in the contributions to this special issue, we will quickly review in this introduction the more important lines of research in Natural Language Processing and Computational Linguistics, both from the point of view of natural language conversation and dialogue analysis, and from that of the perspective of dialogue systems that interact with users. We then present this issue's accepted papers.

2. Analysis of dialogues

Computational analysis of linguistic interaction is not a new topic, as it went hand in hand with progress in speech recognition, which gave rise to interactive applications, and a demand for a more complex understanding of conversations. Another shift has taken place with the increase of written interactions that came with the widespread use of the internet: forums, chat rooms, microblogging, and emails are all manifestations of more or less synchronous dialogues involving two or more participants. This has provided a boost in available data as the volume of written conversations largely surpasses available speech data, and they are also much easier to process. This is then reflected in the growth of annotated corpora, for instance the Ubuntu IRC corpus (Kummerfeld *et al.*, 2019) which includes relations between utterances in a multi-party technical chat discussion, or the STAC corpus (Asher *et al.*, 2016), consisting of chat negotiation dialogues with annotated discourse and conversation structures. In parallel, a rising number of (corpus) linguistic studies have engaged in trying to uncover the linguistic specificities of "online discourse". Thus, according to Baron (2010), a "persistent question intriguing Internet researchers has been whether the stylistic features of CMC [computer-mediated communication] are more like those of informal speech or paradigmatic writing". In this strand of research, interactive text-based computer-mediated conversations have been shown to share many characteristics with informal spoken conversations, especially in terms of conversational and discourse mechanisms such as turn-taking, grounding, and coherence marking.

While the bridge between linguistic descriptive (corpus-based) research and computational work is not yet fully crossed, some of these linguistic findings have found their way to NLP research applications: identifying speech acts (Mohiuddin *et al.*, 2019), analyzing the structure of interactions (Shi *et al.*, 2019; Badene *et al.*, 2019), understanding the flow of information in context, among others. Other questions might be more relevant for oral data (speaker recognition for instance), or

show up in different ways in various types of interactions (disfluencies, monitoring, feedback). Written-text oriented models or annotation standards tend now to be more concerned with integrating oral phenomena, as is shown by the new Universal dependencies model for syntax (Sanguinetti *et al.*, 2020). New interesting problems also appear: with more speakers/writers involved, overlapping threads of conversation complicate understanding (Kummerfeld *et al.*, 2019).

The current (industrial) applications are numerous: for instance, a lot of companies provide chat interaction for Customer Relationship Management instead of telephone services, and are interested in the information they can thus gain about their customers in order to improve their interaction with them. Another example is the analysis of meetings and producing the minutes, which was first undertaken with the AMI corpus (Carletta, 2007) and the ICSI corpus (Janin *et al.*, 2003), and is now being revisited, also with progress made in automated summarization (Li *et al.*, 2019).

On a technical level, neural networks have enabled powerful intermediate, learnable representations at each level of a conversation: tokens, utterances and speech acts, and sequences of such, which pervade the predictive models applied to conversation aspects (Kumar *et al.*, 2018).

3. Dialogue systems

Of course, theoretical advances in dialogue models impact dialogue systems, but interactive dialogue systems include more than just the modelling of interactions. They generally consist of three main components: (1) understanding user input (either from speech or written text); (2) managing the interaction: keeping track of the dialogue state and planning actions; and (3) generating a linguistic output form such as text or speech to interact with the user.

The natural language understanding component is more directly tied to progress in dialogue modelling, but within dialogue systems it is usually tailored to a specific application, focusing on a particular part of the input or a classification of the goal of a speech act (sometimes called *intents*), with the goal of extracting predefined pieces of information relevant to the intent (*slot filling*), a task akin to semantic parsing in more general NLP, cf. the survey of recent work on these aspects by Louvan and Magnini (2020). There is also a lot of interest in so-called open-domain systems, which would have the ability to naturally interact with human participants and converse on any subject, but they raise a lot more issues that only partly concern applied systems: background knowledge, and emotional engagement (Huang *et al.*, 2020).

At the level of dialogue management, one can distinguish two tasks that together define the behaviour of the system and the way it responds to users: (a) dialog state tracking (DST), and (b) determining an optimal dialog policy (what is the best move in a given context). DST covers all models and representations of speaker and agent beliefs, goals, and the state of conversation (questions under discussion, common ground, commitments), see a survey by Williams *et al.* (2016). It is the subject of

an ongoing series of annual challenges[2], with varying subtasks. Recent approaches involve neural networks for their flexibility with respect to integrating different levels of representations (Mrkšić *et al.*, 2017). Dialog policy optimization is the planning and/or selection of dialogue actions, and their types and content, before generation of the linguistic output. Technical approaches have been based on reinforcement learning, first with partially observable Markov decision problems (Young *et al.*, 2013), superseded now by deep RL approaches, for instance Li *et al.* (2017).

These components can also be integrated into a single architecture, especially in neural models that try to enforce an end-to-end architecture, from user language input to system output, with intermediate latent representations for dialogue states and for the policy to follow, all supervised by the end result of the interaction with respect to the task. In this case the tasks are simple and focused, for example in the case of interactive question answering (Dhingra *et al.*, 2017; Wen *et al.*, 2017). Answering a user is then either based on dedicated Natural Language Generation (NLG) approaches, which in general is the problem of taking a formal representation to produce a well-formed linguistic output. As mentioned above, this can also be integrated in a general architecture in which it is the final output. In fact, recent work tends to develop end-to-end architectures, where the only input is the user's last utterance(s), in so-called sequence-to-sequence neural models. These are often trained on existing dialogue corpora or social media exchanges, where there is no clear "task" or objective (Zhao *et al.*, 2017; Chan *et al.*, 2019). The recent survey by Dusek *et al.* (2020) underlines the difficulty of evaluating end-to-end NLG systems; and while seq2seq models give excellent results on some metrics (naturalness), they also can lack in semantic fidelity and diversity in their outputs.

Evaluation is an important issue in dialogue systems in general, and a complicated one, as systems involve different components and different objectives, as shown by Deriu *et al.* (2021). There is thus a variety of automatic metrics and human evaluation procedures specific to the different subtasks mentioned above.

The variety of approaches and applications can sometimes make generalization difficult from one domain to another, between tasks and contexts of use. This is also reflected in the richness and diversity of available data useful for designing systems, and a good view is given in Serban *et al.* (2018). The prevalence of data-driven models also means there is less *a priori* control on the behaviour of the systems, which can lead to undesirable outcomes and raise ethical questions (Henderson *et al.*, 2018).

2. `https://dstc9.dstc.community/past-challenges`.

4. Papers

4.1. *Dialogue management with linear logic: the role of metavariables in questions and clarifications*

The paper by Maraev, Bernardy and Ginzburg focuses on the dialogue management component of dialogue systems: they are concerned by the modelling of dialogue states, consisting here in several elements: a set of *questions under discussion*, recording unresolved interactions (mostly questions waiting for an answer), the history of speech acts and their content, as interpreted by the system. The manager is also supposed to take care of an *agenda* of planned moves by the system. They focus on a type of interaction that is common in information-seeking conversations: question and answers, including embedded sequences of questions when *clarification questions* occur.

The model they present uses formal representations for these elements, and proposes to characterize updates of the dialogue state and its agenda as proof derivations in a linear logic, in which dialogue possible operations are linear logic formulas: processing a question, processing a potential answer, generating a clarification question when no unambiguous answer exists in the system database.

This gives an interesting formal framework to understand these dialogue moves, and gives a blueprint for a rule-based dialogue manager, as they implemented a prototype of the dialogue acts covered in the paper.

While rule-based formal approaches to dialog management were once very popular, they now tend to be in the shadow of probabilistic approaches, among them mostly neural approaches. There is nonetheless a growing awareness that the two kinds of approaches can benefit each other: empirical methods help achieving better coverage and robustness, while injecting knowledge in systems relying on supervised learning helps diminish the demands on huge amounts of data (Lison, 2015; Williams *et al.*, 2017).

4.2. *Situated meaning in multimodal dialogue: human-robot and human-computer interactions*

The paper by Pustejovsky & Krishnaswamy focuses on the important topic of “situated" conversation, where interaction between humans or between humans and a system takes into account the specific context of the interaction, including user location, and activities that connect users to each other and to their environment (e.g. a common task or a game), and non-linguistic interactions. A model of this kind of conversation must address different issues, most importantly linguistic references to the context (deixis), reasoning about the environment, tying the perceptions of agents and their actions to the conversation (Hunter *et al.*, 2018).

In order to study these sorts of interactions, the paper presents a system which provides a simulated physical environment for agent interactions and a few scripted

tasks involving the manipulation of objects. The system demonstrates the kind of knowledge that is needed to explain conversation moves and references to the situation in which the conversation takes place. To this end, it proposes an ontology for physical objects that makes explicit how they might be manipulated and discussed.

An important topic of the paper is how to address paralinguistic conversation, with gestures associated with speech acts: interpreting the other speaker's gestures jointly with the linguistic message. They thus provide a formal representation for the semantics of speech acts and accompanying gestures.

The interest of such a platform and associated models is two-fold: it provides a simulation environment to record situated conversations with a trace of the situation: knowledge of the environment, the geometry of objects and agents, and their gestures superimposed on the conversation content. Such a platform could prove useful in collecting rich conversational data. Furthermore, depending on the ease with which it might be configured, the platform could provide an environment and testable model for situated conversational agents.

4.3. *Comparaison linguistique et neuro-physiologique de conversations humain-humain et humain-robot*

The paper by Hallart, Maes, Spatola, Prévot and Chaminade addresses questions about the behaviour of speakers in a conversation, comparing a context involving an automated dialog agent (or perceived as such, in a Wizard-of-Oz experimental setup) and a more natural context with two humans. By observing various parameters, linguistic and conversation patterns, but also cognitive aspects through fMRI scans, experiments show how human behave differently when facing a perceived robot with limited linguistic capabilities, and how they adapt their linguistic behaviour to the agent. For instance, humans facing robots show a stronger lexical *alignment*, i.e. their vocabulary converges more towards the other conversant's vocabulary. In general, language complexity decreases when talking to a perceived robot.

Conversational aspects that are observed include speech time, interaction with the other speaker with feedback moves, specific discourse markers. Linguistic aspects are mostly related to language complexity: lexical, syntactic, and also *descriptive*, related to the use of adjectives and adverbs. The paper proposes or refines quantitative measures of those complexity types. An original aspect of this study is to compare the linguistic observations with fMRI scans to see what regions of the brain are more or less active during a conversation, and if differences can be identified when talking to another human or to a robot. In particular, lexical complexity seems correlated negatively with the activation of certain regions that would indicate cognitive resources (i.e. memory) are less mobilized when talking to a simple robot agent.

While it is only a perspective of the present work, studying linguistic and cognitive patterns of interaction raises potentially interesting lines of research to improve dialog agents and anticipate unforeseen or undesirable human reactions to a system.

Acknowledgements

We wish to thanks the editors-in-chief for their unfailing support, especially Sophie Rosset, the scientific committee for this special issue for their constructive work and reactivity: Nicholas Asher (IRIT, CNRS), Fred Bechet (Aix Marseille University), Christophe Cerisara (LORIA, CNRS), Nancy Chen (A*STAR, Singapore), Laurence Devillers (Paris Sorbonne University, LIMSI), Yannick Esteve (LIUM, University of Le Mans), Raquel Fernández (ILLC, University of Amsterdam), Kerstin Fischer (Hamburg University), Jonathan Ginzburg (Paris University), Casey Kennington (Boise State University), Nicolas Hernandez (LS2N, Nantes University), Julie Hunter (Linagora), Frédéric Landragin (CNRS, ENS), Pierre Lison (Norwegian Computing Center), Laurent Prévot (Aix Marseille University), Lina Maria Rojas Barahona (Orange Labs, Lannion), David Schlangen (Bielefeld University), Noël Nguyen Trong (Aix Marseille University), and finally Kate Thompson for proof-reading and suggestions.

5. References

Asher N., Hunter J., Morey M., Benamara F., Afantenos S. D., "Discourse Structure and Dialogue Acts in Multiparty Dialogue: the STAC Corpus", *in* N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (eds), *Proceedings of the 10th International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*, European Language Resources Association (ELRA), 2016.

Badene S., Thompson K., Lorré J.-P., Asher N., "Weak Supervision for Learning Discourse Structure", *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, Hong Kong, China, p. 2296-2305, November, 2019.

Baron N. S., "Discourse structures in Instant Messaging: The case of utterance breaks", *Language@Internet*, 2010.

Carletta J., "Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus", *Lang. Resour. Evaluation*, vol. 41, n° 2, p. 181-190, 2007.

Chan Z., Li J., Yang X., Chen X., Hu W., Zhao D., Yan R., "Modeling Personalization in Continuous Space for Response Generation via Augmented Wasserstein Autoencoders", *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, Hong Kong, China, p. 1931-1940, November, 2019.

Deriu J., Rodrigo Á., Otegi A., Echegoyen G., Rosset S., Agirre E., Cieliebak M., "Survey on evaluation methods for dialogue systems", *Artif. Intell. Rev.*, vol. 54, n° 1, p. 755-810, 2021.

Dhingra B., Li L., Li X., Gao J., Chen Y.-N., Ahmed F., Deng L., "Towards End-to-End Reinforcement Learning of Dialogue Agents for Information Access", *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

Papers), Association for Computational Linguistics, Vancouver, Canada, p. 484-495, July, 2017.

Dusek O., Novikova J., Rieser V., "Evaluating the state-of-the-art of End-to-End Natural Language Generation: The E2E NLG challenge", *Comput. Speech Lang.*, vol. 59, p. 123-156, 2020.

Henderson P., Sinha K., Angelard-Gontier N., Ke N. R., Fried G., Lowe R., Pineau J., "Ethical Challenges in Data-Driven Dialogue Systems", *in* J. Furman, G. E. Marchant, H. Price, F. Rossi (eds), *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2018, New Orleans, LA, USA, February 02-03, 2018*, ACM, p. 123-129, 2018.

Huang M., Zhu X., Gao J., "Challenges in Building Intelligent Open-domain Dialog Systems", *ACM Trans. Inf. Syst.*, vol. 38, n° 3, p. 21:1-21:32, 2020.

Hunter J., Asher N., Lascarides A., "A formal semantics for situated conversation", *Semantics and Pragmatics*, 2018.

Janin A., Baron D., Edwards J., Ellis D., Gelbart D., Morgan N., Peskin B., Pfau T., Shriberg E., Stolcke A., Wooters C., "The ICSI Meeting Corpus", *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '03, Hong Kong, April 6-10, 2003*, IEEE, p. 364-367, 2003.

Kumar H., Agarwal A., Dasgupta R., Joshi S., "Dialogue Act Sequence Labeling Using Hierarchical Encoder With CRF", *in* S. A. McIlraith, K. Q. Weinberger (eds), *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, AAAI Press, p. 3440-3447, 2018.

Kummerfeld J. K., Gouravajhala S. R., Peper J. J., Athreya V., Gunasekara C., Ganhotra J., Patel S. S., Polymenakos L. C., Lasecki W., "A Large-Scale Corpus for Conversation Disentanglement", *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, p. 3846-3856, July, 2019.

Li M., Zhang L., Ji H., Radke R. J., "Keep Meeting Summaries on Topic: Abstractive Multi-Modal Meeting Summarization", *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, p. 2190-2196, July, 2019.

Li X., Chen Y.-N., Li L., Gao J., Celikyilmaz A., "End-to-End Task-Completion Neural Dialogue Systems", *Proceedings of the 8th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Asian Federation of Natural Language Processing, Taipei, Taiwan, p. 733-743, November, 2017.

Lison P., "A hybrid approach to dialogue management based on probabilistic rules", *Comput. Speech Lang.*, vol. 34, n° 1, p. 232-255, 2015.

Louvan S., Magnini B., "Recent Neural Methods on Slot Filling and Intent Classification for Task-Oriented Dialogue Systems: A Survey", *in* D. Scott, N. Bel, C. Zong (eds), *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, International Committee on Computational Linguistics, p. 480-496, 2020.

Mohiuddin T., Nguyen T.-T., Joty S., "Adaptation of Hierarchical Structured Models for Speech Act Recognition in Asynchronous Conversation", *Proceedings of the 2019 Conference of*

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (Volume 1: Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, p. 1326-1336, June, 2019.

Mrkšić N., Ó Séaghdha D., Wen T.-H., Thomson B., Young S., "Neural Belief Tracker: Data-Driven Dialogue State Tracking", *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Vancouver, Canada, p. 1777-1788, July, 2017.

Sanguinetti M., Bosco C., Cassidy L., Çetinoğlu Ö., Cignarella A. T., Lynn T., Rehbein I., Ruppenhofer J., Seddah D., Zeldes A., "Treebanking User-Generated Content: A Proposal for a Unified Representation in Universal Dependencies", *Proceedings of the 12th Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, p. 5240-5250, May, 2020.

Serban I. V., Lowe R., Henderson P., Charlin L., Pineau J., "A Survey of Available Corpora For Building Data-Driven Dialogue Systems: The Journal Version", *Dialogue Discourse*, vol. 9, n° 1, p. 1-49, 2018.

Shi W., Zhao T., Yu Z., "Unsupervised Dialog Structure Learning", *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (Volume 1: Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, p. 1797-1807, June, 2019.

Wen T.-H., Vandyke D., Mrkšić N., Gašić M., Rojas-Barahona L. M., Su P.-H., Ultes S., Young S., "A Network-based End-to-End Trainable Task-oriented Dialogue System", *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Valencia, Spain, p. 438-449, April, 2017.

Williams J. D., Asadi K., Zweig G., "Hybrid Code Networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning", *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Vancouver, Canada, p. 665-677, July, 2017.

Williams J. D., Raux A., Henderson M., "The Dialog State Tracking Challenge Series: A Review", *Dialogue Discourse*, vol. 7, n° 3, p. 4-33, 2016.

Young S., Gašić M., Thomson B., Williams J. D., "POMDP-Based Statistical Spoken Dialog Systems: A Review", *Proceedings of the IEEE*, vol. 101, n° 5, p. 1160-1179, 2013.

Zhao T., Zhao R., Eskenazi M., "Learning Discourse-level Diversity for Neural Dialog Models using Conditional Variational Autoencoders", *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Vancouver, Canada, p. 654-664, July, 2017.

Situated Meaning in Multimodal Dialogue: Human-Robot and Human-Computer Interactions

James Pustejovsky* — **Nikhil Krishnaswamy****

* *Department of Computer Science, Brandeis University*
** *Department of Computer Science, Colorado State University*

ABSTRACT. *The demand for more sophisticated natural human-computer and human-robot interactions is rapidly increasing, as users become more accustomed to conversation-like interactions with their devices. This requires not only the robust recognition and generation of expressions through multiple modalities (language, gesture, vision, action), but also the encoding of situated meaning: (a) the situated grounding of expressions in context; (b) an interpretation of the expression contextualized to the dynamics of the discourse; and (c) an appreciation of the actions and consequences associated with objects in the environment. In this paper, we introduce VoxWorld, a multimodal simulation platform for modeling human-computer interactions. It is built on the language VoxML, and offers a rich platform for studying the generation and interpretation of expressions, as conveyed through multiple modalities, including: language, gesture, and the visualization of objects moving and agents acting in their environment.*

RÉSUMÉ. *La demande d'interactions naturelles homme-ordinateur et homme-robot plus sophistiquées augmente rapidement, car les utilisateurs s'habituent davantage aux interactions de type conversation avec leurs appareils. Cela nécessite non seulement la reconnaissance et la génération robustes d'expressions à travers de multiples modalités (langage, geste, vision, action), mais aussi l'encodage du sens situé : (a) l'ancrage situé des expressions dans le contexte; (b) une interprétation de l'expression contextualisée à la dynamique du discours; et (c) une appréciation des actions et des conséquences associées aux objets dans l'environnement. Nous présentons VoxWorld, une plateforme de simulation multimodale pour la modélisation des interactions homme-machine. Il est construit sur le langage VoxML et offre une plate-forme riche pour étudier la génération et l'interprétation d'expressions, telles qu'elles sont véhiculées à travers de multiples modalités, notamment : le langage, le geste et la visualisation des objets en mouvement et des agents agissant dans leur environnement.*

KEYWORDS: *Multimodal dialogue, affordances, qualia structure, continuations, gesture, simulations, common ground, situated meaning, semantic grounding, referring expressions.*

MOTS-CLÉS : *Dialogue multimodal, affordances, structure qualia, continuations, geste, simulations, terrain d'entente, sens situé, ancrage sémantique, expressions de référence.*

1. Introduction

When humans communicate with each other through language, there is a shared understanding of both an utterance meaning (content) and the speaker's meaning in the specific context (intent). The ability to link these two is the act of situationally grounding meaning to the local context, typically referred to as "establishing the common ground" between interlocutors (Stalnaker, 2002; Asher, 1998). Language use may reflect only a subset of all properties of the current situation, where a full description may be impossible or at least unwieldy. Some kinds of information may in fact be more efficiently communicated using other modalities, such as gesture (e.g., deixis for pointing), demonstration or action, images, or some other visual modality. A central component to the contextualized interpretation of meaning in a discourse is the situational determination of the meanings of expressions given the common ground. It is this notion of *situated meaning* that is missing in most current human-computer and human-robot interaction models, and the focus of the present paper.

In this paper, we argue that the problem of situational awareness and the creation of *situated meaning* in discourse involves at least three components: (a) the situated *grounding* of expressions in context; (b) an interpretation of the expression contextualized to the *dynamics* of the discourse; and (c) an appreciation of the *actions and consequences* associated with objects in the environment. In Section 2, we expand on these aspects of meaning in some detail, and then in Section 3, we adopt the modeling language, VoxML, designed to encode non-linguistic, multimodal aspects of meaning associated with concepts. In section 4, we present a computational framework, VoxWorld, within which these components are operationalized to facilitate multimodal communication between humans and robots or computers. Section 5 outlines a framework within which to interpret multimodal expressions, while Section 6 presents experimental evidence from single and mixed modality dialogues, illustrating the different ways in which meaning is situated in goal-directed dialogues.

2. Interactions in the Common Ground

There has been a growing interest in the Human-Robot Interaction community on how to contextually resolve ambiguities that may arise from communication in situated dialogues, from earlier discussions on how HRI dialogues should be designed (Fischer, 2011; Scheutz *et al.*, 2011), how perception and grounding can be integrated into language understanding (Landragin, 2006), to recent work on task-oriented dialogues (Williams *et al.*, 2019). This is the problem of identifying and modifying the *common ground* between speakers (Clark and Brennan, 1991; Stalnaker, 2002; Asher, 1998). It has long been recognized that an utterance's meaning is subject to contextualized interpretation; this is also the case with gestures in task-oriented dialogues. E.g., depending on the situation, an oriented hand gesture could refer either to an action request ("move it") or a dismissive response ("forget it") (Williams *et al.*, 2019). Even a request for action can be underspecified, denoting either a continuous movement or a movement to a specific location. Similarly, depending on the situation, the definite description in the command "Open the box." may uniquely refer or not, depending on how many boxes are in the context. These and similar miscommunications or the need for clarification in dialogue have been

called *situated grounding problems* (Marge and Rudnicky, 2013), and can be viewed as problematic in a model that appeals to and encodes both a visual modality and situational information into the dialogue state. What the occurrence of these issues makes apparent is the complexity underlying the interpretation of referential expressions in actual situated dialogues. The richness provided by situationally grounding computer or robot behaviors brings to the surface interpretive questions similar to those of a human in the same scenario.

Some recent efforts have been made to provide contextual grounding to linguistic expressions. For example, work on "multimodal semantic grounding" within the natural language processing and image processing communities has resulted in a number of large corpora linking words or captions with images (cf. Chai *et al.* (2016)). In this paper, we argue that language understanding and linking to abstract instances of concepts in other modalities is insufficient; *situated grounding* entails knowledge of situation and contextual entities beyond that provided by a multimodal linking approach (cf. Kennington *et al.* (2013)).

Actual situated meaning is much more involved than aligning captions and bounding boxes in an image: e.g., Hunter *et al.* (2018) discuss the contribution of non-linguistic events in situated discourse, and also whether they can be the arguments to discourse relations. Similarly, it is acknowledged that gesture is part of either the direct content of the utterance (Stojnić *et al.*, 2019) or cosuppositional content (Schlenker, 2020). Hence, we must assume that natural interactions with computers and robots have to account for interpreting and generating language and gesture.

Figure 1. *Mother and son interacting in a shared task of icing cupcakes.*

Consider the joint activity shown in Fig. 1 above between a mother and her son, where they are engaged in icing cupcakes in a kitchen setting. The dialogue in Fig. 2 illustrates some possible multimodal expressions used in such a context of joint activity between two agents.

SITUATED MEANING IN A JOINT ACTIVITY

– SON: *Put it there (gesturing with co-attention)?*
– MOTHER: *Yes, go down for about two inches.*
– MOTHER: *OK, stop there. (co-attentional gaze)*
– SON: *Okay. (stops action)*
– MOTHER: *Now, start this one (pointing to another cupcake).*

Figure 2. *Dialogue.*

Viewed as a multi-agent collaborative task interaction, there are some obvious elements constituting the common ground between the two agents in Fig. 1. These include reference to: the participants (agents); shared beliefs and assumptions; shared goals and intentions; the accompanying objects in the situation; the shared perception

of these objects; and the surrounding space within which the situation unfolds. Some of these elements are given below in Fig. 3.

Agents	mother, son
Shared goals Beliefs, desires, intentions	baking, icing Mother knows how to ice, icing goes on cupcakes, etc. Mother is teaching son
Objects	cupcakes, plate, knives, pastry bag, icing, gloves
Shared perception	the objects on the table
Shared Space	kitchen

Figure 3. *Elements from the common ground for Figure 1.*

From this example, it is apparent that we can identify three core aspects of meaning that contribute to the common ground in a multimodal dialogue:

1) *co-situatedness* and *co-perception* of the agents, such that they can interpret the same situation from their respective frames of reference. This might be a human and an avatar perceiving the same virtual scene from different perspectives; or a human sharing the perspective of a robot as it navigates through a disaster zone;

2) *co-attention* of a shared situated reference, which allows more expressiveness in referring to the environment (i.e., through language, gesture, visual presentation, etc.). The human and avatar might refer to objects in multiple modalities with a common model of differences in perspective-relative references (e.g., "your left, my right"); or the human sharing the robot's perspective might be able to direct its motion using reference in natural language ("go through the second door on the left") or gesture ("go this way," with pointing);

3) *co-intent* of a common goal, such that misaligned relationships between agents reflect a breakdown in the common ground. A human and avatar interacting around a table might seek to collaborate to build a structural pattern known to one or both of them; or the human and robot sharing perspective both have a goal to free someone trapped behind a door in a fire. The robot informs the human about the situation and the human helps the robot problem-solve in real time until the goal is achieved.

What this suggests is that any robust communication between humans and computers or robots will require at least three capabilities: (a) a robust recognition and generation within multiple modalities; (b) an understanding of contextual grounding and co-situatedness in the conversation; and (c) an appreciation of the consequences of behavior and actions taking place throughout the dialogue. To this end, in our work, we have developed a platform making use of semantically interpreted *multimodal simulations*, which provides an approach to modeling human-computer communication by both situating and contextualizing the interaction, thereby visually demonstrating what the co-agent computer or robot is hearing, seeing, thinking, and doing. This platform is based on VoxML, a modeling language for encoding traditionally non-linguistic, multimodal, aspects of meaning associated with the objects that we encounter, manipulate, and explore in our environment. We turn to this discussion in the next section.

3. VoxML: Encoding Knowledge of Action and Behavior

Here we argue that a significant part of any model for situated communication is an encoding of the semantic type, functions, purposes, and uses introduced by the objects under discussion. I.e., a semantic model of perceived *object teleology*, as introduced by Qualia Structure, for example (Pustejovsky, 1995), as well as *object affordances* (Gibson, 1977) is needed to help ground expression meaning to speaker intent.

Objects under discussion in discourse (cf. Ginzburg (1996)) can be partially contextualized through their semantic type and their qualia structure: e.g., a food item has a TELIC value of eat, a pencil, a TELIC of $write$, a box, a CHAIR of sit_in, and so forth. However, while an artifact may be designed for a specific purpose, this can only be achieved under specific circumstances. To account for this context-dependence, Pustejovsky (2013) enriches the lexical semantics of words denoting artifacts (the TELIC role specifically) by introducing the notion of an object's *habitat*, which encodes these circumstances. For example, an object, x, within the appropriate context $\mathcal{C}$, performing the action π will result in the intended or desired resulting state, $\mathcal{R}$, i.e., $\mathcal{C} \rightarrow [\pi]\mathcal{R}$. That is, if the habitat $\mathcal{C}$ (a set of contextual factors) is satisfied, then every time the activity of π is performed, the resulting state $\mathcal{R}$ will occur. The precondition context $\mathcal{C}$ is necessary to specify, since this enables the local modality to be satisfied.

The habitat for an object is situated within an *embedding space* and then contextualized within it. For example, in order to use a glass to drink from, the concavity has to be oriented upward, the interior must be accessible, and so on. Similarly, a chair must also be oriented up, the seat must be free and accessible, it must be large enough to support the user, etc. An example of what the resulting knowledge structure for the habitat of a chair is shown below, where these constraints are superscripted with " *".

These distinctions in habitats facilitate both Gibsonian and telic affordances and transfer learning of Gibsonian affordances relies on information taken from telic affordances (its use for sitting), and vice versa (see Section 6.4): below, the F and C values specify size and part structure, respectively.

(1) $\lambda x \left[\begin{array}{l} \textbf{chair(x)} \\ \text{F} = [phys(x), on(x, y_1)^*, in(x, y_2)^*, clear(x_1)^*, orient(x, up)^*, \\ \qquad support(x_1, y_3)^*] \\ \text{C} = [seat(x_1), back(x_2), legs(x_3)] \\ \text{T} = \lambda z \lambda e[\mathcal{C} \rightarrow [sit(e, z, x)]\mathcal{R}_{sit}(x)] \\ \text{A} = [made(e', w, x)] \end{array} \right]$

The notion of habitat and the attached behaviors that are associated with an object are further developed in Pustejovsky and Krishnaswamy (2016), where an explicit connection to Gibson's ecological psychology is made, along with a direct encoding of the *affordance structure* for the object (Gibson, 1977). The affordance structure available to an agent, when presented with an object, is the set of actions that can be performed with it. We refer to these as GIBSONIAN affordances, and they include "grasp", "move", "hold", "turn", etc. This is to distinguish them from more goal-directed, intentionally situated activities, what we call TELIC affordances.

VoxML (Visual Object Concept Modeling Language) is a modeling language for constructing 3D visualizations of concepts denoted by natural language expressions,

and is being used as the platform for creating multimodal semantic simulations in the context of human-computer and human-robot communication (Pustejovsky and Krishnaswamy, 2016; Krishnaswamy and Pustejovsky, 2016). It adopts the basic semantic typing for objects and properties from Generation Lexicon and the dynamic interpretation of event structure developed in Pustejovsky and Moszkowicz (2011), along with a continuation-based dynamic interpretation for both sentence and discourse composition (De Groote, 2001; Barker and Shan, 2014; Asher and Pogodalla, 2010).

VoxML forms the scaffolding we use to encode knowledge about objects, events, attributes, and functions by linking lexemes to their visual instantiations, termed the "visual object concept" or *voxeme*. Voxemes representing humans or IVAs are lexically typed as *agents*, but agents, due to their embodiments, ultimately inherit from physical objects and so fall under objects in the taxonomy. In parallel to a lexicon, a collection of voxemes is termed a *voxicon*. There is no requirement on a voxicon to have a one-to-one correspondence between its voxemes and the lexemes in the associated lexicon, which often results in a many-to-many correspondence. That is, the lexeme *plate* may be visualized as a [[SQUARE PLATE]], a [[ROUND PLATE]], or other voxemes, and those voxemes in turn may be linked to other lexemes such as *dish* or *saucer*. Each voxeme is linked to either an object geometry, a program in a dynamic semantics, an attribute set, or a transformation algorithm, which are all structures easily exploitable in a rendered simulation platform.

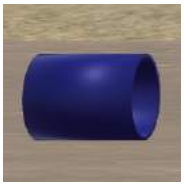


Figure 4. *Cup in habitat allowing rolling.*

An OBJECT voxeme's semantic structure provides *habitats*, which are situational contexts or environments conditioning the object's *affordances*, which may be either "Gibsonian" affordances (Gibson, 1977) or "Telic" affordances (Pustejovsky, 1995; Pustejovsky, 2013). A habitat specifies how an object typically occupies a space. When we are challenged with computing the embedding space for an event, the individual habitats associated with each participant in the event will both define and delineate the space required for the event to transpire. Affordances are used as attached behaviors, which the object either facilitates by its geometry (Gibsonian) or purposes for which it is intended to be used (Telic). For example, a Gibsonian affordance for [[CUP]] is "grasp," while a Telic affordance is "drink from." This allows procedural reasoning to be associated with habitats and affordances, executed in real time in the simulation, inferring the complete set of spatial relations between objects at each frame and tracking changes in the shared context between human and computer.

Indeed, object properties and the events they facilitate are a primary component of situational context. In Fig. 4, we understand that the cup in the orientation shown can be *rolled* by a human. Were it not in this orientation, it might be able to be only *slid* across its supporting surface (cf. (2)). This voxeme for [[CUP]] gives the object appropriate lexical predicate and typing (a *cup* is a PHYSICAL OBJECT and an ARTIFACT). It denotes that the cup is roughly cylindrical and concave, has a surface and an interior, is symmetrical around the Y-axis and across associated planes (VoxML

adopts 3D graphics convention where the Y-axis is vertical), and is smaller than and movable by the agent. The remainder of VoxML typing structure is devoted to habitat and affordance structures, which we discuss below.

(2) Objects encoding semantic type, habitat, and affordances:

$$\left[\begin{array}{l}\textbf{cup}\\ \text{LEXICAL} = \left[\begin{array}{l}\text{PREDICATE} = \textbf{cup}\\ \text{TYPE} = \textbf{physobj, artifact}\end{array}\right]\\ \text{TYPE} = \left[\begin{array}{l}\text{HEAD} = \textbf{cylindroid[1]}\\ \text{COMPONENTS} = \textbf{surface, interior}\\ \text{CONCAVITY} = \textbf{concave}\\ \text{ROTATIONAL\_SYMMETRY} = \{Y\}\\ \text{REFLECTION\_SYMETRY} = \{XY, YZ\}\end{array}\right]\\ \text{HABITAT} = \left[\begin{array}{l}\text{INTRINSIC} = [2]\left[\begin{array}{l}\text{CONSTR} = \{Y > X, Y > Z\}\\ \text{UP} = align(Y, \mathcal{E}_Y)\\ \text{TOP} = top(+Y)\end{array}\right]\\ \text{EXTRINSIC} = [3]\left[\text{UP} = align(Y, \mathcal{E}_{\perp Y})\right]\end{array}\right]\\ \text{AFFORDANCE\_STRUCTURE} = \left[\begin{array}{l}\text{A}_1 = H_{[2]} \rightarrow [put(x, on([1]))]support([1], x)\\ \text{A}_2 = H_{[2]} \rightarrow [put(x, in([1]))]contain([1], x)\\ \text{A}_3 = H_{[2]} \rightarrow [grasp(x, [1])]hold(x, [1])\\ \text{A}_4 = H_{[3]} \rightarrow [roll(x, [1])]\mathcal{R}\end{array}\right]\\ \text{EMBOD} = \left[\begin{array}{l}\text{SCALE} = <\textbf{agent}\\ \text{MOVABLE} = \textbf{true}\end{array}\right]\end{array}\right]$$

In VoxML encodings like 2, bracketed numbers, e.g., [1] are reentrancy indices, such that terms annotated with the same number refer to the same entity. For instance, in habitat 2 ($H_{[2]}$), the intrinsic habitat where the cup has an upward orientation, if an agent puts some x inside the cup's cylindroid geometry ([1]), the cup contains x.

One of the major improvements to the notion of habitat developed in VoxML over that given originally in Pustejovsky (2013) is how the preconditions to actions are encoded and scoped. Notice how in the example in (1), the constraint on relative size of the chair to its user (along with all constraints) is specified outside the modal context in the TELIC, while the VoxML representation using Habitats in (3) provides a reentrant binding for the situational variables.

(3) Habitat and affordance structure for *chair*:

$$\left[\begin{array}{l}\textbf{chair}\\ \text{HABITAT} = \left[\begin{array}{l}\text{INTR} = [2]\left[\begin{array}{l}\text{CONSTR} = \{Y > X, Y > Z\}\\ \text{UP} = align(Y, \mathcal{E}_Y)\\ \text{TOP} = top(+Y)\end{array}\right]\end{array}\right]\\ \text{AFFORD\_STR} = \left[\text{A}_1 = H_{[2]} \rightarrow [sit(y, on([1]))]support([1], y)\right]\end{array}\right]$$

VoxML treats actions and events within a dynamic event semantics as programs (Pustejovsky and Moszkowicz, 2011; Mani and Pustejovsky, 2012). The advantage of adopting a dynamic interpretation of events is that one can map linguistic expressions directly into simulations through an operational semantics (Miller and Johnson-Laird, 1976). Models of processes using updating typically make reference to the notion of

a state transition (Harel, 1984). Each event, such as *put* in (4), can be seen as a traced structure over a Labeled Transition System. The approach is similar in many respects to that developed in both Fernando (2009) and Naumann (2001).

This also allows the system to reason about objects and actions independently. When simulating the objects alone, the simulation presents how the objects change in the world. By removing the objects and presenting only the actions that the viewer would interpret as *causing* the intended object motion (i.e., an embodied agent pantomiming the object motion), the system presents a "decoupled" interpretation of the action, for example, as an animated gesture that traces the intended path of motion. By composing the two, it demonstrates a particular instantiation of the complete event. This allows an embodied situated simulation approach to easily compose objects with actions by directly interpreting at runtime how the two interact.

For the simulation to run, all parameters (e.g., object location, agent motion, etc.) must have values assigned. The simulation environment itself facilitates the calculation of these values, including a common path that the object and agent's manipulator must follow while completing an action; adhering to these common paths and positional values keeps the two synchronized.

(4) Events as Programs:

$$\begin{bmatrix} \textbf{put} \\ \text{LEX} = \begin{bmatrix} \text{PRED} = \textbf{put} \\ \text{TYPE} = \textbf{transition\_event} \end{bmatrix} \\ \text{TYPE} = \begin{bmatrix} \text{HEAD} = \textbf{transition} \\ \text{ARGS} = \begin{bmatrix} \text{A}_1 = \textbf{x:agent} \\ \text{A}_2 = \textbf{y:physobj} \\ \text{A}_3 = \textbf{z:location} \end{bmatrix} \\ \text{BODY} = \begin{bmatrix} \text{E}_1 = grasp(x,y) \\ \text{E}_2 = [while(hold(x,y), move(x,y)] \\ \text{E}_3 = [at(y,z) \rightarrow ungrasp(x,y)] \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

The logic of event structure encodes only minimal temporal constraints on how the subevents interact or play out. The rendering engine itself maintains an internal clock and regulates frame rate, and therefore the time it takes to conduct movements, obviating the need to regularly model this temporal aspect in operationally defined events in VoxML, although scalara attributives like *faster* or *slower* can provide temporal modifiers.

4. VoxWorld: A Platform for Multimodal Simulations

In this section, we introduce a simulation framework, VoxWorld, that situates an embodied agent in a multimodal simulation, with the capability of understanding and generating language and gesture, and the ability to synthetically perceive an interlocutor human as well as objects in its virtual surroundings, and act on them through a limited inventory of actions.

4.1. *Modes of Simulation*

The concept of simulation has played an important role in both AI and cognitive science for over forty years. The two most common uses for the term *simulation* as used in computer science and AI include: (a) *computational simulation modeling*, where variables in a model are set, the model is run, and the consequences of all possible computable configurations become known; and (b) *situated embodied simulations*, where an environment allows a user to interact with objects in a "virtual or simulated world", where the agent is embodied as a dynamic point-of-view or avatar in a proxy situation. Such simulations are used for training humans in scripted scenarios, such as flight simulators, battle training, and of course, in video gaming, where the goal is to simulate an agent within a situation.

Simulation has yet another meaning, where starting with Craik (1943), we encounter the notion that agents carry a mental model of external reality in their heads. Johnson-Laird (1987) develops his own theory of a mental model, which represents a situational possibility, capturing what is common to all the different ways in which the situation may occur. This is used to drive inference and reasoning, both factual and counterfactual. Simulation Theory, as developed in philosophy of mind, has focused on the role "mind reading" plays in modeling the mental representations of other agents and the content of their communicative acts (Goldman, 2006). Simulation semantics (Feldman, 2010; Narayanan, 2010) argues that language comprehension is accomplished by means of such mind reading operations. Similarly, within psychology, there is an established body of work arguing for "mental simulations" of future or possible outcomes, as well as interpretations of perceptual input (Barsalou, 1999). These approaches we refer to as *embodied theories of mind*.

4.2. *VoxWorld*

VoxWorld integrates the functionality and the goals of all three approaches above. The platform situates an embodied agent in a multimodal simulation, with *mind-reading* interpretive capabilities, facilitated through assignment and evaluation of object and context parameters within the environment being modeled.

4.2.1. *Architecture*

VoxWorld is based on the semantic scaffold provided by the VoxML modeling language (Pustejovsky and Krishnaswamy, 2016), which provides a dynamic, interpretable model of objects, events, and their properties. This allows us to create visualized simulations of events and scenarios that are rendered analogues to the "mental simulations" discussed above. We can restrict mind-reading to events that are tangible and perceptually reflective or transparent. So, mental events (desires, beliefs by themselves, etc.) will not be modeled here as simulations themselves, but rather as modal signatures or propositional content of a common ground—that is an agent's desire for food may manifest as holding their stomach or opening the refrigerator, themselves modeled as distinct events stemming from that cause. VoxSim (Krishnaswamy and Pustejovsky, 2016) serves as the event simulator within which these simulations are created and rendered in real time, serving as the computer's method of visually presenting its interpretation of a situation or event. Because modalities are modes of

presentation, a multimodal simulation entails as many presentational modes as there are modalities being modeled. The visual modality of presentation (as in embodied gaming) necessitates "situatedness" of the agent, as do the other perceptual modalities. Therefore, when we speak of *multimodal simulations*, they are inherently situated. In a human-computer interaction using such a simulation, the simulation is a demonstration of the computational agent's "mind-reading" capabilities (an *agent simulation*). If the two are the same (where the agent is a proxy for the player or user), then the "mind-reading" is just a demonstration of the scenario. It observes what the user can. The user observes the agent act as if they share the same perspective. If, on the other hand, the two are separate (agent is *not* proxy for the user), then the simulation/demonstration communicates the agent's understanding of the user and the interaction. In this case, this demonstration entails the illustration of both epistemic and perceptual content of the agent. The agent's actions within the scene facilitate the human's "mind-reading" based on the agent's demonstrated interpretation of propositional content within the scene. We assume an agent has present epistemic knowledge and the relevant inferences reasonably associated with/derivable from these propositions. The agent may know that an object is graspable and can be held in a certain way. This also means that the agent "knows" that it is touchable and moveable, similarly for propositional knowledge associated with logical entailments, etc.

The current architecture of the VoxWorld system is shown in Fig. 5.

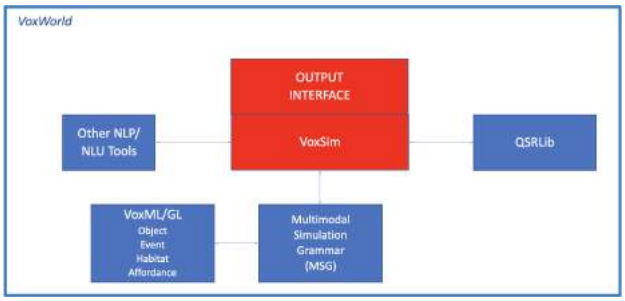


Figure 5. *VoxWorld Architecture schematic.*

At the center is VoxSim, the software that handles visual event simulation in three dimensions, written with the Unity game engine. VoxSim connects to a number of other default VoxWorld components, including some native natural language processing capabilities, VoxML encodings/GL knowledge as interpreted through the multimodal semantics discussed in Section 5, and 3rd-party libraries, e.g., QSRLib (Gatsoulis *et al.*, 2016). Individual agent, such as the interactive avatar Diana (discussed below), are arbitary output interfaces that can also connect to 3rd-party endpoints; in the case of Diana, this is custom gesture and affect recognition (Narayana *et al.*, 2018).

4.2.2. *Usage*

VoxSim contains scenes in a Blocks World domain, plus a set of more complicated or interesting everyday objects (e.g., cups, plates, books, etc.). In scenes without an avatar, the user can direct the computer to manipulate objects in space or create an avatar that can act upon objects and respond to the user's input. VoxWorld includes

other software, models, and interfaces, e.g., to consume input from CNN-based gesture recognizers (Narayana *et al.*, 2018), or to track the agent's epistemic state or knowledge of what its interlocutor knows.

It is straightforward to create new scenes with 3D geometries with packaged code that creates and instantiates voxemes, handles their interactions and performs basic spatial reasoning over them. VoxWorld contains a library of basic motion predicates and methods to compose them into more complex actions using VoxML.

4.2.3. *Situated Reasoning in VoxWorld*

Situational embodiment takes place in real time, so in a situation where there may be too many variables to predict the state of the world at time t from initial conditions at time 0, situational embodiment within the simulation allows the agent to reason forward about a specific subset of consequences of actions taken at time t, given the agent's current conditions and surroundings. Situatedness and embodiment is required to arrive at a complete, tractable interpretation given any element of non-determinism. E.g., an agent trying to navigate a maze from start to finish could easily do so with a map that provides complete or sufficient information about the scenario. However, if the scene is disrupted (e.g., the floor crumbles, or doors open and shut randomly), the agent would be unable to plot a course to the goal. It would have to start moving, assess circumstances at every timestep, and choose the next move(s) based on them. Situated embodiment allows the agent to assess the next move based on the current set of relations between itself and the environment (e.g., ability to move forward but not leftward at the current state). This allows reasoning that saves computational resources and performs more analogously to human reasoning.

Given the continuous tracking of object parameters such as position and orientation, facilitated by a game engine or simulation, and the knowledge of object, event, and functional semantics facilitated by a formal model, an entity's interpretation at runtime can be computed in conjunction with the other entities it is currently interacting with and their properties. One such canonical example would be placing an object [[SPOON]] in an [[IN]] relation with another object [[MUG]] (Fig. 6).

Figure 6. *[[*SPOON IN MUG*]].*

The mug has an intrinsic top, which is aligned with the upward Y-axis of the world or embedding space (denoted in VoxML as $\{align(Y, \mathcal{E}_Y)$, $top(+Y)\}$). The mug is a concave object, and the mug's geometry (the [[CUP]], excluding the handle) has reflectional symmetry across its inherent (object-relative) XY- and YZ-planes, and rotational symmetry around its inherent Y-axis such that when the object is situated in its inherent *top* habitat, its Y-axis is parallel to the world's. From this we can infer that the *opening* (e.g., access to the concavity) must be along the Y-axis. Encoding the object's concavity allows fast computation for physics and collisions using bounding boxes, while still facilitating reasoning over concave objects.

An embodied simulation model such as VoxWorld is an approach that integrates all three aspects of simulation: a situated embodied environment built on a game engine platform. The computer, either as an embodied agent distinct from the viewer,

or as the totality of the rendered environment itself, presents an interpretation (*mind-reading*) of its internal model, down to specific parameter values, which are often assigned for the purposes of testing that model. As such, it provides a rich environment within which to experiment with task-oriented dialogues, such as those explored in Section 6, because of the requirement that the agent have a situated embodiment in which it interprets its environment and its interlocutor. This in turn requires the creation of common ground (CG) between the human and the AI that allows them to communicate. The parameters within this CG structure can be varied and set according to various experimental configurations, allowing us to both qualitatively and quantitatively measure the effect of different CG structures on the communication. For example, we can experiment with variable settings for the composition of multimodal referring descriptions as well as action or event predicates; that is, what aspects of the content of the expression are conveyed through each modality, speech or gesture? Another variation involves the degree of alignment of information in each modal channel; that is, whether a linguistic expression and gesture are synchronous or asynchronous when generated. The interaction in Fig. 7 illustrates a person directing an avatar to pick up a block, using an asynchronous multimodal expression.

Figure 7. *Asynchronous ensemble dialogue: Human grasping gesture precedes his linguistic utterance, "Grab it".*

We assume that a simulation is a contextualized 3D virtual realization of both the situational environment and the co-situated agents, as well as the most salient content denoted by communicative acts in discourse between them. The encoding that VoxML provides for objects, with its rich semantic typing and action affordances, enables VoxWorld to describe agent actions as multimodal programs, as well as identifying and tracking the elements of the common ground that are revealed in the interaction between parties, be they humans or artificially intelligent agents.

5. Multimodal Semantics for Common Ground

The theory of common ground has a rich and diverse literature concerning what is shared or presupposed in human communication (Clark and Brennan, 1991; Stalnaker, 2002; Asher, 1998; Ginzburg and Fernández, 2010). With the presence of a common ground during shared experiences, embodied communication assumes agents can understand one another in a shared context, through the use of co-situational and co-perceptual anchors, and a means for identifying such anchors, such as gesture, gaze, intonation, and language. In this section, we develop a computational model of common ground for multimodal communication.

We assume generally a model of discourse semantics as proposed in Asher and Lascarides (2003), as it facilitates the adoption of a continuation-based semantics for our phrase-level compositional semantics (Barker and Shan, 2014), as well for discourse, as outlined in De Groote (2001) and Asher and Pogodalla (2010). For the present discussion, however, we will not refer to SDRT representations, but focus

instead on the semantics integrated multimodal expressions in the context of task oriented dialogue, as presented first in (Pustejovsky, 2018) and extended here.

Here, we introduce the notion of a *common ground structure*, the information associated with a state in a dialogue or discourse. We model this as a state monad (Unger, 2011), as illustrated in (5).

(5) State Monad: $\mathbf{M}\alpha = State \rightarrow (\alpha \times State)$

A state monad corresponds to computations that read and modify a particular state, in this case a state in the discourse. **M** is a type constructor that constructs a function type taking a state as input and returns a pair of a value and a new or modified state as output. This monad consists of the following state information:

(6) a. the communicative act, C_a, performed by an agent, a: a tuple of expressions from the modalities involved. For our present discussion, we restrict this to a linguistic utterance, S (speech) and a gesture, G. There are hence three possible configurations in performing a C: $C_a = \{(G), (S), (S, G)\}$;
b. **A**: the agents engaged in communication;
c. **B**: the shared belief space;
d. **P**: the objects and relations that are jointly perceived in the environment;
e. $\mathcal{E}$: the embedding space that both agents occupy in the communication.

The common ground structure (CGS) can be represented graphically as in (7), where an agent, a_i, makes a communicative act either through gesture, $\mathcal{G}$ in (7a), or linguistically, as in (7b.) [1]

(7) a. $\dfrac{\textbf{A:}a_1, a_2 \;\; \textbf{B:}\Delta \;\; \textbf{P:}b \;\; \mathcal{E} : E}{\mathcal{G}_{a_1}}$ b. $\begin{array}{|l|}\hline \textbf{A:}a_1, a_2 \;\; \textbf{B:}\Delta \;\; \textbf{P:}b \;\; \mathcal{E} : E \\ \mathcal{S}_{a_1} = \text{“You}_{a_2} \text{ see it}_b\text{”} \\ \hline \end{array}$

(7a) specifies that two agents, a_1 and a_2, co-inhabiting an embedding space, E, within which the experience is embodied, share a set of beliefs, Δ, where they can both see the object, b. Given this representation, the gesture is now situated to refer to objects and knowledge within the CG structure. In (7b), the linguistic expression, $\mathcal{S}_{a_1}$, is grounded relative to the parameters of common ground, where the indexical *you* will denote the agent, a_2, and the pronoun *it* will denote the object, b.

We have augmented and extended the approach taken in Kendon (2004) and Lascarides and Stone (2009), where gestures are simple schemas consisting of distinct sub-gestural phases, where **Stroke** is the content-bearing phase of the gesture.

(8) $G \rightarrow (Prep)\ (Pre\_stroke\ Hold)\ \mathbf{Stroke}\ Retract$

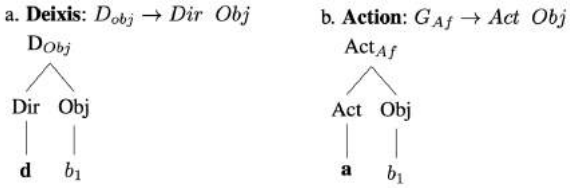
[1] This is similar in many respects to the representations introduced in Cooper and Ginzburg (2015), Ginzburg and Fernández (2010) and Dobnik *et al.* (2013) for modeling action and control with robots.

In the context of multimodal dialogues and interactions with computational agents and robots, gesture's **Stroke** will denote a range of primitive action types, $\mathcal{ACT}$, e.g., *grasp*, *pick up*, *move*, *throw*, *pull*, *push*, *separate*, and *put together*. There are many ways to convey intent to carry out these actions, but they all involve two characteristics: (a) the action's object is an embodied reference in the common ground; and (b) the gesture sequence must be interpreted dynamically, to correctly compute the end state of the event. To this end, we model two kinds of gestures in our dialogues: (a) establishing a reference; and (b) depicting an action-object pair.

(9) a. **Deixis**: $D_{obj} \rightarrow Dir \;\; Obj$
b. **Action**: $G_{Af} \rightarrow Act \;\; Obj$

We introduce the notion of an interpreted gesture tree in (10a), which indicates that the gesture D_{obj} functionally consists of a deictic orientation, Dir, with the demonstratum, **d**, and the referenced or denoting entity, Obj, denoting b_1.

(10) Interpreted Gesture Tree:



As gesture is intended for visual interpretation, it is directly interpretable by the interlocutor in the context if and only if the value is clearly evident in the common ground, most likely through visual inspection. Directional or orientational information conveyed in a gesture identifies a distinct object or area of the embedding space, E, by directing attention to the End of the designated pointing ray (or cone) trace (Lascarides and Stone, 2009; Lücking *et al.*, 2015; Pustejovsky, 2018).

(11) $[\![\mathbf{D}_{obj}]\!] = [\![End(ray(\mathbf{d}))]\!]$

We model the interpretation function, $[\![.]\!]$, as fully determining the value of the deixis in the context, supplied by the common ground, which we discuss below. In (10b), the action gesture type, G_{Af}, consists of an action-object pairing, where the action, **a**, is applied to the object, b_1, in some prototypical manner. The strategies available are outlined in (12-14).

(12) a. ACTION-OBJECT: e.g., *grab* [**Object**]
b. $GvP_1 \rightarrow G_{Af} \;\; D_{obj}$ (Action Focus)
$\rightarrow \;\; D_{obj} \;\; G_{Af}$ (Object Focus)

(13) a. ACTION-RESULT: e.g., *put* [**Object**] at [**Location**]
b. $GvP_2 \rightarrow G_{Af} \;\; D_{obj} \;\; D_{loc}$ (Action Focus)
$\rightarrow \;\; D_{obj} \;\; G_{Af} \;\; D_{loc}$ (Object Focus)
$\rightarrow \;\; D_{obj} \;\; D_{loc} \;\; G_{Af}$ (Transition Focus)

(14) a. ACTION-RESULT: e.g., *move* [**Object**] [**Direction**]
b. $GvP_3 \rightarrow G_{Af} \;\; D_{obj} \;\; D_{dir}$

As mentioned above, the deictic gesture in (9a) and (10a) actually serves to indicate both a location and objects within that location, suggesting that deixis denotes a *dot object*, viz., PHYSOBJ•LOCATION (Pustejovsky, 1995). Either of these type components may be exploited by the deictic reference, which is then interpreted in context, either as a selection (exploiting the PHYSOBJ) or as a destination (exploiting either). For example, should an object b_1 already be selected through a deixis $\mathbf{d}_a$, as in (10a), a subsequent deixis $\mathbf{d}_b$ may be interpreted as selecting a destination location in isolation (in which case the interpretation exploits the LOCATION of $\mathbf{d}_b$), or as selecting a location relative to another object (exploiting the PHYSOBJ type of $\mathbf{d}_b$). We discuss this further below.

With conventional treatments of continuation-style passing within the utterance, all linguistic expressions are continuized within the sentence. This has a distinct advantage in multimodal processing, because it allows for an *informational distribution* among the expressions being used in composition to form larger meanings.

By treating the common ground as a state monad, as described above, we can continuize the composition above the level of the sentence as well. Following De Groote (2001), Asher and Pogodalla (2010) and further developments in Van Eijck and Unger (2010), we represent a context as a stack of items and the type of left contexts to be lists of entities, $[e]$. Right contexts will be interpreted as continuations: a discourse that requires a left context to yield a truth value. The type of a right context is therefore $[e] \to t$. Hence, context transitions get the type $[e] \to [e] \to t$; they are characteristic functions of binary relations on contexts. The continuized semantics for gesture phrases is in (15).

(15) a. $\mathbf{S}_G \to (\mathbf{NP})\ \mathbf{GvP}$
$[\![S]\!] = ([\![\mathbf{NP}]\!][\![\mathbf{GvP}]\!])$

b. $\mathbf{GvP}_1 \to \mathbf{G}_{af}\ \mathbf{D}_{Obj}$
$[\![\mathbf{GvP}_1]\!] = \lambda j.([\![\mathbf{D}_{Obj}]\!]; \lambda j'.(([\![\mathbf{G}_{af}]\!]j')j))$

c. $\mathbf{GvP}_2 \to \mathbf{G}_{af}\ \mathbf{D}_{Obj}\mathbf{D}_{Loc}$
$[\![\mathbf{GvP}_2]\!] = \lambda k.([\![\mathbf{D}_{Loc}]\!]; \lambda j.([\![\mathbf{D}_{Obj}]\!]; \lambda j'.(([\![\mathbf{G}_{af}]\!]j')j)k))$

d. $\mathbf{GvP}_3 \to \mathbf{G}_{af}\ \mathbf{D}_{Obj}\mathbf{D}_{Dir}$
$[\![\mathbf{GvP}_3]\!] = \lambda k.([\![\mathbf{D}_{Dir}]\!]; \lambda j.([\![\mathbf{D}_{Obj}]\!]; \lambda j'.(([\![\mathbf{G}_{af}]\!]j')j)k))$

The discourse updating operation is accomplished through continuation-passing as well, as in (Asher and Pogodalla, 2010). We apply a CPS transformation to arrive at the continuized type for each expression, notated as an overlined expression (Van Eijck and Unger, 2010). Given the current discourse, T, and the new utterance, C, we take the integration of C into T as follows:

(16) $[\![\overline{(\mathbf{T}.\mathbf{C})}]\!]^{M,cg} = \lambda k.[\![\overline{\mathbf{T}}]\!](\lambda n.[\![\overline{\mathbf{C}}]\!](\lambda m.k(m\ n)))$

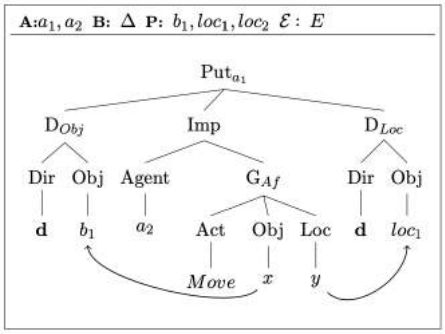
To illustrate how continuations help in the interpretation of gesture sequences, consider a single modality gesture imperative.

SINGLE MODALITY (GESTURE) IMPERATIVE

HUMAN$_1$: $\mathcal{G}$ = [*points to the purple block*]$_{t1}$
HUMAN$_2$: $\mathcal{G}$ = [*makes move gesture*]$_{t2}$
HUMAN$_3$: $\mathcal{G}$ = [*points to the red block*]$_{t3}$

Through its own continuation, the referent identified in the first deixis, $\mathbf{D}_{Obj}$, is passed to the action ($\lambda k.k([\![\mathbf{Move}]\!])$), while the continuized interpretation of the action delays the computation of its argument until the appropriate binding has been identified. Finally, the goal location for the movement selected for by the *move* gesture is identified through the action of the continuized location deixis, $\mathbf{D}_{Loc}$. This is illustrated in (18), along with the common ground structure that is computed, shown in (17).

(17)



(18) $[\![\mathbf{D}_{Obj}.\mathbf{Move}.\mathbf{D}_{Loc}]\!] = \lambda k.([\![\mathbf{D}_{Loc}]\!]; \lambda j.([\![\mathbf{D}_{Obj}]\!]; \lambda j'.(([\![\mathbf{Move}]\!]j')j)k))$

Given a description of the gesture grammar as used in our multimodal dialogues, let us explore a communicative act that exploits a combination of both speech and gesture, (S, G). We identify three configurations for how a language-gesture *ensemble* can be interpreted, depending on which modality carries the majority of semantic content: (a) language with *co-speech gesture*, where language conveys the bulk of the propositional content and gesture adds situated grounding, affect, effect, and presuppositional force (Cassell *et al.*, 2000; Lascarides and Stone, 2009; Schlenker, 2020); (b) *co-gestural speech*, where gesture plays this role (Pustejovsky, 2018); and (c) a truly mixed modal expression, where both language and gesture contribute equally to the meaning. In practice, while many of the interaction in our dialogues have this property, the discourse narrative is broadly guided by gesture. For this reason, we model the multimodal interactions as content-bearing gesture with *co-gestural speech*.

A multimodal communicative act, C, consists of a sequence of gesture-language ensembles, (g_i, s_i), where an ensemble is temporally aligned in the common ground. Let us assume that a linguistic subexpression, s, is either a word or full phrase in the utterance, while a gesture, g, comports with the gesture grammar described above.

(19) **Co-gestural Speech Ensemble**:

$$\begin{bmatrix} \mathcal{G} & g_1 & \dots & g_i & \dots & g_n \\ \mathcal{S} & s_1 & \dots & s_i & \dots & s_n \end{bmatrix}$$

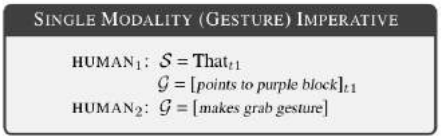
We assume an aligned language-gesture syntactic structure, for which we provide a continuized semantic interpretation. Both

of these are contained in the common ground state monad introduced above (6). For each temporally indexed and aligned gesture-speech pair, (g, s), we have a continuized interpretation, as shown below. Each modal expresssion carries a continuation, k_g or k_s, and we denote alignment of these two continuations as $k_s \otimes k_g$, seen (20).

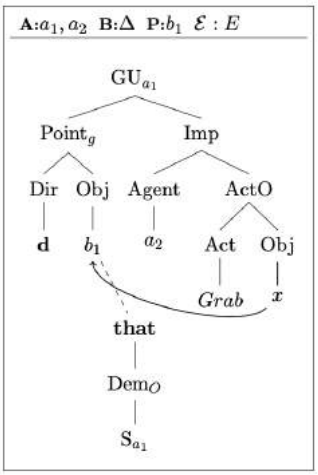
(20) $\lambda k_s.k_s([\![\mathbf{s}]\!])$
$\lambda k_g.k_g([\![\mathbf{g}]\!])$
$\lambda k_s \otimes k_g.k_s \otimes k_g([\![\mathbf{(s,g)}]\!])$

We bind co-gestural speech to specific gestures in the communicative act, within a common ground, CGS. A dashed line in (21) indicates that a co-gestural speech element, $\mathcal{S}$, is aligned with a particular gesture, $\mathcal{G}$. For example, consider the co-gestural speech expression below.

The CG structure for this expression, $\begin{bmatrix} \mathcal{G} & D_{Obj} & Grab_g \\ \mathcal{S} & \text{THAT} & \_\_\_ \end{bmatrix}$, is shown in (21).



(21) $[\![\langle \text{THAT}, \mathbf{D}_{Obj}\rangle.\langle \_, \mathbf{Grab}\rangle]\!] = \lambda k_s \otimes k_g.([\![\mathbf{D}_{Obj}]\!]; \lambda j_g.((([\![\mathbf{Grab}]\!]j_g)k_s \otimes k_g))$



Common ground updates will also include executing modal operations over the belief space **B**, where each new element from the discourse is introduced via a *public announcement logic* (PAL) formula, and each new perceived object or relation is introduced into **P** via an analogous *public perception logic* (PPL) formula (Plaza, 2007; Van Ditmarsch *et al.*, 2007; Van Benthem, 2011). We will use $[\alpha]\varphi$ to denote that an agent "α *knows* φ". Public announcements are implemented as: $[!\phi_1]\phi_2$. Any proposition, φ, in the common knowledge held by two agents, α and β, is computed as: $[(\alpha \cup \beta)^*]\varphi$.

Similarly, an agent α's perception is encoded as sets of accessibility relations, α, between situations. What is seen in a situation is encoded as either a proposition, φ, or existential statement of an object, x, $\hat{x}$. $[\alpha]_\sigma\varphi$ denotes that agent "α perceives that φ". $[\alpha]_\sigma\hat{x}$ denotes that agent "α perceives that there is an x."

(22) a. **block**: Pick me up!, Move me!
b. **cup**: Pick me up!, Drink what's in me!
c. **knife**: Pick me up!, Cut that with me!

Given the theory of two-level affordances proposed here (Gibsonian/Telic), we can naturally think of objects as *antecedents to the actions performable on them*. For each object in (22), we identify attached behaviors. This naturally suggests that affordances are a subclass of continuations. For example, both *cup* and *block* have similar Gibsonian affordance values, but quite distinct Telic affordance values. This can be distinguished by the nature of their respective Telic continuation sets as follows, where **sel** is a function that selects a suitable discourse antecedent inside the continuation set (Asher and Pogodalla, 2010): $\lambda k_{Gib} \otimes k_{Telic}.k_{Gib} \otimes k_{Telic}(cup)$, $grab \subseteq$ **sel** k_{Gib}, $drink \subseteq$ **sel** k_{Telic}, $\lambda k_{Gib} \otimes k_{Telic}.k_{Gib} \otimes k_{Telic}(block)$, $grab \subseteq$ **sel** k_{Gib}, $pick\_up \subseteq$ **sel** k_{Gib}, $move \subseteq$ **sel** k_{Gib}. This is the subject of ongoing research.

6. Experiments with Multimodal Dialogues

6.1. *Aspects of Multimodal Compositionality*

In this section, we provide additional formal analysis of experimental data gathered from multimodal dialogues between a human and a computational agent, represented as an avatar in VoxWorld. We examine extracts from dialogues between humans and computational agents in various tasks, in order to examine the nature of the communicative act in the context of the common ground structure. We illustrate how the situated meaning of the multimodal expression is constructed in each case. In particular, we look at three aspects of multimodal compositionality in these examples:

(23) a. generating referring expressions using different modalities;
b. generating and interpreting action and event expressions;
c. generating full action descriptions using both gesture and language.

Recall that a multimodal communicative act, C, consists of a sequence of gesture-language ensembles, (g_i, s_i), where an ensemble is temporally aligned in the common ground. For the examples below, we annotate the dialogue with the contribution of both speech and gesture for each agent. Each dialogue turn encodes a multimodal ensemble, $\begin{bmatrix} \mathcal{S} \\ \mathcal{G} \end{bmatrix}$, which may or may not be realized in both modalities. In the annotation below, alignment between the modalities is indicated through a temporal indexing on the appropriate modal expression, e.g., t_i.

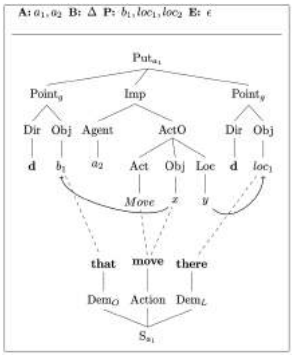
Figure 8. *Co-gestural speech imperative.*

Since we can use speech and gesture to indicate objects, location, and actions, we bias our speech recognition toward syntactic categories that represent partial information (e.g., NPs for objects, PPs for locations, VPs for actions), using incrememntal predictivity (cf. Hough *et al.* (2015)). We parse input in both directions, so we can take inputs like "put a block on the purple block" without resolving "a block" to the purple block, to prevent the agent from putting the purple block on itself.

(24)

MIXED MODALITY IMPERATIVE

HUMAN$_1$: $\mathcal{S}$ = That$_{t1}$
$\mathcal{G}$ = [*points to the purple block*]$_{t1}$
HUMAN$_2$: $\mathcal{S}$ = Move$_{t2}$
$\mathcal{G}$ = [*makes move gesture*]$_{t2}$
HUMAN$_3$: $\mathcal{S}$ = There$_{t3}$
$\mathcal{G}$ = [*points to the red block*]$_{t3}$



$\lambda r_s \otimes r_g.([\![\langle \text{THERE}, \mathbf{D}_{Loc} \rangle]\!] \lambda k_s \otimes k_g.([\![\langle \text{THAT}, \mathbf{D}_{Obj} \rangle]\!]; \lambda j_s \otimes j_g.((([\![\langle \text{MOVE}, \mathbf{Move} \rangle]\!] j_s \otimes j_g) k_s \otimes k_g)) r_s \otimes r_g)$

6.2. *Multimodal Referring Expressions*

The *E*mbodied *M*ultimodal *R*eferring *E*xpressions (EMRE) dataset (Krishnaswamy and Pustejovsky, 2019) consists of 1,500 visual simulated situations showing an agent (Diana) indicating various object in a scene each accompanied by a definite referring expression. Referring expressions may take the form of deictic gesture only, a spoken description only with no demonstratives (e.g., "the red block in front of the knife and left of the green block"), or a mixed-modality referring expression as in Fig. 9 (right). Fig. 9 (left) shows a sample still that accompanies the utterance, with an equivalent common ground structure one the right.

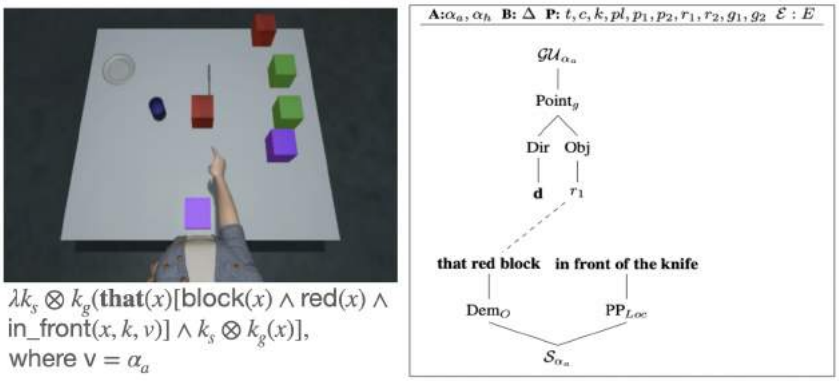


Figure 9. *Left: Sample still from the EMRE dataset (L), with CGS (R) and semantics of the RE (below), showing a continuation for each modality, k_s and k_g, which apply over the object subsequently in the dialogue.*

Amazon Mechanical Turk workers evaluted the EMRE dataset on a Likert-type scale for naturalness of the depicted referring expression for the indicated object. We found a clear preference for the multimodal referring expressions, suggesting that the redundancy provided by co-occurring language and gesture made for the clearest, most natural references to objects.

In Krishnaswamy and Pustejovsky (2020), we extracted formal features from the data as one-hot vectors representing elements of common ground structures. If in one of visualized REs, the avatar points to b, one of the jointly perceived objects $\in \mathbf{P}$, such that $\forall b\ (b \in \mathbf{P} \rightarrow \mathcal{K}_{\alpha_h}\mathcal{P}_{\alpha_a} b \wedge \mathcal{K}_{\alpha_a}\mathcal{P}_{\alpha_h})$. This demonstrates the avatar can point, and knows that b is the target $[C_{\alpha_a} = Point_g \rightarrow Dir\ b!]\mathcal{K}_{\alpha_h}\mathcal{K}_{\alpha_a}(Point_g \wedge target(b))$, which is encoded as a single feature. An agent may introduce a new object into the dicussion, making common the knowledge of its existence. Or an agent a uses a term t to make public the knowledge of a's interpretation of t.

We used these CGS-extracted features to train a neural net to predict the naturalness of a given referring expression, using the naturalness judgments from the EMRE dataset as ground truth. The EMRE dataset contains situational information about the specific configuration in which the referring expression was generated, and the linguistic referring expression itself, so we tested the effects of including formal, CGS-derived features by training classifiers on combinations of the symbolic situational features, embedding vectors of the linguistic RE, and the CGS-derived features.

We trained a multilayer preception, a simple, fast architecture that can distinguish dependencies in linearly-inseparable regions of data. This architecture consists of three fully-connected hidden layers of 32, 128, and 64, respectively, prior to a $softmax$ output layer. The layers use $tanh$, ELU, and $tanh$ activation, respectively, cross-entropy loss and Adam optimization, and is trained for 1,000 epochs with a batch size of 50. We perform 7-fold cross-validation in order to achieve a more balanced sample across all classes of annotator judgments. $k = 7$ is chosen here to approximate a leave-one-out cross-validation approach over the 8 annotator judgments on each visualized referring expression. The "most likely" annotator judgment in the EMRE dataset is a probability distribution so, we regard a "correct" prediction by the classifier as one that falls within the correct quintile of the distribution over all annotator judgments of that visualized referring expression.

	Raw features	**Raw feat. + SE**
μ **Acc. (1K)**	0.6757	0.6429
σ **Acc. (1K)**	0.0230	0.0111

	Raw + form.	**Raw + form. + SE**	**Formal only**
μ **Acc. (1K)**	0.7214	0.6671	0.7471
σ **Acc. (1K)**	0.0398	0.0243	0.0269

Figure 10. *Classification accuracy using formal features (mean and standard deviation).*

Fig. 10 shows that inclusion of formal features derived from the elements of common-ground structures improved classifier prediction accuracy by between 7% and 11% relative to baseline predictions that used the raw features of the EMRE dataset, plus sentence embedding representations of the referring expression itself. This suggests that common ground structures provide a dense, interpretable representation of the dialogue state, facilitating generation of natural, situation-appropriate referring expressions, and predicts the natural quality of a referring expression beyond other strong predictors of naturalness, e.g., modality.

6.3. *Interruptions and Corrections in Dialogue*

Figure 11. *Correcting and undoing an action.*

Establishing entities in a common ground structure so they can be recombined appropriately and interpreted in context allows us to build asynchronous agent behaviors capable of interruption and correction. Correction (Fig. 12) is currently implemented by performing three functions: (a) **Undo**, which recontinuizes an expression which has saturated its parameters, i.e., **undo** $k = \lambda k.k(grab)$; (b) **Rewind**, which reintroduces the previous monad; and (c) **Reassign**, which takes the corrected value and assigns it, resulting in $M, cg_2 \models grab(white)$.

In this manner, parameters can be unbound from either object or location argument, depending on the typing of the content communicated. Fig. 11 shows one such situation, where the replacement content "*on* the white one" is evaluated to a location. The state monad containing the location on the blue block is rewound, and the argument reassigned to the location on top of the white block. Had the utterance been "the

REFERENCE REPAIR

H: $\mathcal{G}$ = [*points to area around yellow and white blocks*]
D: $\mathcal{S}$ = Okay.$_{t1}$
$\mathcal{G}$ = [*picks up yellow block*]$_{t1}$
H: $\mathcal{S}$ = No, the white one.
D: $\mathcal{S}$ = Okay.$_{t2}$
$\mathcal{G}$ = [*picks up white block*]$_{t2}$

The user ambiguously points to yellow and white blocks. Diana chooses the yellow block ($\lambda k.k(grab) \Rightarrow M, cg_1 \models grab(yellow)$). The user corrects her, focus is unbound from the yellow block and assigned to the white block.

Figure 12. *Correcting deictic reference*

white one," the action would be reassigned with the white block as the theme, with the previously-existing target location, and Diana would put down the yellow block and put the *white* block on the blue block.

6.4. *Affordance Structure and Transfer Learning*

Diana may come across objects with different affordances from the typical Blocks World scenario. In these cases, the semantics of each object provided by VoxML allows Diana to learn new gestures associated with specific affordances of specific objects. Fig. 13 specifies such an interaction.

AFFORDANCE LEARNING IN KITCHENWORLD

HUMAN: $\mathcal{S}$ = The plate.
DIANA: $\mathcal{S}$ = Okay.$_{t1}$
$\mathcal{G}$ = [*points to the plate*]$_{t1}$
HUMAN: $\mathcal{G}$ = [*makes "claw down" gesture*]
DIANA: $\mathcal{S}$ = Should I grasp it like this$_{t2}$?
$\mathcal{G}$ = [*grasps plate from the side*]$_{t2}$
HUMAN: $\mathcal{S}$ = Yes.
DIANA: $\mathcal{S}$ = Is there a gesture for that?
HUMAN: $\mathcal{G}$ = [*makes "grasp plate" gesture*]

Figure 13. *Diana and human interacting.*

Using a random forest classifier, the gesture the human makes to associate with the specific affordance is situated in the search space defining the existing known gestures. Those learned grasp semantics can then be propagated down to any other event containing [[GRASP]] as a subevent, as shown in (25).

while(C, A) states that an activity, A, is performed only if a constraint, C, is satisfied at the same moment. Thus, if the agent encounters a [[SLIDE]] action with an outstanding variable ($\lambda y.slide(y, loc)$), and the human makes a gesture denoting $grasp(plate)$, the agent can directly lift $grasp(plate)$ to the slide action and apply the argument $plate$ to y: $\lambda y.slide(y, loc)@plate \Rightarrow \lambda y.slide(y, loc)$.

(25) $grasp(e_1, \text{AG}, y)$; **while**$(hold(\text{AG}, y) \wedge on(y, \text{SURF}) \wedge \neg at(y, \text{LOC}))$, $move\_to(e_2, \text{AG}, y, \text{LOC}))$; **if**$(at(y, \text{LOC}), ungrasp(e_3, \text{AG}, y))$

Model	% correct cluster
MLP (Habitats)	78.82
MLP (Affordances)	**84.71**
CNN (Habitats)	78.82
CNN (Affordances)	81.18

Figure 14. *Prediction accuracy w/ 6 means.*

Affordance properties can also be transferred between objects. Given that similar habitats serve as necessary (but not sufficient) preconditions to behaviors (e.g., to be rolled, an apple, cup, and bottle must all be turned on their sides), the ability to assess an unknown object relative to known ones allows an agent to transfer properties between them, to gain a handle on interacting with and discussing a novel object. Consider Fig. 15, where Diana has no semantics for what we recognize as a bottle.

Diana observes similarities in the cup's habitats and the bottle's (e.g., similar orientation, symmetry and size constraints), infers they may share behaviors, and so grasps one like the other. Links between habitats and affordances allows inference of similar objects and behaviors in the current situation.

TRANSFER OF OBJECT AFFORDANCES

HUMAN: $\mathcal{S}$ = What is that?$_{t1}$
$\mathcal{G}$ = [*points to the cup*]$_{t1}$
DIANA: $\mathcal{S}$ = That's a cup$_{t2}$
$\mathcal{G}$ = [*points to the cup*]$_{t2}$
HUMAN: $\mathcal{S}$ = What is that?$_{t3}$
$\mathcal{G}$ = [*points to the bottle*]$_{t3}$
DIANA: $\mathcal{S}$ = I don't know, but I could grasp it like a cup$_{t4}$.
$\mathcal{G}$ = [*points to the bottle*]$_{t4}$
HUMAN: $\mathcal{S}$ = Grab it.
DIANA: $\mathcal{G}$ = [*grasps bottle from the side*]$_{t5}$

Figure 15. *Transferring affordance properties through dialogue.*

Over 17 VoxML objects (e.g., Fig.2), we trained 200D *habitat and affordance embeddings* using a Skip-Gram model for 50,000 epochs with window size 3. Objects were represented as averaged habitat or affordance vectors. These embeddings were run through a 7-layer MLP and a 4-layer (1D) CNN, that chose the known object most similar to the unlabeled vector. E.g., a vector representing a *plate*'s affordances was predicted to be similar to a *cup* or *bottle* due to its containment affordance.

For each object, 8 annotators chose the 2 most similar objects in the vocabulary, in terms of their afforded behaviors, and we performed k-means clustering over these annotations. Our models trained on habitat or affordance embedding vectors successfully predicted an object in the correct cluster 80% of the time (Fig. 14). Diana then enacted known behaviors over novel objects (Fig. 15, top right). Further analysis of these models and their properties are ongoing but these early results show how affordances can be used to train useful models over small sample sizes.

7. Conclusion

Multimodal peer-to-peer interfaces require robust integration of conversational modalities in a naturalistic fashion. We have outlined the first steps toward such integration, based on the logic of our multimodal simulation semantics and 3D environment as the platform for shared common ground. We give our computational agent a framework for major faculties natively available to humans using computer vision techniques to recognize gesture and by laying the groundwork for a modal logic of synthetic vision. The result is a framework and platform that interweaves linguistic and non-linguistic modalities in the completion of a shared task by exploiting the relative strengths of linguistic and non-linguistic context to exchange information in a situated communication. We have also developed this framework into an interaction with a mobile robot mediated by a virtual rendition of the environment the robot sees as it explores. The human then gestures to objects and locations on the screen and gives the robot grounded instructions with spoken English and gesture.

We hope to have demonstrated that the notion of situatedness involves embedding linguistic expressions and their grounding within a multimodal semantics. This approach allows environmentally-aware models that can be validated; if one model of expression (e.g., gesture) is insufficiently communicative, another (e.g., language) can be used to examine where it went wrong. Each additional modality provides an avenue through which to validate models of other modalities.

Acknowledgements

This work was supported by Contract W911NF-15-C-0238 with the US Defense Advanced Research Projects Agency (DARPA) and the Army Research Office (ARO). Approved for Public Release, Distribution Unlimited. The views expressed herein are ours and do not reflect the official policy or position of the Department of Defense or the U.S. Government. We would like to thank Ken Lai, Bruce Draper, Ross Beveridge, Francisco Ortega, and Lucia Donatelli for their comments and suggestions.

8. References

Asher N., "Common ground, corrections and coordination", *Journal of Semantics*, 1998.

Asher N., Lascarides A., *Logics of conversation*, Cambridge University Press, 2003.

Asher N., Pogodalla S., "SDRT and continuation semantics", *JSAI International Symposium on Artificial Intelligence*, Springer, p. 3-15, 2010.

Barker C., Shan C.-c., *Continuations and natural language*, vol. 53, Oxford studies in theoretical linguistics, 2014.

Barsalou L. W., "Perceptions of perceptual symbols", *Behavioral and brain sciences*, vol. 22, n° 4, p. 637-660, 1999.

Cassell J., Stone M., Yan H., "Coordination and context-dependence in the generation of embodied conversation", *Proc. of 1st Int. Conf. on NLG*, ACL, p. 171-178, 2000.

Chai J. Y., Fang R., Liu C., She L., "Collaborative language grounding toward situated human-robot dialogue", *AI Magazine*, vol. 37, n° 4, p. 32-45, 2016.

Clark H. H., Brennan S. E., "Grounding in communication", *Perspectives on socially shared cognition*, vol. 13, n° 1991, p. 127-149, 1991.

Cooper R., Ginzburg J., "Type Theory with Records for Natural Language Semantics", *The handbook of contemporary semantic theory*p. 375, 2015.

Craik K. J. W., *The nature of explanation*, Cambridge University, Cambridge UK, 1943.

De Groote P., "Type raising, continuations, and classical logic", *Proceedings of the 13th Amsterdam Colloquium*, p. 97-101, 2001.

Dobnik S., Cooper R., Larsson S., "Modelling language, action, and perception in type theory with records", *Constraint Solving and Language Processing*, Springer, p. 70-91, 2013.

Feldman J., "Embodied language, best-fit analysis, and formal compositionality", *Physics of life reviews*, vol. 7, n° 4, p. 385-410, 2010.

Fernando T., "Situations in LTL as strings", *Information and Computation*, vol. 207, n° 10, p. 980-999, 2009.

Fischer K., "How people talk with robots: Designing dialog to reduce user uncertainty", *AI Magazine*, vol. 32, n° 4, p. 31-38, 2011.

Gatsoulis Y., Alomari M., Burbridge C., Dondrup C., Duckworth P., Lightbody P., Hanheide M., Hawes N., Hogg D., Cohn A. *et al.*, "QSRlib: a software library for online acquisition of Qualitative Spatial Relations from Video", 2016.

Gibson J. J., "The Theory of Affordances", *Perceiving, Acting, and Knowing: Toward an ecological psychology*p. 67-82, 1977.

Ginzburg J., "Interrogatives: Questions, facts and dialogue", *The handbook of contemporary semantic theory*p. 359-423, 1996.

Ginzburg J., Fernández R., "Computational Models of Dialogue", *The handbook of computational linguistics and natural language processing*, vol. 57, p. 1, 2010.

Goldman A. I., *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*, Oxford University Press, 2006.

Harel D., "Dynamic Logic", *in* M. Gabbay, F. Gunthner (eds), *Handbook of Philosophical Logic, Volume II: Extensions of Classical Logic*, Reidel, p. 497-604, 1984.

Hough J., Kennington C., Schlangen D., Ginzburg J., "Incremental semantics for dialogue processing: Requirements, and a comparison of two approaches", 2015.

Hunter J., Asher N., Lascarides A., "A formal semantics for situated conversation", *Semantics and Pragmatics*, 2018.

Johnson-Laird P., "How could consciousness arise from the computations of the brain", *Mindwaves. Oxford: Basil Blackwell*p. 247-257, 1987.

Kendon A., *Gesture: Visible action as utterance*, Cambridge University Press, 2004.

Kennington C., Kousidis S., Schlangen D., "Interpreting situated dialogue utterances: an update model that uses speech, gaze, and gesture information", *Proceedings of SigDial 2013*, 2013.

Krishnaswamy N., Pustejovsky J., "VoxSim: A Visual Platform for Modeling Motion Language", *Proceedings of COLING 2016*, ACL, 2016.

Krishnaswamy N., Pustejovsky J., "Generating a Novel Dataset of Multimodal Referring Expressions", *Proc. of 13th Int. Conference on Computational Semantics*, p. 44-51, 2019.

Krishnaswamy N., Pustejovsky J., "A Formal Analysis of Multimodal Referring Strategies Under Common Ground", *Proceedings of The 12th LREC*, p. 5919-5927, 2020.

Landragin F., "Visual perception, language and gesture: A model for their understanding in multimodal dialogue systems", *Signal Processing*, vol. 86, n° 12, p. 3578-3595, 2006.

Lascarides A., Stone M., "A formal semantic analysis of gesture", *Journal of Semantics*p. ffp004, 2009.

Lücking A., Pfeiffer T., Rieser H., "Pointing and reference reconsidered", *Journal of Pragmatics*, vol. 77, p. 56-79, 2015.

Mani I., Pustejovsky J., *Interpreting Motion: Grounded Representations for Spatial Language*, Oxford University Press, 2012.

Marge M., Rudnicky A. I., "Towards evaluating recovery strategies for situated grounding problems in human-robot dialogue", *2013 IEEE RO-MAN*, IEEE, p. 340-341, 2013.

Miller G. A., Johnson-Laird P. N., *Language and perception.*, Belknap Press, 1976.

Narayana P., Krishnaswamy N., Wang I., Bangar R., Patil D., Mulay G., Rim K., Beveridge R., Ruiz J., Pustejovsky J., Draper B., "Cooperating with Avatars Through Gesture, Language and Action", *Intelligent Systems Conference (IntelliSys)*, 2018.

Narayanan S., "Mind changes: A simulation semantics account of counterfactuals", *Cognitive Science*, 2010.

Naumann R., "Aspects of changes: a dynamic event semantics", *Journal of semantics*, vol. 18, p. 27-81, 2001.

Plaza J., "Logics of public communications", *Synthese*, vol. 158, n° 2, p. 165-179, 2007.

Pustejovsky J., *The Generative Lexicon*, MIT Press, 1995.

Pustejovsky J., "Dynamic Event Structure and Habitat Theory", *Proc. of the 6th International Conference on Generative Approaches to the Lexicon (GL2013)*, ACL, p. 1-10, 2013.

Pustejovsky J., "From actions to events: Communicating through language and gesture", *Interaction Studies*, vol. 19, n° 1-2, p. 289-317, 2018.

Pustejovsky J., Krishnaswamy N., "VoxML: A Visualization Modeling Language", *Proceedings of LREC*, 2016.

Pustejovsky J., Moszkowicz J., "The qualitative spatial dynamics of motion", *The Journal of Spatial Cognition and Computation*, 2011.

Scheutz M., Cantrell R., Schermerhorn P., "Toward humanlike task-based dialogue processing for human robot interaction", *Ai Magazine*, vol. 32, n° 4, p. 77-84, 2011.

Schlenker P., "Gestural grammar", *Natural Language & Linguistic Theory*p. 1-50, 2020.

Stalnaker R., "Common ground", *Linguistics and philosophy*, vol. 25, n° 5-6, p. 701-721, 2002.

Stojnić U., Stone M., Lepore E., "Pointing things out: in defense of attention and coherence", *Linguistics and Philosophy*p. 1-10, 2019.

Unger C., "Dynamic semantics as monadic computation", *JSAI International Symposium on Artificial Intelligence*, Springer, p. 68-81, 2011.

Van Benthem J., *Logical dynamics of information and interaction*, Cambridge, 2011.

Van Ditmarsch H., van Der Hoek W., Kooi B., *Dynamic epistemic logic*, vol. 337, Springer Science & Business Media, 2007.

Van Eijck J., Unger C., *Computational semantics with functional programming*, Cambridge, 2010.

Williams T., Bussing M., Cabrol S., Boyle E., Tran N., "Mixed reality deictic gesture for multimodal robot communication", *IEEE Int'l Conf. on HRI*, IEEE, p. 191-201, 2019.

Dialogue management with linear logic: the role of metavariables in questions and clarifications

Vladislav Maraev* — **Jean-Philippe Bernardy*** — **Jonathan Ginzburg****

* *Centre for Linguistic Theory and Studies in Probability (CLASP), Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg*
** *Laboratoire de Linguistique Formelle (LLF), CNRS – UMR 7110, Université de Paris*

ABSTRACT. *In this paper, we study the formalisation of a dialogue management system using proof-search on top of a linear logic. We argue that linear logic is the natural formalism to implement information-state dialogue management. We give particular attention to modelling question-answering sequences, including clarification requests, and argue that metavariables, arising from unification in the proof search, play a decisive role in providing a natural formalisation. We show that our framework is not only well suited from a theoretical perspective, but it is also suitable for implementation which we exemplify with a small scale implementation.*

RÉSUMÉ. *Cet article propose une formalisation de la gestion de dialogue via la recherche de preuves de formules de la logique linéaire. C'est-à-dire que nous proposons que la logique linéaire constitue une base naturelle de la formalisation de systèmes de gestion de dialogue basé sur un état d'information. Nous prêtons une attention particulière à la modelisation des séquences de questions-reponses (y compris les demandes de clarification), et nous arguons que les métavariables, résultant des unifications issue de la recherche de preuves, jouent un rôle décisif dans la formalisation. Nous montrons que notre système est non seulement adéquat d'un point de vue théorique, mais également d'un point de vue pratique. Ainsi, nous complétons notre argument d'une implementation d'un système de recherche de preuve générique, ainsi que d'un exemple de gestion de dialogue l'utilisant.*

KEYWORDS: *Symbolic dialogue management, Linear logic, Metavariables, Question answering, Clarification requests*

MOTS-CLÉS : *Gestion de dialogue symbolique, Logique linéaire, Métavariables, Question/Réponse, Demande de clarification*

TAL. Volume 61 – n°3/2020, pages 43 à 67

1. Introduction

A key aspect of dialogue systems design is the coherence of the system's responses. In this respect, a key component of a dialogue system is the dialogue manager, which selects appropriate system actions depending on the current state and the external context.

Two families of approaches to dialogue management can be considered: hand-crafted dialogue strategies (Larsson, 2002; Jokinen, 2009) and statistical modelling of dialogue (Rieser and Lemon, 2011; Young *et al.*, 2010; Huang *et al.*, 2020). Frameworks for hand-crafted strategies range from finite-state machines and form-filling to more complex dialogue planning and logical inference systems, such as Information State Update (ISU) (Larsson, 2002) that we employ here. Statistical models help to contend with the uncertainty that arises in human interaction; from noisy signals from speech recognition and other sensors to pragmatic ambiguities.

End-to-end systems that do not specify a dialogue manager as an explicit component have gained lots of attention recently (Huang *et al.*, 2020). Although most of them are focused on chit-chat dialogues, coherence plays a crucial role there too. Typically the main issues associated with such systems are related to memory limitations which cause repetition, contradiction and forgetfulness. Having a policy for dialogue coherence would be beneficial for such systems.

Although there has been a lot of development in dialogue systems in recent years, only a few approaches reflect advancements in *dialogue theory*. Our aim is to closely integrate dialogue systems with work in theoretical semantics and pragmatics of dialogue. This field has provided accounts for linguistic phenomena intrinsic to dialogue such as non-sentential utterances (Schlangen, 2003; Fernández *et al.*, 2007; Ginzburg, 2012), clarification requests (Purver, 2006; Ginzburg, 2012) and self-repair (Ginzburg *et al.*, 2014; Hough and Purver, 2012), where the resolution is intuitively tied to the coherence of what is being said.

To this end, a formal and in particular a logical representation is instrumental. This paper is concerned with the representation of participant states and transitions in a unified logical framework.

Even though the progress in bridging dialogue management and theoretical research of dialogue is promising, we believe that it is crucial to use formal tools which are most appropriate for the task, so that the formalisation and implementation of dialogue semantics closely matches the mental picture that experts have. In the view of Dixon *et al.* (2009) this is best done by representing the information-state of the agents as updatable sets of propositions. Subsets of propositions in the information state can be treated independently, and, therefore, a suitable and flexible way to represent updates is as propositions in linear logic. We adopt this view here, and further argue for it in the body of the paper.

We further extend the framework of Dixon *et al.* (2009) to deal with unclarity (and certain cases of non-probabilistic ambiguity). Indeed, asking a question is typi-

cally not done in one utterance which leaves nothing to interpretation. Typically, in a conversation a question and its answer may be many utterances apart, and the intermediate utterances form insertion sequences (Schegloff, 1972), for instance a follow-up clarification request and a corresponding answer. The insertion sequences are, in turn, conditioned on the preliminaries for the original question (Levinson, 1983, Chapter 6). In this paper, we deal with question-answering and clarification requests in a unified way, within a framework of dialogue management and using linear logic formalisation.

To deal with unclarity of an initial question, we propose here the use of *metavariables*, thereby leveraging much research on unification and proof search in various logical frameworks. That is, metavariables will stand in for any piece of information which is left to further interpretation. In particular, in this paper we explore the potential of using metavariables in the representation of question/answer exchanges.

By using a solid logical basis (Bratko, 2001; Girard, 1995) which corresponds well with the intuition of information-state based dialogue management, we are able to provide a fully working prototype [1] of the components of our framework:

1) a new proof-search engine based on linear logic, modified to support inputs from external systems (representing inputs and outputs of the agent);

2) a set of rules which function as a core framework for dialogue management (in the style of KoS (Ginzburg, 2012) theoretical account). The rules which we present below are provided to this engine, in the same form (modulo typesetting);

3) several examples which use the above to construct potential applications of the system. The engine is able to run domain-specific rules and generic rules together, forming a working system.

The rest of the paper is structured as follows. In section 2 we review important background for formalisation and implementation theories: dialogue management, linear logic and proof search. In section 3 we sketch a treatment of question, answers and clarification using the aforementioned formalisms. This treatment ignores certain dialogue management complexities, which we address in section 4. We discuss related work in section 5. Concluding remarks are provided in section 6.

2. Background

2.1. *Dialogue management*

2.1.1. *KoS*

KoS (not an acronym but loosely corresponds to Conversation Oriented Semantics) (Ginzburg, 2012) provides one of the most detailed theoretical treatments of domain-general conversational relevance, especially for query responses—see the work of

1. Source code and documentation are available at https://github.com/GU-CLASP/ProLin.

Purver (2006) on Clarification Requests, and Łupkowski and Ginzburg (2017) for a general account—and this ties into the KoS treatment of non-sentential utterances, again a domain crucial for naturalistic dialogue systems and where KoS has one of the most detailed analyses (Fernández *et al.*, 2007; Ginzburg, 2012).

In KoS (and other dynamic approaches to meaning), language is compared to a game, containing players (interlocutors), goals and rules. KoS represents language interaction by a dynamically changing context. The meaning of an utterance is then how it changes the context. Compared to most approaches, which represent a single context for both dialogue participants, KoS keeps separate representations for each participant, using the *Dialogue Game Board* (DGB). Thus, the information states of the participants comprise a private part and the dialogue gameboard that represents information arising from publicised interactions. The DGB tracks, at the very least, shared assumptions/visual field, moves (= utterances, form and content), and questions under discussion.

KoS is based on Cooper's formalism, Type Theory with Records (TTR) thus can leverage a wide range of work based on it, including the modelling of intentionality and mental attitudes (Cooper, 2005), generalised quantifiers (Cooper, 2013), co-predication and dot types in lexical innovation, frame semantics for temporal reasoning, reasoning in hypothetical contexts (Cooper, 2011), spatial reasoning (Dobnik and Cooper, 2017), enthymematic reasoning (Breitholtz, 2014), clarification requests (Purver, 2006; Ginzburg, 2012), negation (Cooper and Ginzburg, 2012), non-sentential utterance resolution (Fernández *et al.*, 2007; Ginzburg, 2012) and iconic gesture (Lücking, 2016). Being based on types and record-like contexts, we hope that our framework can also benefit from all this literature.

2.1.2. *Information-state update approach*

In this work we are employing an information-state update (ISU) approach, following several authors, including Larsson (2002) and Ginzburg (2012). In this view we present the information available to each participant of the dialogue (either a human or an artificial agent) in a rich information state. Being rich entails that the information state contains a hierarchy of facts, including the ones that are thought to be shared and the ones that have not been yet publicised.

Let us now consider the *update*, another essential component of ISU. In this case, we rely on a set of rules that will govern the updates. For instance, Ginzburg (2012) defines one of the most basic rules – the rule of QUD-incrementation – the procedure of updating the current set of questions under discussion (QUD) if the latest utterance is a question. This operation is salient to a user and therefore it constitutes the update of the public part of the information state.

The main benefit of using a rich representation of the information state with underspecified components is to be able to address a wide range of clarifications from both parties. This is especially beneficial in the case of automatic speech recognition or natural language understanding errors. But even putting such errors aside, we can

also consider topically relevant follow-up questions by the system, or contributions when the user provides more information than they were asked (over-answering).

2.1.3. *Questions and clarifications*

One of the greatest challenges in theoretical semantics and pragmatics is the treatment of interrogatives in the context of dialogue (Wiśniewski, 2015; Ginzburg, 2012). Here we distinguish *questions* as a general surface form and more contextualised forms of them, such as questions that initiate side sequences and constitute clarification requests (CRs). Side sequences usually refer to introducing some new question under discussion, for instance, requesting some additional information, whereas clarification requests generally account for cases of non-understanding, but the boundaries between them are often blurred. In the current study we exemplify our approach by accounting for requests for additional information, but it is only tested for the cases of system-initiated CRs.

For spoken dialogue systems it is crucial to be able to produce and process clarifications requests (Purver, 2004). Even though this is not our focus here, in the context of the low confidence of speech recognition and NLU, the system could clarify its input with the user. Further, with recent advances in speech recognition and statistical NLU, users expect to be able to initiate CRs themselves. Because our theory is symmetric with respect to users and systems roles, it can be useful in this context.

2.2. *Proof search as a programming language*

The prevailing tradition in formal semantics, including in most pieces of work cited above, is to represent (declarative) statements as propositions, formalised in an underlying logic (often first-order logic). In particular, in linguistic theories based on intuitionistic logic (such as TTR), true statements corresponds to propositions which admit a proof.

There is a long history of using proof search as a declarative programming paradigm, where the programmer specifies *axioms* and *rules of inference* which model their application domain. Typically such a system of axioms and rules represents a database of facts. For example, the axiom $(Leave\ 55\ Valand\ 11.50)$ can model the fact that bus 55 leaves from Valand at 11:50. The rule $(Leave\ x\ Valand\ y \rightarrow Arrive\ x\ CentralStationen\ (y + 45\ minutes))$ can represent travelling times on a certain line.

Then, the user may define a query (or goal) as a logical formula. The system can then search for a proof of a goal as a way to query the database of facts. Often, goals contain *metavariables,* [2] which play the role of unknowns for unification: their value can be fixed to any term for a goal to be reached. For example, the goal

2. Here, we use the convention that metavariables start with a lowercase letter, and constants (including predicates) with an upper case.

$(Leave\ x\ Valand\ y)$ corresponds to a request to list all the buses leaving from Valand (as x) together with their departure time (as y).

Because statements are propositions, it is only natural to use proof-search as a means to represent possible moves in dialogue seen as a game (Larsson, 2002).

2.3. *Linear logic as a Dialogue Management Framework*

Typically, and in particular in the archetypal logic programming language prolog (Bratko, 2001), axioms and rules are expressed within the general framework of first-order logic. However, several authors (Dixon *et al.*, 2009; Martens, 2015) have proposed to use linear logic (Girard, 1995) instead. For our purpose, the crucial feature of linear logic is that hypotheses may be used *only once*. For example, one could have a rule $IsAt\ x\ Valand\ y \multimap IsAt\ x\ CentralStationen\ (y + 45\ minutes)$. Consequently, after firing the above rule, the premiss $(Is\ x\ Valand\ y)$ becomes unavailable for any other rule. Thereby the linear arrow $\multimap$ can be used to conveniently model that a bus cannot be at two places simultaneously. [3]

In general, the linear arrow corresponds to *destructive state updates*. Thus, the hypotheses available for proof search correspond to the *state* of the system. In our application they will correspond to the *information state* of the dialogue participant. [4]

This way, firing a linear rule corresponds to triggering an *action* of an agent, and a complete proof corresponds to a *scenario*, i.e. a sequence of actions, possibly involving action from several agents. Hence, the actions realised as actual interactions constitute the observable dialogue. That is, an action can result in sending a message to the outside world (in the form of speech, movement, etc.). Conversely, events happening in the outside world can result in updates of the information state (through a model of the perceptory subsystem).

At any point in the scenario, the multiset of available *linear hypotheses* represents the current information-state of the agent which is modelled. To clarify, the information-state (typically in the literature and in this paper as well), corresponds to the state of a *single* agent. Thus, a scenario is conceived as a sequence of actions and updates of the information state of a single agent a, even though such actions can be attributed to any other dialogue participant b. (That is, they are a's representation of actions of b.)

3. If several arrows are present in a rule (such as $A \multimap B \multimap C$) then both A and B are consumed and C is produced.

4. We note, that in linear logic, facts (or hypotheses) do not come in a hierarchy. Either we have a fact, or we don't. However, in second-order variants of intuitionistic logic, like the one we use, one can conveniently wrap propositions in constructors, to indicate that they come with a qualification. For example, we can write $Unsure\ P$ to indicate that the proposition P may hold (for example if clarification is required).

To reiterate, in our implementation, the information-state can be queried using *rules* (such as those we list below). Because they are linear, these hypotheses can also be removed from the state, as we discuss in detail in section 4.

It is important to note that we will not forego the unrestricted (i.e. non-linear) implication ($\rightarrow$). Rather, both implications will co-exist in our implementation, thus we can represent simultaneously transient facts, or states (introduced by the linear arrow) and immutable facts (introduced by the unrestricted arrow). Besides, we have a *fixed* set of rules (they remain available even after being used), such as ($IsAt\ x\ Valand\ y \multimap IsAt\ x\ CentralStationen\ (y + 45\ minutes)$) above. Each such rule manipulates a part of the information state (captured by its premisses) and leaves everything else in the state unchanged.

3. Questions and clarifications

3.1. *Question-answering with metavariables*

In this subsection we show how a metavariable can represent what is being asked, as the unknown in a proposition. A first use for metavariables is to represent the requested answer of a question.

In this paper, we represent a question by a predicate P over a type A. That is, using a typed intuitionistic logic:

$$A : Type \qquad\qquad P : A \rightarrow Prop$$

The intent of the question is to find out about a value x of type A which makes $P\ x$ true, or at least entertained by the other participant. We provide several examples in Table 1. It is worth stressing that the type A can be large (for example asking for any location) or as small as a boolean (if one requires a simple yes/no answer). We note in passing that, typically, polar questions can be answered not just by a boolean but by qualifing the predicate in question, for example "maybe", "on Tuesdays", etc. (Table 1, last two rows). In this instance $A = Prop \rightarrow Prop$.

One complication are polar questions phrased in the negative (Cooper and Ginzburg, 2012); for example: "Doesn't John like bananas?". In this instance, a simple "no" answer can be ambiguous, and a possible model would be a multi-valued kind of answer ("yes he does" represented as $DefiniteYes$; "no he doesn't", represented as $DefiniteNo$, "no" as $AmbiguousNo$, and "He does in the weekend" as $Qualifier\ OnWeekend$):

$$\begin{array}{lll} Q\ Multi\ (\lambda x.\mathbf{case}\ x\ \mathbf{of} & AmbiguousNo & \rightarrow Trivial \\ & DefiniteNo & \rightarrow \neg\ (Like\ John\ Bananas) \\ & DefiniteYes & \rightarrow Like\ John\ Bananas \\ & Qualifier\ m & \rightarrow m\ (Like\ John\ Bananas)) \end{array}$$

To represent ambiguity in the case of $AmbiguousNo$, we make the answer provide no information, in the form of a trivial proposition (which is always true regardless

of context). This is a natural account, because the meaning of short answers (such as "no") always depends on the context. ("Paris" does not mean the same thing in the context of "Where do you live?" as in the context "Where were you born?".) Additionally, in the framework of a full dialogue management system, the $AmbiguousNo$ case should be treated as unresolving (the question effectively remains unanswered). However, in such a framework, it is always possible to receive a biasing answer ("I don't know") or no answer whatsoever. Even more complications are possible, by introduction of cases such as rhetorical questions. We deem such complications out of the scope of the current paper.

Within the state of the agent, if the value of the requested answer is represented as a metavariable x, then the question can be represented as: $Q\ A\ x\ (P\ x)$. That is, the pending question (Q denotes a question constructor) is a triple of a type, a metavariable x, and a proposition where x occurs. We stress that $P\ x$ is *not* part of the information state of the agent yet, rather the fact that the above question is *under discussion* is a fact. For example, after asking "Where does John live?", we have:

$$haveQud : QUD\ (Q\ Location\ x\ (Live\ John\ x))$$

Resolving a question can be done by communicating an answer. An answer to a question $(A : Type; P : A \rightarrow Prop)$ can be of either of the two following forms: i) A **ShortAnswer** is a pair of an element $X : A$ and its type A, represented as $ShortAnswer\ A\ X$ or ii) An **Assertion** is a proposition $R : Prop$, represented as $Assert\ R$. Therefore, one way to process a short answer is by the $processShort$ rule:

$$processShort : (a : Type) \rightarrow (x : a) \rightarrow (p : Prop) \rightarrow \\ ShortAnswer\ a\ x \multimap QUD\ (Q\ a\ x\ p) \multimap p$$

Above we use Π type binders to declare (meta)variables (written here $(a : Type) \rightarrow$, $(x : a) \rightarrow$, etc.). This terminology will make sense to readers familiar with dependent types. For the others, such binders can be thought as universal quantification ($\forall a, \forall x$, etc.), the difference is that the type of the bound variable is specified. (The reader worried about any theoretical difficulty regarding mixing linear and dependent types is directed to Atkey (2018) and Abel and Bernardy (2020).)

We demand in particular that types in the answer and in the question match (a occurs in both places). Additionally, because x occurs in p, the information state will mention the concrete x which was provided in the answer. For example, if the QUD was $(Q\ Location\ x\ (Live\ John\ x))$ and the system processes the answer $ShortAnswer\ Location\ Paris$, then x unifies with $Paris$, and the new state will include $Live\ John\ Paris$.

To process assertions, we can use the following rule:

$$processAssert : (a : Type) \rightarrow (x : a) \rightarrow (p : Prop) \rightarrow \\ Assert\ p \multimap QUD\ (Q\ a\ x\ p) \multimap p$$

That is, if (1) p was asserted, and (2) the proposition q is part of a question under discussion, and (3) p can be unified with q (we ensure this unification by simply using

question	A	P	reply	x
Where does John live?	$Location$	$\lambda x.Live\ John\ x$	in London	$ShortAnswer\ Location\ London$
Does John live in Paris?	$Bool$	$\lambda x.\mathbf{if}\ x\ \mathbf{then}\ (Live\ John\ Paris)$ $\mathbf{else}\ Not\ (Live\ John\ Paris)$	yes	$ShortAnswer\ Bool\ True$
What time is it?	$Time$	$\lambda x.IsTime\ x$	It is 5am.	$Assert\ (IsTime\ 5.00)$
Does John live in Paris?	$Prop \to Prop$	$\lambda m.m\ (Live\ John\ Paris)$	yes	$ShortAnswer\ (Prop \to Prop)$ $(\lambda x.x)$
Does John live in Paris?	$Prop \to Prop$	$\lambda m.m\ (Live\ John\ Paris)$	from January	$ShortAnswer\ (Prop \to Prop)$ $(\lambda x.FromJanuary\ (x))$

Table 1. *Examples of questions and the possible corresponding answers. The type A is the type of possible short answers. The proposition $P\ x$ is the interpretation of a short answer x. The x column shows the formal representation of a possible answer, either in short form or assertion form.*

the same metavariable p in both roles in the above rule), then the assertion resolves the question. Additionally, the metavariable x is made ground to a value provided by p, by virtue of unification of p and q. For example, "John lives in Paris" answers both questions "Where does John live?" and "Does John live in Paris?" (there is unification), but, not, for example "What time is it?" (there is no unification). Note that, in both cases ($processAssert$ and $processShort$), the information state is updated with the proposition posed in the question.

3.2. *Notion of unique and concrete values*

However, one should consider the question resolved only if the answer is "unique". For example, the assertion "John lives somewhere" generally does not resolve the question "Where does John live?". That is, if "somewhere" is represented by a metavariable, then the answer is not resolving.

Assume a two-place predicate Eat with agent as first argument and object as second argument. The phrase "John eats Mars" could then be represented as $(Eat\ John\ Mars)$. According to our theory, one can then represent the phrase "John eats" as $(Eat\ John\ x)$, with x being a metavariable. Assume now a system with the following state:

$$Eat\ John\ Mars$$

Then the question "What does John eat?, represented as $(Q\ Food\ x\ (Eat\ John\ x))$, can be answered. From the point of view of modelling with linear logic, we could attempt to model the answering by the rule as follows:

$$(a : Type) \to (x : a) \to (p : Prop) \to \\ QUD\ (Q\ a\ x\ p) \to p \multimap (p \otimes Answer\ x\ (Q\ a\ x\ p))$$

Note: taking a linear argument and producing it again is a common pattern, which can be spelled out $A \multimap (A \otimes P)$. It is so common that from here on we use the syntactic sugar $A \multimap P$ for it, so the above rule will be written:

$$(a : Type) \to (x : a) \to (p : Prop) \to \\ QUD\ (Q\ a\ x\ p) \to p \multimap Answer\ x\ (Q\ a\ x\ p)$$

The above states that if x makes the proposition p true (more precisely, provable — we require that p is a fact in the last argument) then it is valid to answer x if $Q\ a\ x\ p$ is under discussion. However, there is an issue with the above rule: there are several values making p true, i.e. if x is *not unique*, then intuitively one would not consider x a suitable answer. Indeed, assume instead that the system is in the state:

$$Eat\ John\ x$$

Then the question cannot be answered, because x stands for some unknown thing. The proper answer is then "I do not know".

Hence, we introduce another type-former $(x : A) \rightarrow_! B$. As for $(x : A) \rightarrow B$, it introduces the metavariable x. However, the rule fires only when x is made *ground* (it is bound to a term which does not contain any metavariable) and *unique* by matching the rule — this is what we call a unique and concrete value. That is, it won't match in the previous example, because the answer is not made ground (it contains unknowns). Additionally, it won't match if the state of the system is composed of the two hypotheses $(Eat\ John\ Mars)$ and $(Eat\ John\ Twix)$: the answer is not unique.

Thus, the rule for answering can be written like so:

$$produceAnswer : (a : Type) \rightarrow (x : a) \rightarrow_! (p : Prop) \rightarrow \\ QUD\ (Q\ a\ x\ p) \rightarrow p \multimap ShortAnswer\ a\ x$$

For example, if we have the following state:

$$QUD\ (Q\ Food\ x\ (Eat\ John\ x)) \\ Eat\ John\ Mars$$

The system can unify $QUD\ (Q\ Food\ x\ (Eat\ John\ x))$ and $QUD\ (Q\ a\ x\ p)$, yielding $a = Food$ and $p = (Eat\ John\ x)$. Then, we search for a proof p, and to do this, we can unify $(Eat\ John\ x)$ with $(Eat\ John\ Mars)$, giving finally the answer $x = Mars$ and therefore the state becomes:

$$Eat\ John\ Mars \\ ShortAnswer\ Food\ Mars$$

Note that the fact $Eat\ John\ Mars$ is found both as hypothesis and a conclusion of $produceAnswer$, and therefore it remains in the information state.

3.3. *Clarification requests and follow-up questions*

In this section we discuss an alternative kind of responding, which is to issue clarification requests. To see how they can occur, consider again the question "What does John eat?", in the information state $Eat\ John\ Mars$ and $Eat\ John\ Twix$. A proper answer could be "Mars and Twix" or even "Mars or Twix". However we consider here a third possibility: instead of answering, the agent can issue a clarification request.

To illustrate, consider the question "What is being eaten?" represented as $Q\ x\ (Eat\ y\ x)$), with the state

$$Eat\ John\ Mars \\ Eat\ Mary\ Mars$$

Then the agent can unambiguously answer "Mars": even if we do not know who we're talking about, it does not matter: only Mars is being eaten. However, if the state is

$Eat\ John\ Mars$
$Eat\ Mary\ Twix$

then, a probable answer would be a *clarification request*, namely "By whom?".

To detect situations where a clarification request can be issued, we can use the following rule (we leave unspecified the exact form of the CR abstract for now and come back to it below in section 4):

$$[\,a : Type; x : a; p : Prop; qud :: QUD\ (Q\ x\ p); proof :: p\,] \rightarrow_?\ CR$$

The conditions are similar to that of the answering rule. The principal difference is the use of the $\rightarrow_?$ operator, which takes as left operand the specification of a request and tests whether it has a non-unique solution or cannot be made fully ground. Essentially this does the opposite of the $\rightarrow_!$ operator. However, because the components of the query are indeterminate, they cannot be fixed when firing the rule, and therefore the state update cannot depend on them. Therefore we use a record syntax to limit their scope, ensuring that they won't occur in the state update. Such a record can be understood as a conjunction which additionally binds components to field names. Additionally, note the use of the single colon (:) for metavariables and the double colon for information-state hypotheses (::).

We can then turn our attention to the formulation of this clarification request. It is itself a question, and has a tricky representation:

$$Q\ Person\ z\ (z = y)$$

That is, the question is asking about some aspect which was left implicit in the original question (what is being eaten). In our terms, it must refer to the metavariable (y) which the original question included. After getting an answer (say $Mary$), z will be bound to a ground term, and, in turn, the fact $z = y$ will ensure that y becomes ground.

$Eat\ John\ Mars$
$Eat\ Mary\ Twix$
$ori :: QUD\ (Q\ Food\ x\ (Eat\ y\ x))$
$cr\ \ :: QUD\ (Q\ Person\ z\ (z = y))$
$a\ \ \ :: ShortAnswer\ Person\ Mary$

after applying $processShort$:

$Eat\ John\ Mars$
$Eat\ Mary\ Twix$
$ori :: QUD\ (Q\ Food\ x\ (Eat\ y\ x))$
$r\ \ \ :: Mary = y$

This means the original question will, by unification, become $Q\ \ Food\ \ x\ \ (Eat\ \ Mary\ \ x)$, and it can be unambiguously answered using the

$produceAnswer$ rule. We note that the logical form of the question (z such that $z = y$) is typically realised in a complicated way. In our example, it could be "By whom?"; echoing part of the original question and assuming cooperative communication so that the questioner properly relates the clarification request to the implicits of the original questions. In practice, the form of clarification questions will greatly vary depending on the context (Purver, 2004).

The above presupposes a clear-cut distinction: if an answer is unique, it is given; otherwise a clarification request is issued. However, answers could simply be exhaustive ("Mars or Twix"). If the original questioners are unhappy with the ambiguity, they are free to issue more precise questions. In practice, one can easily imagine an ambiguity threshold after which clarification requests are preferred. In the simplest form, this ambiguity threshold could be expressed by the length of the answer. In our example, if one has to list, say, 20 different kinds of food, it is easy to imagine that the answer won't be fully given. In fact, this question can be the topic of an experimental study.

3.3.1. *Clarification via adding extra arguments*

The scope of what is subject to clarification is anything which can be represented as an argument in a relation. For instance, consider the question "Where does John live?" with the short answer "Paris". The questionee may decide that there is some ambiguity about *which* location one is talking about — after all there are several places with this name. To be able to model this, the $Live$ relation needs to be generalised to be a 3-place predicate, where the country is specified.

However most of the time one may choose to leave this parameter implicit. This is what is done for example when asking the above question:

$$Q\ Location\ x\ (Live\ John\ x\ y)$$

If the question can be answered without regard for the country, then the metavariable will remain free for the duration of the dialogue. If on the other hand, answering the question demands clarification, this can be done using the mechanisms described above. In sum, in our model, to support clarification requests, a system must integrate many arguments and use metavariables.

The same technique can apply to polar questions. Considering "Does John live in Paris?", we can assume that the question can be encoded (for simplicity) as $\lambda x.\mathbf{if}\ x\ \mathbf{then}\ (Live\ John\ Paris\ y)\ \mathbf{else}\ Not\ (Live\ John\ Paris\ y)$.

If the system has the following facts:

$$Live\ John\ Paris\ France$$
$$Not\ (Live\ John\ Paris\ Denmark)$$

then both "True" and "False" are valid answers, and a clarification requests should be issued: $Q\ Country\ z\ (z = y)$. We see again that the realisation of the clarification request depends highly on the formulation of the question and the context. In this case "Do you mean Paris, France?" would be suitable.

3.3.2. *Clarification via adding named contextual parameters*

The above presentation (using a ternary predicate) is useful conceptually, but not ideal in practice: in the most general case one would end up with predicates with lots of arguments, for example country, county, district, etc.

However, there is a standard solution to the issue: because the country is functionally dependent on the location, these two concepts should be linked directly together rather than involve the $Live$ predicate. Using an intermediary entity type for locations and binary predicates, one can represent the question "Does John live in Paris?" as follows:

$$\lambda x.\mathbf{if}\ x\ \mathbf{then}\ (Live\ John\ y \rightarrow Name\ y\ Paris)\\ \mathbf{else}\ Not\ (Live\ John\ y \rightarrow Name\ y\ Paris)$$

Literally, "Does John live in a place called Paris?". The ambiguity of the $Paris$ name can be represented by several locations named $Paris$, X and Y in our illustration: [5]

$$Name\ X\ Paris$$
$$Name\ Y\ Paris$$
$$Live\ John\ X$$
$$Not\ (Live\ John\ Y)$$
$$Country\ France\ X$$
$$Not\ (Country\ France\ Y)$$

Because John lives in X but not in Y the question is ambiguous. One way to lift the ambiguity is to raise the clarification request as above. Here it can be phrased as a polar question [6] again:

$$Q\ Bool\ (\lambda x.\mathbf{if}\ x\ \mathbf{then}\ Country\ France\ y\ \mathbf{else}\ Not\ (Country\ France\ y))$$

3.3.3. *Summary*

In sum, we leverage a feature of linear-logic proof search: at any point in the scenario, the context can refer to metavariables. In a dialogue application, metavariables represent a certain amount of flexibility in the scenario: *so far* the scenario works for any value which could be assigned to the metavariable. This means that at a further point the metavariable can be instantiated to some other value.

4. KoS-inspired dialogue management with linear logic

In this section we integrate our question/answering framework within more complete dialog manager (DM). We stress that this DM models the information-state of

5. The combination of negation and proof search leads to complications which are out of scope here, for this reason we simply assume that negated predicates are available in the information-state.

6. Here we use the simpler version of the treatment of polar questions.

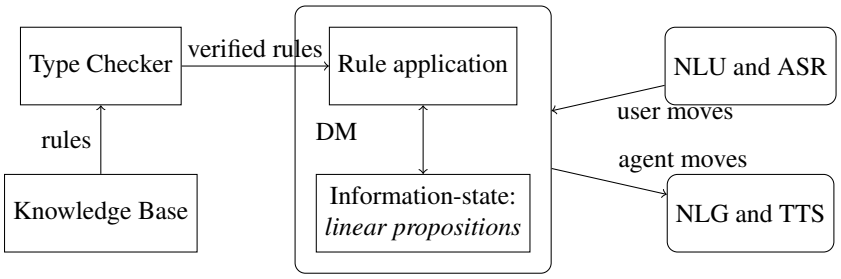


Figure 1. *Architecture of a spoken dialogue system with a dialogue manager based on a linear logic framework.*

only one participant. Regardless, this participant can record its own beliefs about the state of other participants. Figure 1 shows how such a DM can be integrated into a spoken dialogue system. In general, the core of DM is comprised of a set of linear-logic rules which depend on the domain of application. However, many rules will be domain-independent (such as generic processing of answers). We show these generic rules first, and then illustrate them with an example application.

4.1. *Domain-independent rules*

4.1.1. *Interface with language understanding and generation*

To be useful, a DM must interact with the outside world, and this interaction cannot be represented using logical rules, which can only manipulate data which is already integrated in the information state. Here, we assume that the information that comes from sources which are external to the dialogue manager is expressed in terms of semantic interpretations of moves, and contains information about the speaker and the addressee in a structured way. We provide 5 basic types of moves, specified with a speaker and an addressee, as an illustration:

$$
\begin{array}{ll}
Greet & spkr\ addr \\
CounterGreet & spkr\ addr \\
Ask & question\ spkr\ addr \\
ShortAnswer & vtype\ v\ spkr\ addr \\
Assert & p\ spkr\ addr
\end{array}
$$

These moves can either be received as input or produced as outputs. If they are inputs, they come from the NLU component, and they enter the context with $Heard : Move \rightarrow Prop$ predicate. For example, if one hears a greeting, the proposition $Heard\ (Greet\ S\ A)$ is added to the information state/context, without any rule being fired — this is what we mean by an external source.

If they are outputs, to be further used by the NLG component, some rule will place them in $Agenda$. For example, to issue a countergreeting, a rule will place the proposition $Agenda\ (CounterGreet\ A\ S)$ in the information state.

Thereby each move is accompanied by the information about who has uttered it, and towards whom was it addressed. All the moves are recored in the $Moves$ part of the participant's dialogue gameboard, as a $Cons$-list (stack).

Additionally, we record any move m which one has yet to actively react to, in an hypothesis of the form $Pending\ m$. We cannot use the $Moves$ part of the state for this purpose, because it is meant to be static (not to be consumed). $Pending$ thus allows one to make the difference between a move which is fully processed and a pending one.

4.1.2. *Initial state*

In general, we start with empty QUD and $Agenda$. A non-empty QUD can be prepared if, in a certain domain, some open questions are assumed from the start. The $Agenda$ might not be empty if one wants the system to initiate the conversation. There are also no moves: nothing has been said by either party.

$$\_ :: QUD\ Nil; \_ :: Agenda\ Nil; \_ :: Moves\ Nil;$$

(We often do not care about the proof object witnessing a propositions, in which case we denote it with an underscore.)

4.1.3. *Hearing*

The capacity of "hearing" or, in other words, starting the processing of semantic representations of utterances from the NLU component is implemented with the following rule:

$$\begin{array}{l} hearAndRemember: \\ \quad (m : DP \to DP \to Move) \to (x\ y : DP) \to (ms : List\ Move) \to \\ \quad Heard\ (m\ x\ y) \quad \multimap Moves\ ms \multimap HasTurn\ x \multimap \\ \quad [\_ :: Moves\ (Cons\ (m\ x\ y)\ ms); \_ :: Pending\ (m\ x\ y); \_ :: HasTurn\ y] \end{array}$$

where $(m\ \ x\ \ y)$ is a semantic representation of the utterance. Here we produce a record, whose fields will all be added to the information state. The rule demands that participant x has the turn and, as a result, turn was taken by his partner y.[7] The DP type stands for *dialogue participant*. As a result we do several things: i) place the move in a move list for further references ($PushMove$), ii) record the turn-switching (which in a complete system may not apply to all cases — then additional hypotheses would be added), and iii) prepare to process the move ($Pending$).

7. For now we have a very simple model of turn-taking, which can be improved in many ways: certain moves may not induce turn-change, there can be more than two participants, etc.

4.1.4. *Uttering*

The capacity of "uttering" represents an ability to generate information for the NLG component. NLP component is represented by $Agenda$ that contains a move that is just about to be uttered.

$$\begin{array}{l} utterAndRemember : \\ \quad (m : DP \to DP \to Move) \to (ms : List\ Move) \to (x\ y : DP) \to \\ \quad Agenda\ (m\ x\ y) \multimap Moves\ ms \multimap HasTurn\ x \multimap \\ \quad [\_ :: Utter\ (m\ x\ y); \_ :: Moves\ (Cons\ (m\ x\ y)\ ms); \_ :: HasTurn\ y] \end{array}$$

Here also we take care of turn-taking in the same rule. As a result, the system consumes the $Agenda$ and passes the move to the NLG component. The move is also memorised in the $Moves$ stack.

4.1.5. *Basic adjacency: greeting*

We can show how basic move adjacency can be defined in the example of counter-greeting preconditioned by a greeting from the other party:

$$\begin{array}{l} counterGreeting : (x\ y : DP) \to HasTurn\ x \multimap Pending\ (Greet\ y\ x) \multimap \\ \qquad\qquad Agenda\ (CounterGreet\ x\ y) \end{array}$$

4.1.6. *QUD incrementation*

Another important rule accounts for pushing the content of the last move, in the case if it is an Ask move, on top of the questions under discussion (QUD) stack.

$$\begin{array}{l} pushQUD : (q : Question) \to (qs : List\ Question) \to (x\ y : DP) \to \\ \qquad Pending\ (Ask\ q\ x\ y) \multimap QUD\ qs \multimap QUD\ (Cons\ q\ qs) \end{array}$$

4.1.7. *Integrating the answers*

If the user asserts something that relates to the top QUD, then the QUD can be resolved and therefore removed from the stack. The corresponding proposition p is saved as a $UserFact$.[8] This rule extends the abstract rule that were introduced in section 3.3.

$$\begin{array}{l} processAssert : (a : Type) \to (x : a) \to (p : Prop) \to (qs : List\ Question) \to \\ \quad (dp\ dp1 : DP) \to Pending\ (Assert\ p\ dp1\ dp) \multimap \\ \quad QUD\ (Cons\ (Q\ dp\ a\ x\ p)\ qs) \multimap [\_ :: UserFact\ p; \_ :: QUD\ qs] \end{array}$$

Short answers are processed in a very similar way to assertions:

8. For the current purposes we only remove the top QUD, but in a more general case we can implement the policy that can potentially resolve any QUD from the stack.

$$
\begin{array}{l}
processShort : (a : Type) \to (x : a) \to (p : Prop) \to (qs : List\ Question) \to \\
\quad (dp\ dp1 : DP) \to Pending\ (ShortAnswer\ a\ x\ dp1\ dp) \multimap \\
\quad QUD\ (Cons\ (Q\ dp\ a\ x\ p)\ qs) \multimap [\_ :: UserFact\ p; \_ :: QUD\ qs]
\end{array}
$$

4.1.8. *Questions and clarifications*

Just as we described in 3.2, we use uniqueness check to determine whether system can resolve the question ($produceAnswer$) or it needs to initiate a clarifying side sequence ($produceCR$).

$$
\begin{array}{l}
produceAnswer : \\
\quad (a : Type) \to (x : a) \to_! (p : Prop) \to (qs : List\ Question) \to \\
\quad QUD\ (Cons\ (Q\ USER\ a\ x\ p)\ qs) \multimap p \multimap \\
\quad [\_ :: Agenda\ (ShortAnswer\ a\ x\ SYSTEM\ USER); \_ :: QUD\ qs; \\
\quad\ \_ :: Answered\ (Q\ USER\ a\ x\ p)] \\
produceCR : \\
\quad [a : Type; x : a; p : Prop; qs : List\ Question; \\
\quad\ \_ :: QUD\ (Cons\ (Q\ USER\ a\ x\ p)\ qs); \_ :: p] \to_? CR
\end{array}
$$

The clarifying side sequence itself (CR) is meant to be specified by a dialogue developer, possibly informed by machine-learning systems, because it is domain-specific and the choice of the spectrum of possible options is wide. We provide an example of a domain-specific CR in the section 4.2 below.

4.2. *Example*

We now show how the generic system of rules above can handle the exchange:

U: Hello!
S: Hello, U.
U: When is there a bus from Valand?
S: In 15 minutes.

Let us further assume the following system context, which contains up-to-date public transport information in the following format:

$$TT\ Bus\ Time\ Origin\ Destination$$

This is added to the initial domain-independent context outlined above. We also assume that the user has the turn at the start.

$$
\begin{array}{ll}
QUD & Nil \\
Agenda & Nil \\
HasTurn & U \\
Moves & Nil
\end{array}
$$

When the system hears the greeting it can be integrated into the state using $hearAndRemember$ rule, therefore system updates its state accordingly:

$$\begin{array}{ll} QUD & Nil \\ Agenda & Nil \\ HasTurn & S \\ Moves & [Greet\ U\ S] \end{array}$$

(To save space we use a list notation from now on, [A, B, C] is a shorthand for $(Cons\ A\ (Cons\ B\ (Cons\ C)))$.) In this context the system can issue a counter-greeting by firing the $counterGreeting$ rule:

$$\begin{array}{ll} Agenda & (CounterGreet\ S\ U) \\ HasTurn & S \\ Moves & [Greet\ U\ S] \end{array}$$

Everything which is on the agenda can be uttered using $utterAndRemember$ rule, given that the system has the turn. System also hands the turn over to the user. Therefore, the state becomes (we use bracket syntax instead of $Cons$ for readability):

$$\begin{array}{ll} HasTurn & U \\ Moves & [CounterGreet\ S\ U, Greet\ U\ S] \end{array}$$

Now the system hears the question $(Ask\ (Q\ U\ Time\ t0\ (TT\ n0\ t0\ Valand\ d0)))$. It is domain-specific, and basically requests the timetable information for the given departure station. Again, we use $hearAndRemember$ rule to itegrate it into state, but also, because the move is Ask, the system sets its QUD to the question that the move contains with the $pushQUD$ rule.

$$\begin{array}{ll} QUD & [\ Q\ U\ Time\ t0\ (TT\ n0\ t0\ Valand\ d0)] \\ HasTurn & S \\ Moves & [\ Ask\ (Q\ U\ Time\ t0\ (TT\ n0\ t0\ Valand\ d0))\ U\ S, \\ & \quad CounterGreet\ S\ U, Greet\ U\ S] \end{array}$$

Now, depending on the state of the knowledge base, the system will have two options: i) produce the answer straight away, or ii) integrate a clarifying side sequence.

4.2.1. *Straight answer*

For this case we will consider a knowledge base that includes information just about the unique (w.r.t. the time) entry in the timetable:

$$TT\ B18\ T15\ Valand\ Johanneberg$$

Therefore the question can be resolved and the resolving short answer can be put on the $Agenda$.

$$\begin{array}{ll} Answered & (Q\ U\ Time\ T15\ (TT\ B18\ T15\ Valand\ Johanneberg)) \\ QUD & Nil \end{array}$$

$HasTurn\ \ S$
$Agenda\ \ (ShortAnswer\ Time\ T15\ S\ U)$
$Moves\ \ [\dots]$ -- same as above

4.2.2. *Clarifying side sequence*

In contrast, we can extend our minimal timetable example with another entry, therefore making it non-unique, w.r.t. time.

$TT\ B18\ T15\ Valand\ Johanneberg$
$TT\ B55\ T20\ Valand\ SciencePark$

In order to make it unique we can either clarify the bus number or the destination. For the bus number the rule for clarification can be formulated as follows:

$$
\begin{array}{l}
specificCR: \\
\quad (t : Time) \to (n : Bus) \to (s\ d : Location) \to (qs : List\ Question) \to \\
\quad CR \multimap QUD\ (Cons\ (Q\ U\ Time\ t\ (TT\ n\ t\ s\ d))\ qs) \multimap \\
\qquad [\_ :: QUD\ (Cons\ (Q\ S\ Bus\ n\ (WantBus\ n)) \\
\qquad\quad (Cons\ (Q\ U\ Time\ t\ (TT\ n\ t\ s\ d))\ qs)); \\
\qquad\ \ \_ :: Agenda\ (Ask\ (Q\ S\ Bus\ n\ (WantBus\ n))\ S\ U)]
\end{array}
$$

As a result of applying it, the state becomes:

$Agenda\ \ (Ask\ (Q\ S\ Bus\ n0\ (WantBus\ n0))\ S\ U)$
$QUD\ \ [Q\ S\ Bus\ n0\ (WantBus\ n0),$
$\qquad\ Q\ U\ Time\ t0\ (TT\ n0\ t0\ Valand\ d0)]$
$HasTurn\ S$
$Moves\ \ [\dots]$ -- same as above

Then, the system can utter the clarification request ($utterAndRemember$ rule):

$QUD\ [Q\ S\ Bus\ n0\ (WantBus\ n0), Q\ U\ Time\ t0\ (TT\ n0\ t0\ Valand\ d0)]$
$HasTurn\ S$
$Moves\ [Ask\ (Q\ S\ Bus\ n0\ (WantBus\ n0))\ S\ U$
$\qquad\ Ask\ (Q\ U\ Time\ t0\ (TT\ n0\ t0\ Valand\ d0))\ U\ S$
$\qquad\ CounterGreet\ S\ U, Greet\ U\ S]$

The user can reply to this with a short answer $ShortAnswer\ Bus\ B55\ U\ S$ or an assertion $Assert\ (WantBus\ B55)\ U\ S$, which can be integrated using $processShort$ or $processAssert$ rule respectively. We show the state after processing the short answer:

$QUD\ [Q\ U\ Time\ t0\ (TT\ B55\ t0\ Valand\ d0)]$
$UserFact\ (WantBus\ B55)$
$HasTurn\ S$
$Moves\ [ShortAnswer\ Bus\ B55\ U\ S,$
$\qquad\ Ask\ (Q\ S\ Bus\ B55\ (WantBus\ B55))\ S\ U, \dots]$

The reader can see that the metavariable $n0$ from the previous state is now unified with $B55$ in the QUD, therefore it now corresponds to one unique entry in the knowledge base. Hence, the answer can be issued by the $produceAnswer$ rule.

$$
\begin{array}{ll}
Answered & (Q\ U\ Time\ T20\ (TT\ B55\ T20\ Valand\ SciencePark)) \\
QUD & Nil \\
Agenda & (ShortAnswer\ Time\ T20\ S\ U) \\
UserFact & (WantBus\ B55) \\
HasTurn & S \\
Moves & [\dots] \quad \text{-- same as above}
\end{array}
$$

5. Related work

The present work provides a minimal and fine-grained account for clarification requests initiated by any conversational party, following accounts of and supporting a subset of cases thoroughly investigated in the CLARIE Prolog-based system (Purver, 2006), following corpus studies by Purver (2004) and Rodríguez and Schlangen (2004).

One of our main sources of inspiration is Ginzburg's KoS (Ginzburg, 2012). However we recast it in the framework of proof search, and linear logic. We have argued that this has many advantages. First, it affords the use of metavariables to represent uncertainty, which is absent from TTR. Second, expressing updates using linear logic rules means that only the relevant parts of the information state must be dealt with in any given rule. Cooper's TTR has a special "asymmetric merge" operator for this purpose, but it is a less-studied *ad-hoc* addition to type-theory, though see *inter alia* (Grover *et al.*, 1994). As it stands, KoS is lacking implementations, with the exception of the work of Maraev *et al.* (2018), who adapt KoS to eschew the assymetric merge operation. An oft-touted advantage of TTR is that propositions are witnessed by proof objects. We benefit from the same advantage: we use an intuitionistic system, and as such every proposition in the information state is associated a witness, even if we have not shown them for concision (they play little role in our analysis).

Larsson (2002) proposed the use of Prolog (and hence, proof search), as a dialogue management framework. However, the lack of linear hypotheses means that destructive information-state updates are sometimes awkward to represent. Besides, he does not consider the use of metavariables to represent uncertainty — even though Prolog in principle has the capacity to do it.

To our knowledge Dixon *et al.* (2009) were the first to advocate the use of linear logic for dialogue management and planning. Compared to the present work, they focus primarily on the planning part of dialogue rather than question-answering. In particular, they do not discuss the role of metavariables and clarification requests. We additionally propose the extension of linear logic with special-purpose operators $X \to_! Y$ and $X \to_? Y$ to distinguish the presence or the absence of ambiguity.

6. Evaluation/Discussion/Future work

A kind of dialogue move often studied in parallel to clarifications are *corrections*. It would be elegant if corrections could be formalised in a way similar to clarifications. However, in our analysis, metavariables disappear once they have been grounded. Therefore, corrections cannot involve metavariables and thus require a different treatment. A solution could be to keep metavariables in terms (apply unification substitutions only at the point of testing equality between such variables). We leave a detailed study to further work.

We note that the use of (meta)variables to refer to discourse objects is a very general device. Anything which can be subject to clarification can occur as an argument to predicates. We already showed how "Paris" can be clarified. But we could also clarify "Live" by making the verb be an argument to a general $Apply$ predicate, taking say a verb and its arguments.

Prior studies have noted the phenomenon of semantic dependency relations between questions (Wiśniewski, 2015), e.g. "Who killed Bill?" can be responded by "Who was in town?". The cases of dependencies covered in this study are limited to clarification of metavariables from the original question. This is meant to serve as a proof-of-concept rather than thorough coverage of all possible cases of question dependence. A similar issue concern follow-up questions that are meant to clarify the type of the metavariable, e.g. "What does John like? Do you mean foodwise?". Generally, further work is needed to be carried out in order to extend our system to full-scale coverage of interrelations between QUDs.

A natural progression of this work is to allow the assignment of probabilities to rules and to the components of the state, and to train the probabilities according to the new observations. Our approach follows Lison (2015), which is based on probabilistic rules, but in our case the structure of information state is rich and derived from the theoretical outlook on dialogue, and dialogue management has a core set of domain-independent rules. We can also imagine combining such ideas with probabilistic meaning for sentences (Goodman and Lassiter, 2015; Bernardy *et al.*, 2018).

An important dimension of dialogue processing that the current work does not address is providing a detailed utterance processing of the user and word-by-word incremental processing. This means we cannot deal with form-based parallelism needed for various types of acknowledgements, CRs, and self-repair. Nor, as things stand, do we engage in grounding interaction, modelled extensively by Larsson (2002).

Table 2 originates from Ginzburg and Fernández (2010), who proposed a series of benchmarks for comparing different approaches to developing dialogue systems (see section 2 of that paper). For each approach the symbol ✓ indicates that the current approach safisfies the benchmark in the corresponding row; $\sim$ that the benchmark could be met with some caveats, as explained in the text above for most cases; and — that the benchmark is not met by a standard version of the current approach.

	Benchmark		Example
query and assertion	Q1 simple answers	$\sim$	A: Who slept? B: Bo/Not Bo
	Q2a non-resolving answers	✓	A: Who slept? B: A student.
	Q2b follow up queries	✓	B: A student. A: Who?
	Q3 overinformative answers	✓	A: Who? B: Bo on his own.
	Q4 sub-questions	✓	A: Who? B: Who was here?
	Q5 topic changing	—	
	A1 propositional content update	✓	
	A2 disagreement	$\sim$	A: A student. B: A teacher.
	SC scalability	$\sim$	
	DA domain adaptability	✓	
metacommunication	Ack1 completed acknowledgements	—	A: Move right. B: Mhm.
	Ack2 continuation ack.	—	A: Move- B: mm A: -to the left.
	Ack3 gestural ack.	—	
	CR1 repetition CRs	—	A: Did Bo leave? B: What?
	CR2 confirmation CRs	—	A: Bill left. B: Bill? A: Yes.
	CR3 intended content CRs	✓	A: Where is Bo? B: Which Bo?
	CR4 intention recognition CRs	—	A: Where is the bus? B: Why?
	SND distinct updates	$\sim$	
	FG fine-grained representations	$\sim$	
fragments	SF1 wide coverage of NSUs	$\sim$	
	SF2 basic answer resolution	$\sim$	
	SF3 reprise fragment resolution	—	Bo? $\mapsto$ Who is Bo?
	SF4 long distance short answers	$\sim$	
	SF5 genre sensitive initiating NSUs	$\sim$	(dialogue initially) The Aix bus?
	D1 recognize and repair disfluencies	—	
	D2 keep disfluencies in context	—	

Table 2. *System evaluation. Q5—understand that irrelevant answers imply "change the topic", A2—disagree with user if her utterance is incompatible with own belief, SND—an utterance can give rise to distinct updates across participants. SC—ensure approach scales down to monologue and up to multilogue. For other, more obvious benchmarks we refer our readers to (Ginzburg and Fernández, 2010).*

Acknowledgements

This research was supported by a grant from the Swedish Research Council for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg. We also acknowledge support by a public grant overseen by the French National Research Agency (ANR) as part of the program "Investissements d'Avenir" (reference: ANR-10-LABX-0083). It contributes to the IdEx Université de Paris — ANR-18-IDEX-0001. We also acknowledge a senior fellow-

ship from the Institut Universitaire de France to Ginzburg. In addition, we would like to thank our anonymous reviewers for their useful comments.

7. References

Abel A., Bernardy J.-P., "A unified view of modalities in type systems", *Proceedings of the ACM on Programming Languages*, 2020.

Atkey R., "Syntax and Semantics of Quantitative Type Theory", *Proceedings of the 33rd Annual ACM/IEEE Symposium on Logic in Computer Science, LICS 2018, Oxford, UK*, p. 56-65, 2018.

Bernardy J.-P., Blanck R., Chatzikyriakidis S., Lappin S., "A Compositional Bayesian Semantics for Natural Language", *Proceedings of the International Workshop on Language, Cognition and Computational Models, COLING 2018, Santa Fe, New Mexico*, p. 1-11, 2018.

Bratko I., *Prolog programming for artificial intelligence*, Pearson education, 2001.

Breitholtz E., "Reasoning with topoi–towards a rhetorical approach to non-monotonicity", *Proceedings of the 50th anniversary convention of the AISB, University of London*, 2014.

Cooper R., "Austinian truth, attitudes and type theory", *Research on Language and Computation*, vol. 3, n° 4, p. 333-362, 2005.

Cooper R., "Copredication, quantification and frames", *in* S. Pogodalla, J.-P. Prost (eds), *Logical Aspects of Computational Linguistics (LACL 2011)*, Springer, 2011.

Cooper R., "Clarification and generalized quantifiers", *Dialogue and Discourse*, vol. 4, p. 1–25, 2013.

Cooper R., Ginzburg J., "Negative inquisitiveness and alternatives-based negation", *Logic, Language and Meaning*, Springer, p. 32-41, 2012.

Dixon L., Smaill A., Tsang T., "Plans, actions and dialogues using linear logic", *Journal of Logic, Language and Information*, vol. 18, n° 2, p. 251-289, 2009.

Dobnik S., Cooper R., "Interfacing language, spatial perception and cognition in Type Theory with Records", *Journal of Language Modelling*, vol. 5, n° 2, p. 273-301, 2017.

Fernández R., Ginzburg J., Lappin S., "Classifying Ellipsis in Dialogue: A Machine Learning Approach", *Computational Linguistics*, vol. 33, n° 3, p. 397-427, 2007.

Ginzburg J., *The Interactive Stance*, Oxford University Press, 2012.

Ginzburg J., Fernández R., "Computational Models of Dialogue", *The Handbook of Computational Linguistics and Natural Language Processing*, vol. 57, p. 1, 2010.

Ginzburg J., Fernández R., Schlangen D., "Disfluencies as Intra-Utterance Dialogue Moves", *Semantics and Pragmatics*, vol. 7, n° 9, p. 1-64, 2014.

Girard J.-Y., *Linear Logic: its syntax and semantics*, London Mathematical Society Lecture Note Series, Cambridge University Press, p. 1–42, 1995.

Goodman N., Lassiter D., "Probabilistic Semantics and Pragmatics: Uncertainty in Language and Thought", *in* S. Lappin, C. Fox (eds), *The Handbook of Contemporary Semantic Theory, Second Edition*, Wiley-Blackwell, Malden, Oxford, p. 655-686, 2015.

Grover C., Brew C., Manandhar S., Moens M., "Priority Union and Generalization in Discourse Grammars", *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*, ACL '94, Association for Computational Linguistics, USA, p. 17–24, 1994.

Hough J., Purver M., "Processing Self-Repairs in an Incremental Type-Theoretic Dialogue System", *Proceedings of SemDial 2012 (SeineDial)*, p. 136-144, September, 2012.

Huang M., Zhu X., Gao J., "Challenges in building intelligent open-domain dialog systems", *ACM Transactions on Information Systems (TOIS)*, vol. 38, nº 3, p. 1-32, 2020.

Jokinen K., *Constructive dialogue modelling: Speech interaction and rational agents*, vol. 10, John Wiley & Sons, 2009.

Larsson S., Issue-based dialogue management, PhD thesis, University of Gothenburg, 2002.

Levinson S. C., *Pragmatics*, Cambridge University Press, Cambridge, U.K., 1983.

Lison P., "A hybrid approach to dialogue management based on probabilistic rules", *Computer Speech & Language*, vol. 34, nº 1, p. 232-255, 2015.

Lücking A., "Modeling Co-Verbal Gesture Perception in Type Theory with Records", *Proceedings of the 2016 Federated Conference on Computer Science and Information Systems*, p. 383-392, 2016.

Łupkowski P., Ginzburg J., "Query responses", *Journal of Language Modelling*, vol. 4, nº 2, p. 245-292, 2017.

Maraev V., Ginzburg J., Larsson S., Tian Y., Bernardy J.-P., "Towards KoS/TTR-based proof-theoretic dialogue management", *Proceedings of the 22nd Workshop on the Semantics and Pragmatics of Dialogue*, SEMDIAL, Aix-en-Provence, France, November, 2018.

Martens C., Programming Interactive Worlds with Linear Logic, PhD thesis, Carnegie Mellon University Pittsburgh, PA, 2015.

Purver M., "CLARIE: Handling Clarification Requests in a Dialogue System", *Research on Language & Computation*, vol. 4, nº 2, p. 259-288, 2006.

Purver M. R. J., The theory and use of clarification requests in dialogue, PhD thesis, University of London, 2004.

Rieser V., Lemon O., *Reinforcement learning for adaptive dialogue systems: a data-driven methodology for dialogue management and natural language generation*, Springer Science & Business Media, 2011.

Rodríguez K. J., Schlangen D., "Form, intonation and function of clarification requests in German task-oriented spoken dialogues", *Proceedings of Catalog (the 8th workshop on the semantics and pragmatics of dialogue; SemDial04)*, 2004.

Schegloff E. A., "Sequencing in conversational openings", *Directions in sociolinguistics*, 1972.

Schlangen D., A coherence-based approach to the interpretation of non-sentential utterances in dialogue, PhD thesis, University of Edinburgh. College of Science and Engineering, 2003.

Wiśniewski A., "Semantics of questions", *The Handbook of Contemporary Semantic Theory*, Wiley Online Library, p. 273-313, 2015.

Young S., Gašić M., Keizer S., Mairesse F., Schatzmann J., Thomson B., Yu K., "The hidden information state model: A practical framework for POMDP-based spoken dialogue management", *Computer Speech & Language*, vol. 24, nº 2, p. 150-174, 2010.

Comparaison linguistique et neuro-physiologique de conversations humain humain et humain robot

Charlie Hallart* — **Juliette Maes*** — **Nicolas Spatola**** — **Laurent Prévot***** — **Thierry Chaminade***

* *Aix-Marseille Université, CNRS, INT, Marseille, France, prénom.nom@univ-amu.fr,*
** *Italian Institute of Technology, Genova, Italy,*
*** *Aix-Marseille Université, CNRS, LPL, Aix-en-Provence, France.*

RÉSUMÉ. *Nous décrivons l'analyse d'un corpus de conversations humain-humain et humain-robot. Vingt et un participants ont été scannés en imagerie par résonance magnétique fonctionnelle (IRMf) pendant qu'ils discutaient soit avec un humain, soit avec un robot. En s'inspirant de ce qui est communément utilisé pour étudier les conversations, huit variables linguistiques adaptées aux spécificités du corpus ont permis de mettre en évidence les compétences linguistiques limitées du système de magicien d'Oz utilisé pour contrôler le robot. Nous avons également adapté une variable d'alignement lexical qui nous permet d'étudier l'alignement conversationnel, plus important dans les interactions avec le robot qu'avec l'humain. Enfin, nos résultats de neuro-imagerie suggèrent une réduction du contrôle cognitif associée à l'augmentation de l'alignement lexical du participant sur l'interlocuteur.*

ABSTRACT. *Here we describe the analysis of a unique corpus of conversations with a human or a robot. Twenty-one participants were scanned with functional Magnetic Resonance Imaging (fMRI) while talking either with a human or with a robot. Inspired by what is commonly studied in natural conversations, eight linguistic variables adapted to the specifics of the corpus highlight the limited linguistic skills of the Wizard of Oz system used to control the robot. We also calculate a lexical alignment variable which allows us to study the phenomenon of conversational alignment, increased with the robot compared to the human. Finally, our neuroimaging results suggest a reduction in cognitive control associated with an increase in the participant's lexical alignment with the interlocutor.*

MOTS-CLÉS : *conversation, humain, robot, neurosciences, alignement lexical.*

KEYWORDS: *conversation, human, Robot, neurosciences, lexical alignment.*

TAL. Volume 62 – n°1/2021, pages 69 à 93

1. Introduction

La conversation, activité complexe et ordonnée, est généralement décrite comme un échange de propos libres et spontanés entre plusieurs locuteurs autour d'un thème commun. Défi déjà complexe, l'analyse conversationnelle voit grandir, avec le développement de l'intelligence artificielle et des interfaces humain-machine, l'importance d'étudier non seulement les conversations naturelles, mais également les conversations avec des agents artificiels. En effet, que les individus conversent avec l'assistant vocal de leur téléphone ou avec un robot humanoïde, les conversations entre les humains et les interfaces artificielles s'intègrent de plus en plus au quotidien. Il est pour cela primordial de comprendre les mécanismes qui sous-tendent ces interactions.

Dans ce contexte, cet article décrit l'exploitation d'un corpus qui combine des données linguistiques, comportementales et neurophysiologiques synchronisées de vingt et un participants (terme utilisé dans cet article pour parler du conversant installé dans le scanner IRM dont on mesure l'activité cérébrale). Le corpus rassemble les enregistrements de conversations naturelles qui ont eu lieu entre chacun des participants et un interlocuteur (terme utilisé pour le conversant situé dans une pièce annexe), qui peut être soit un humain, soit un robot.

Nous partons de l'observation que l'interlocuteur artificiel possède des compétences linguistiques différentes de celles de l'humain. Le robot est contrôlé par un système de magicien d'Oz doté de compétences linguistiques limitées (nombre de phrases restreint, intonation peu variable, latence dans le temps de réponse...). Ces limitations peuvent sembler fortes au regard de certaines démonstrations de robots humanoïdes mais il faut rappeler que ces démonstrations pour le grand public résultent de nombreuses heures de programmation et suivent des scénarios assez restreints. À l'inverse, il s'agit ici d'une réelle conversation où il faut que les réponses du robot soient adaptées aux interventions parfois imprévisibles du participant. Nous avons utilisé au maximum de ses capacités ce robot conversationnel. La voix utilisée, notamment, était une voix francophone d'Amazon Polly, assez naturelle mais peu expressive.

Cette recherche s'inscrit dans un projet ayant pour but l'amélioration des compétences linguistiques et sociales des robots pour en étudier les effets sur le comportement et l'activité cérébrale. Dans notre protocole, les participants sont amenés à croire que le robot est autonome, ce qui est renforcé par les capacités conversationnelles limitées et la latence de réponse du système de magicien d'Oz utilisé pour le contrôle.

En partant de la proposition que nous adoptons une posture différente selon que nous agissons avec un être doué d'états mentaux ou avec une machine (Dennett, 1987), les participants devraient se comporter différemment avec ces deux interlocuteurs. Pour ces raisons, nous posons l'hypothèse que la nature des interlocuteurs a un effet sur la production langagière des participants, et son évolution au cours du temps.

Dans un second temps, nous nous intéressons au phénomène d'alignement conversationnel, un processus cognitif que la littérature linguistique décrit comme un principe fort et robuste des interactions humaines (Branigan *et al.*, 2000 ; Branigan

et al., 2007). Plus spécifiquement, l'alignement conversationnel se produit à un niveau verbal ou paraverbal lorsque les individus en interaction emploient un lexique commun, des structures syntaxiques identiques ou encore les mêmes patrons prosodiques (Pickering et Garrod, 2004). Notre corpus, constitué de conversations naturelles bidirectionnelles entre deux humains, mais également de conversations comparables entre un humain et un robot, nous permet d'étudier l'alignement selon deux directions. En effet, nous étudions l'alignement conversationnel du participant sur l'interlocuteur selon que ce dernier est humain ou robot, mais nous pouvons également étudier l'alignement conversationnel de l'interlocuteur sur le participant. En partant des travaux de Branigan *et al.* (2000) et Branigan *et al.* (2007) selon lesquels les locuteurs tendent à s'aligner davantage avec un robot dont les capacités langagières sont limitées, nous posons l'hypothèse que le participant s'aligne davantage avec l'interlocuteur robot plutôt qu'avec l'interlocuteur humain. En considérant ensuite que le comportement langagier du robot est limité par le paradigme du magicien d'Oz, et que l'interlocuteur humain est ainsi le seul capable d'adapter finement son discours, nous posons l'hypothèse que l'interlocuteur humain s'aligne davantage sur le participant que l'interlocuteur robot.

Pour vérifier cette hypothèse, nous adaptons une variable d'alignement lexical qui nous permet d'étudier le lexique que les locuteurs utilisent afin de faire référence à des objets ou à des concepts communs. Les données de neuro-imagerie associées à ces interactions nous permettent de tester que cette nouvelle variable d'alignement correspond à une réalité cognitive en recherchant ses corrélats cérébraux. Notre approche statistique est d'abord validée en l'utilisant pour vérifier qu'elle identifie correctement les corrélats cérébraux de la production et de la perception de langage (Price, 2010).

Bonjour Je m'appelle Furhat Comment ça va ?
Oui Non Peut-être
C'est une poire jaune La poire semble triste Peut-être qu'elle est malade et elle devenue pourrie
Tu as une idée du message ? C'est peut-être une campagne pour favoriser les fruits locaux Ça pourrait être une pub pour des producteurs de fruits

Tableau 1. *Exemple de phrases pré-enregistrées prononcées par le robot, groupées selon leur fonction dans la conversation : présentations, réponses génériques, descriptions d'une image (une poire dans la série des fruits pourris), et échanges sur le message de la campagne de pub.*

Participant : du coup là c'est une poire
Participant : euh
Participant : un peu cabossée
Participant : et euh
Participant : du coup il semblerait qu'elle soit avinée
Participant : et
Interlocuteur : ouais
Interlocuteur : ouais ouais
Participant : et euh en fait euh
Participant : au départ je pensais qu'elle était triste mais au final non j'ai l'impression que c'est un petit sourire en coin
Interlocuteur : ah ouais
Participant : ouais je vois pas - fin bon tristesse - fin c'est neutre euh
Participant : c'est pas euh si triste
Interlocuteur : moi elle m'avait l'air euh je sais pas comment dire dépitée peut-être
Participant : comment
Interlocuteur : dépitée peut-être pour moi
Participant : ah oui peut-être ouais
Participant : oui un peu perplexe dépitée mais pas triste au final
Interlocuteur : ouais perdue aussi comme tu disais
Participant : oui peut-être
Participant : ouais
Interlocuteur : ouais
Interlocuteur : ça fait euh ça fait un gros une grosse différence avec euh les fruits de la première campagne
Participant : euh oui
Participant : totalement et aussi avec la framboise qui est plus euh
Participant : mh
Interlocuteur : gélatineuse
Participant : plus d'émotions - fin
Interlocuteur : ah
Participant : oui
Interlocuteur : moi je pensais à @ à l'aspect nourriture
Participant : à la texture

Tableau 2. *Transcription complète des échanges d'un essai entre un participant et l'humain.*

2. Corpus

Les fondements théoriques qui sous-tendent le choix du paradigme expérimental et les procédures d'acquisition et de préparation du corpus, à la fois pour les données linguistiques et neurophysiologiques, ont été publiés ces dernières années

Participant : alors là c'est une fraise qui est encore abîmée
Participant : et et qui avait l'air
Participant : perdue
Participant : et défoncée
Participant : pour moi
Participant : euh
Interlocuteur : comme les deux autres
Participant : euh non
Participant : pas trop non les autres ils avaient plus une expression euh de douleur ou euh
Participant : *
Participant : et euh
Participant : voilà
Interlocuteur : peut-être
Interlocuteur : cette fraise est déformée
Participant : oui
Participant : sur les côtés
Interlocuteur : la fraise est aussi pourrie
Participant : euh
Interlocuteur : qu'est ce que tu en dis
Participant : bah pourrie je sais pas mais déformée oui
Participant : abîmée
Interlocuteur : la fraise est un peu abîmée
Participant : ouh ben là c-
Participant : c'est comme
Participant : comme la poire en fait
Interlocuteur : comme la poire et la framboise

Tableau 3. *Transcription complète d'un échange entre un participant et le robot. Il faut noter qu'il s'agit de l'essai suivant directement celui donné dans le tableau 2.*

(Chaminade, 2017 ; Rauchbauer *et al.*, 2019 ; Chaminade *et al.*, 2018). Toutefois, et même si nous invitons le lecteur à utiliser ces références pour plus de détails, nous souhaitons rappeler les principaux points pour que cet article soit lisible de manière autonome.

2.1. *Cadre de l'expérience*

Afin d'étudier les interactions sociales naturelles, il est essentiel que les participants ne soient pas conscients du véritable objectif de l'expérience. À cette fin, l'expérience est présentée comme une expérience de neuromarketing, dans laquelle une entreprise veut savoir s'il suffit de discuter à propos des images d'une campagne de publicité à venir, soit avec une autre personne, soit avec une intelligence artificielle

incarnée dans un robot conversationnel, pour deviner le message de la campagne (Chaminade, 2017). L'interlocuteur humain est un expérimentateur, du même genre que le participant, mais présenté au participant comme un participant naïf comme lui. Le robot est une tête robotique conversationnelle rétroprojetée (Furhat robotics[1]) Il possède une apparence physique anthropomorphique incluant un visage, un genre, une voix et divers accessoires humains (perruque, casque audio, lunettes) pour ressembler à l'interlocuteur humain (figure 1, gauche). Il est contrôlé par l'interlocuteur humain avec un système de magicien d'Oz : il dispose d'un ensemble fini de réponses pré-enregistrées (voir tableau 1) que l'expérimentateur choisit en appuyant sur des boutons virtuels sur une tablette tactile. Certaines réponses sont génériques et d'autres spécifiques d'une image ou du message de la campagne publicitaire. Ne pouvant pas répondre à des concepts ou des mot-clés qui n'ont pas été programmés, les échanges avec le robot sont donc nécessairement limités en termes de variété et de spontanéité. Il n'y a pas d'intonation sauf pour les phrases interrogatives et le système de magicien d'Oz induit un délai qui perturbe la spontanéité des échanges. Ainsi, les participants discutent en fait avec la même personne dans les conversations avec l'humain et le robot mais la médiation par le robot appauvrit significativement la qualité de la conversation, alors que la croyance en son autonomie modifie la posture intentionnelle (Dennett, 1987) du participant.

2.2. *Acquisition des données*

Vingt et un participants sont inclus dans les analyses présentées ici (15 femmes, m = 27,04 ans, e.t. = 8,19, [21-49]). Le premier participant est exclu puisque les données comportementales de sa quatrième session sont manquantes. Les participants 4 et 23 sont exclus pour manque de participation active à l'expérience. Le participant 19 est exclu pour cause de mauvaise audition de l'interlocuteur robot pendant l'expérience.

Les participants se présentent au centre IRM, où un expérimentateur leur introduit les deux interlocuteurs, l'expérimentateur humain et le robot, ainsi que l'histoire de neuromarketing. Quatre sessions d'IRM sont enregistrées successivement. Dans chacune des sessions, le participant parle trois fois 1 minute avec l'humain, et trois fois 1 minute avec le robot, alternativement, en commençant chaque session avec l'humain. Chaque conversation de 1 minute est appelée un essai. Au total, chaque participant réalise 24 essais, conduisant à 24 minutes de conversation enregistrées (douze fois 1 minute avec l'humain, douze fois 1 minute avec le robot). Avant chaque essai, le participant voit apparaître sur l'écran pendant 8,3 secondes une image qui illustre la campagne publicitaire dont il doit discuter (figure 1). L'image est suivie d'une croix de fixation grise sur un fond noir (3,3 secondes), puis de la conversation d'une minute entre le sujet et l'un des interlocuteurs, et enfin d'un écran noir (4,6 secondes). Ceci se répète jusqu'à ce que le participant ait complété les 6 essais d'une session.

1. `https://www.furhatrobotics.com` (Al Moubayed *et al.*, 2012).

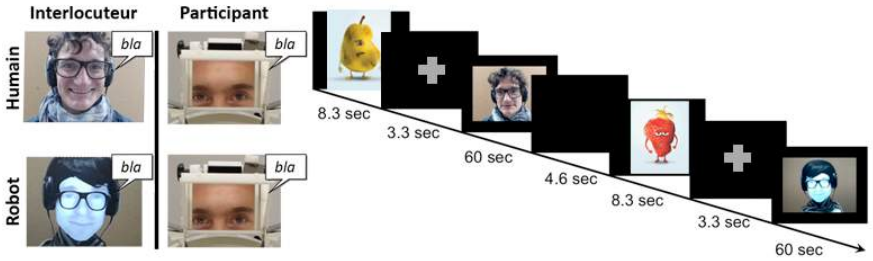


Figure 1. *Présentation des deux conditions expérimentales (participant humain et participant robot) à gauche, et organisation des essais à droite, avec les images des exemples donnés en tableaux 2 (poire) et 3 (fraise).*

Au cours de ces échanges, différentes données brutes sont acquises. Pour les analyses présentées dans cet articles, les données utilisées sont les enregistrements vocaux du participant (à l'aide d'un microphone à réduction active du bruit compatible IRM) et de l'interlocuteur, et l'activité cérébrale enregistrée par le scanner IRM.

2.3. *Préparation des données*

Les données brutes doivent être préparées pour les analyses automatiques décrites dans la suite de cette présentation. Nous nous concentrons sur les deux types de données exploitées ici. Les enregistrements audio des participants sont filtrés pour réduire les bruits du scanner qui n'ont pas été éliminés par le filtre actif du microphone. Les données débruitées des participants et celles des interlocuteurs sont ensuite segmentées en unités interpausales (UIP), qui ont été définies comme des blocs de parole délimités par des silences d'une durée minimale de 200 ms (Blache *et al.*, 2009). L'inspection visuelle du débruitage et de la segmentation a été effectuée à l'aide du logiciel de traitement de la parole oral Praat (Boersma, 2001). Les fichiers audio et les segmentations sont ensuite téléchargés dans SPPAS[2] pour transcription. Les transcriptions orthographiques des productions langagières des participants et des interlocuteurs (fichiers .TextGrid) sont réalisées manuellement sur des fichiers séparés pour les deux interlocuteurs, qui sont déposés dans l'entrepôt de données Ortolang[3]. Ce sont sur ces fichiers que nous réalisons nos analyses linguistiques. Les tableaux 2 et 3 donnent des exemples de discussions reconstruites à partir de ces transcriptions.

Le traitement des données de l'IRMf suit les procédures standard et a déjà été décrit en détail (Rauchbauer *et al.*, 2019).

2. Version 1.9.9 :`www.sppas.org/` (Bigi, 2015).
3. `https://www.ortolang.fr/market/corpora/convers/`

Nous utilisons une parcellisation cérébrale formée sur la base de données fonctionnelles et de connectivité cérébrale, de sorte que les régions d'intérêt représentent des volumes homogènes en termes de fonction (Fan *et al.*, 2016). Dans chacune des 246 régions de l'atlas, l'estimation de l'activité est extraite avec la boîte à outils Marsbar (Brett *et al.*, 2002) et l'ensemble de ces valeurs (21 participants, 24 essais et 246 régions) est utilisé pour les analyses statistiques.

3. Variables linguistiques

Variables temporelles	Variables d'oralité	Variables de complexité
Temps de parole	Feedbacks	Complexité lexicale
Durée moyenne UIP	Pauses remplies	Complexité descriptive
	Marqueurs discursifs	Complexité syntaxique

Tableau 4. *Liste des variables linguistiques étudiées.*

3.1. *Description*

Nous calculons huit variables linguistiques, réparties en trois catégories et présentées dans le tableau 4. L'objectif de ces métriques est double. Dans un premier temps, nous entendons décrire et caractériser les différences entre les productions des interlocuteurs humain et robot. Dans un second temps, nous souhaitons savoir si les participants adaptent leur production selon la production des interlocuteurs, ainsi que selon la nature de ces derniers.

Puisque les interactions enregistrées sont courtes, nous faisons le choix de travailler sur la minute entière pour chaque essai.

3.1.1. *Variables temporelles*

Afin d'étudier la production linguistique des locuteurs dans son ensemble, nous extrayons le temps de parole des locuteurs et la durée moyenne des unités interpausales (UIP) qu'ils produisent.

Le temps de parole correspond à la durée de parole d'un locuteur sur l'essai. Pour le calculer, nous sommons les durées, en secondes, de toutes les UIP du locuteur au cours de l'interaction.

$$temps\,de\,parole = \sum_{i=1}^{N_{UIP}} durée\,UIP_i \qquad [1]$$

Nous calculons la durée moyenne des interventions en divisant le temps de parole du locuteur sur l'essai par le nombre N_{UIP} d'UIP produites.

$$durée\ moyenne\ UIP = \frac{1}{N_{UIP}} \sum_{i=1}^{N_{UIP}} durée\ UIP_i \quad [2]$$

3.1.2. *Variables d'oralité*

Nous réalisons un travail sur les items lexicaux de feedback, sur les pauses remplies et sur les marqueurs discursifs (voir tableau 5 pour des exemples d'items lexicaux les plus fréquemment utilisés dans notre corpus).

Le feedback est un mécanisme linguistique qui permet aux locuteurs en interaction d'échanger des informations sur le processus de communication lui-même, c'est-à-dire sur la perception et la compréhension mutuelles (Allwood *et al.*, 1992 ; Bunt, 1994) et plus globalement sur la gestion du *Common Ground* (Clark *et al.*, 1983). Il est par exemple fréquent lors d'une interaction entre deux locuteurs que l'individu en position d'écoute produise de courts énoncés comme *« oui », « d'accord », « ok »* ou acquiesce d'un mouvement de tête, pour ratifier le discours de celui qui parle et pour signaler à ce dernier la compréhension de ses paroles. La variable que nous calculons réalise un ratio du nombre d'UIP du locuteur commençant par un item lexical de feedback sur le nombre d'UIP prononcés par le locuteur sur la minute.

$$ratio\ feedback = \frac{1}{N_{UIP}} \sum_{i=1}^{N_{UIP}} \chi_F(w_0^i) \quad [3]$$

où w_0^i représente le premier mot de la $i^{ème}$ UIP et χ_Q est la fonction caractéristique de l'ensemble Q. Soit, pour l'ensemble F des marqueurs de feedback :

$$\chi_F(w) = \begin{cases} 1\ si\ w\ est\ un\ marqueur\ de\ feedback\ (w \in F) \\ 0\ sinon \end{cases}.$$

Les pauses remplies découlent des difficultés auxquelles se confrontent les locuteurs lorsqu'ils recherchent un mot, lorsqu'ils ont besoin de temps pour construire leur phrase ou encore quand ils ne savent pas quoi dire. Ces éléments, que l'on peut qualifier de disfluences ou de feedbacks selon le contexte, perturbent le déroulement interactionnel et temporel du discours (Shriberg, 1994 ; Henry et Pallaud, 2003 ; Baiocchi, 2015). La variable que nous calculons réalise un ratio du nombre de marqueurs de pauses remplies (PR) produites par le locuteur par le nombre total n de mots prononcés (W) par le locuteur pendant la minute.

$$ratio\ pauses\ remplies = \frac{1}{n} \sum_{w \in W} \chi_{PR}(w) \quad [4]$$

La dernière analyse concerne les marqueurs discursifs. Ces unités linguistiques lient des propositions syntaxiques entre elles, permettant ainsi de marquer la relation entre les unités du discours (Schiffrin, 1987 ; Roze, 2009). Un locuteur qui souhaite produire un discours structuré va ainsi utiliser des connecteurs comme « mais », « parce

que », « et ». Pour calculer la variable, nous calculons le ratio du nombre de mots qui, parmi tous les mots prononcés par le locuteur pendant la minute, sont des marqueurs discursifs (MD).

$$ratio\, marqueurs\, discursifs = \frac{1}{n} \sum_{w \in W} \chi_{MD}(w) \quad [5]$$

Feedbacks	oui, ouais, non, ok, voilà, d'accord, mh, *rire*
Pauses remplies	euh, heu, mh, hum
Marqueurs discursifs	alors, mais, et, puis, enfin, parce que, parce qu', ensuite

Tableau 5. *Liste non exhaustive des marqueurs linguistiques de l'oral analysés dans notre corpus.*

3.1.3. *Variables de complexité*

Pour extraire les trois variables de complexité, nous utilisons l'outil d'enrichissement de données textuelles MarsaTag (Rauzy *et al.*, 2014). Entraîné sur des corpus français et oraux, l'outil nous permet de réaliser automatiquement la tokenisation, l'étiquetage morphosyntaxique et la lemmatisation de l'ensemble des conversations. La préférence de MarsaTag sur Spacy, bibliothèque Python plus communément utilisée pour ces tâches, résulte de la comparaison que nous avons pu faire des performances de ces deux outils : en effet, Spacy n'étant pas entraîné sur de l'oral, l'étiquetage morphosyntaxique, en particulier, est bien moins performant sur notre corpus. Le tableau 6 présente les performances des deux outils, en se focalisant sur les mots en « a » étiquetés « adjectif » par au moins un des deux outils. Sur l'intégralité du corpus, seuls 3 des 57 mots repérés par MarsaTag comme adjectifs sont mal étiquetés. En revanche, sur 53 tokens en « a » étiquetés adjectifs par Spacy, 15 d'entre eux n'en sont pas. On remarque également que MarsaTag repère 4 adjectifs de plus que Spacy, et que parmi les 57 adjectifs repérés par MarsaTag, 17 n'ont pas été repérés par Spacy (contre 3 repérés par Spacy mais pas MarsaTag).

Cet étiquetage morphosyntaxique nous permet de calculer nos variables de complexité. Estimant que le robot produit des énoncés linguistiques peu complexes lexicalement et syntaxiquement (tableau 1), nous commençons par créer une variable de complexité lexicale qui correspond à la fraction du nombre de mots de contenu (noms + adjectifs + verbes, à l'exception des verbes auxiliaires, semi-auxiliaires et d'état) produits par le locuteur sur le nombre total n de mots prononcés par le locuteur en une minute.

$$complexité\, lexicale = \frac{1}{n} \sum_{w \in W} \mathbb{1}_{adj}(w) + \mathbb{1}_{nom}(w) + \mathbb{1}_{vb\, act}(w) \quad [6]$$

Problème	MarsaTag	Spacy
tokens en « a » faussement étiquetés en adjectif	allo, aubergine, avengers	achète, agissais, ah, allais, allez, allo, amérique, annonces, apprécier, attends, au, auquel, aux, auxquels, avais
adjectifs en « a » étiquetés par un seul des outils	abandonnés, abîmés, acide, affaiblies, affaissée, aimée, allongé, ambigu, ambivalent, amusée, américaines, animé, arrondi, arrondis aseptisés, attachante, attaqué	abîmé, accroché, asiatique

Tableau 6. *Problèmes rencontrés lors de l'étiquetage des tokens en « a » par les outils MarsaTag et Spacy. Les quatre listes de tokens sont exhaustives.*

où $\mathbb{1}_{adj}(w) = \begin{cases} 1\ si\ w\ est\ un\ adjectif \\ 0\ sinon \end{cases}$.

La complexité descriptive, dont la formule est tirée de Ochs *et al.* (2018), correspond au ratio du nombre d'adjectifs et d'adverbes prononcés sur le nombre total de mots n prononcés par le locuteur pendant l'essai.

$$complexité\ descriptive = \frac{1}{n} \sum_{w \in W} \mathbb{1}_{adv}(w) + \mathbb{1}_{adj}(w) \qquad [7]$$

Également inspirée de Ochs *et al.* (2018), nous calculons une variable de complexité syntaxique en divisant le nombre de pronoms, de prépositions et de conjonctions produits par le locuteur par le nombre total de mots n qu'il prononce au cours de l'interaction.

$$complexité\ syntaxique = \frac{1}{n} \sum_{w \in W} \mathbb{1}_{pron}(w) + \mathbb{1}_{prep}(w) + \mathbb{1}_{conj}(w) \qquad [8]$$

3.2. *Comparaison des productions des deux interlocuteurs*

Un modèle linéaire mixte a été employé pour l'analyse de variance sur l'ensemble des variables linguistiques pour les productions de l'interlocuteur. L'intérêt de cette analyse ne porte pas sur l'interlocuteur à proprement parler, car nous savons que les productions sont différentes ; elle permet néanmoins de caractériser et de quantifier ces différences. Nous nous sommes surtout intéressés au facteur temps, représenté par la variable *Essai*, avec l'hypothèse que l'interlocuteur humain s'adapte au participant au fur et à mesure que se construit une familiarité, alors que le robot ne peut pas s'adapter.

Ces analyses ont été produites dans R (R Core Team, 2013) avec le paquet lme4. L'avantage du modèle linéaire mixte, comparativement au modèle linéaire classique, est la prise en compte de la variabilité liée à différents facteurs non contrôlés (par exemple les participants). La sélection de la structure du modèle a été estimée par le maximum de probabilités restreintes (*restricted maximum likelihood*) donnant une solution intégrant la nature de l'interlocuteur *(Interlocuteur)*, des douze essais successifs *(Essai)*, et l'interaction d'intérêt *Interlocuteur* × *Essai* en effets fixes. À cela, une variable aléatoire prenant en compte la variabilité des participants selon les essais *(Essai|Participant)* a été introduite dans le modèle. Cette variable aléatoire permet de prendre en compte la variabilité inter-essais et inter-participants et donc de fournir une meilleure estimation des effets fixes.

$$variable\, interlocuteur \sim Interlocuteur * Essai + (1 + Essai|Participant) \quad [9]$$

Les résultats des analyses statistiques (tests de Fischer) sont donnés dans le tableau 7. Comme attendu, l'*Interlocuteur* a un effet très significatif sur toutes les variables : à cause du contrôle par le système de magicien d'Oz, l'humain et le robot ont des comportements linguistiques différents. Il est probable que des systèmes de contrôle plus élaborés donneraient des effets différents. Nos résultats sont spécifiques à cette implémentation du système de magicien d'Oz. Pour les variables temporelles, l'humain parle plus longtemps, ce qui est en partie causé par le délai introduit par le magicien d'Oz. Il produit aussi des UIP plus longues en moyenne que le robot.

Pour les variables d'oralité, l'humain produit des ratios plus importants de pauses remplies, de feedbacks et de marqueurs discursifs que le robot. Pour les variables de complexité, le robot obtient des moyennes plus importantes que l'humain pour la complexité descriptive et lexicale, mais il est moins élevé pour la complexité syntaxique. Ceci s'explique aussi par des particularités du système de magicien d'Oz, en l'occurrence ses phrases scriptées se rapprochent du langage écrit et évitent donc les disfluences. Les disfluences de l'interlocuteur humain augmentent le nombre de mots comptabilisés au dénominateur de ces variables et diminuent la proportion de signifiants par rapport au nombre total de mots. En corollaire, cette oralité de l'humain implique des structures de phrases plus complexes comme l'indique l'augmentation de la complexité syntaxique. En conclusion, les résultats confirment les hypothèses : le robot n'ayant à disposition que des phrases simples et pas de marqueurs de l'oralité, alors que l'humain parle librement, il a un contenu plus riche, proportionnellement, en termes de lexique mais moins élaboré en termes de syntaxe.

On notera un effet très significatif du temps (facteur *Essai*) et de l'interaction *Interlocuteur* × *Essai* pour les marqueurs discursifs. La Figure 2 indique qu'ils augmentent pour l'humain et diminuent pour le robot au cours du temps. Pour les trois variables de complexité linguistique, le facteur *Essai* et l'interaction *Interlocuteur* × *Essai* sont significatifs. La complexité descriptive augmente au cours du temps pour le robot et diminue pour l'humain ($p < 0{,}001$, figure 2). Des dynamiques différentes apparaissent pour les autres variables de complexité ($p < 0{,}05$), avec une diminution au cours du

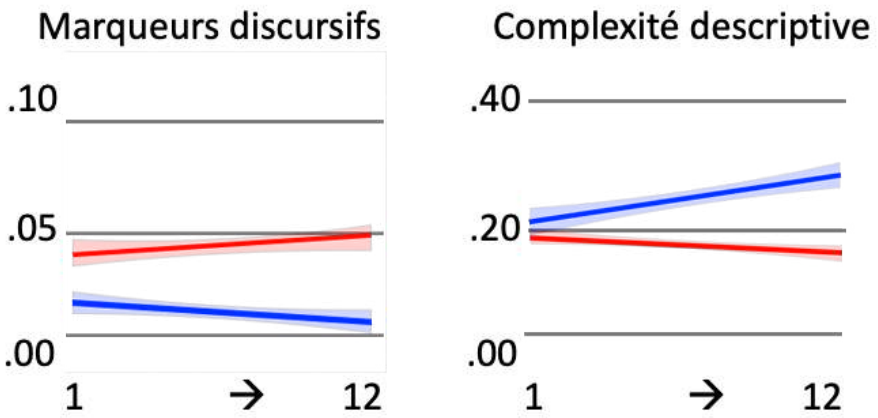


Figure 2. *Évolution des ratio au cours des* Essais *(interlocuteur humain en rouge et robot en bleu)*

temps pour le robot mais pas l'humain pour la complexité lexicale et une augmentation au cours du temps pour l'humain mais pas le robot pour la complexité syntaxique. Aucun autre effet n'impliquant le facteur *Essai* n'est significatif pour les deux autres variables d'oralité, feedbacks et pauses remplies, et pour les variables temporelles. Ces résultats avec le facteur *Essai* impliquent que ces variables de complexité linguistique évoluent au cours du temps différemment selon l'interlocuteur. Une observation des expérimentateurs lors du recueil du corpus pourrait permettre de proposer une explication pour ces résultats. En effet, et sans que cela soit formalisé, en début d'expérience, les conversations concernent surtout la description des images, tandis qu'à la fin, elles portent plus sur le message de la campagne de publicité. En effet, au début les participants découvrent les images et décrivent ce qu'ils perçoivent avec l'interlocuteur, tandis qu'à la fin ils tentent d'accomplir l'objectif qui leur a été donné, à savoir trouver le message de la campagne publicitaire.

Pour l'humain, l'augmentation des marqueurs discursifs et de la complexité syntaxique et la diminution de la complexité descriptive peuvent s'expliquer par cette transition d'une phase descriptive vers une phase argumentative. Les effets opposés pour le robot (diminution des marqueurs discursifs et de la complexité lexicale, et augmentation de la complexité descriptive) s'expliquent par les phrases pré-enregistrées du système de magicien d'Oz.

	Interlocuteur		**Essai**		**Interlocuteur × Essai**	
Variable	**F**	***p***	**F**	***p***	**F**	***p***
Temps de parole	705,537	**0,000**	0,965	0,478	0,493	0,908
Durée moyenne des UIP	217,297	**0,000**	1,552	0,110	1,354	0,192
Feedbacks	114,994	**0,000**	0,794	0,646	0,712	0,728
Pauses remplies	618,650	**0,000**	1,057	0,395	1,800	0,051
Marqueurs discursifs	211,748	**0,000**	3,431	**0,000**	6,533	**0,000**
Complexité lexical	124,758	**0,000**	2,164	*0,015*	1,900	*0,037*
Complexité descriptive	136,494	**0,000**	4,820	**0,000**	3,434	**0,000**
Complexité syntaxique	148,811	**0,000**	2,344	*0,008*	2,332	*0,008*

Tableau 7. *Résultats des analyses statistiques (résultat du test de Fisher avec la formule [9]) sur les variables linguistiques pour l'interlocuteur, en gras les effets significatifs à $p < 0{,}001$ et en italique à $p < 0{,}05$.*

3.3. *Analyses des effets du comportement des interlocuteurs sur le comportement des participants*

Avec ces analyses en modèle linéaire mixte, dont les résultats (tests de Student) sont donnés tableau 8, nous nous interrogeons sur les effets que les variables mesurées pour l'interlocuteur ont sur les mêmes variables mesurées chez le participant, éventuellement en interaction avec la nature de l'interlocuteur. Les résultats doivent se comprendre comme il suit : un effet significatif de la variable de l'interlocuteur sur la variable du participant s'interprète comme l'influence automatique du comportement de l'interlocuteur sur le comportement du participant, alors qu'un effet significatif du facteur *Interlocuteur* s'interprète comme la conséquence de la différence de posture intentionnelle adoptée par le participant en fonction de la nature, humain ou robot, de l'interlocuteur. Une interaction significative signale que les effets « ascendants », ceux de la variable de l'interlocuteur, sont différents selon la nature de l'interlocuteur, c'est-à-dire qu'ils dépendent d'un contrôle « descendant ». Le facteur *Essai* a aussi été inclus dans le modèle (sans interaction avec les autres facteurs) mais ses résultats ne sont pas indiqués car pas discutés ultérieurement, ainsi que la variable aléatoire *(Essai|Participant)* décrite précédemment. La formule complète utilisée dans R pour ces analyses est :

$$variable\ participant \sim$$
$$variable\ interlocuteur * Interlocuteur + Essai + (1 + Essai|Participant) \quad [10]$$

	Var. Inter^{eur}		**Interlocuteur**		**Inter. × Var. Inter^{eur}**	
Var. Part^{nt}	**t**	***p***	**t**	***p***	**t**	***p***
Temps de parole	– 13,184	**0,000**	– 4,708	**0,000**	– 3,633	**0,000**
Durée moyenne UIP	– 1,558	0,119	– 2,556	*0,011*	0,770	0,441
Feedbacks	– 1,838	0,066	– 4,675	**0,000**	– 0,013	0,990
Pauses remplies	0,738	0,461	2,505	*0,012*	– 0,450	0,653
Marqueurs discursifs	– 0,491	0,623	0,656	0,512	0,357	0,721
Complexité lexicale	1,126	0,260	1,238	0,216	– 0,895	0,371
Complexité descriptive	2,357	*0,018*	0,598	0,550	– 0,594	0,552
Complexité syntaxique	2,285	*0,022*	1,881	0,060	– 1,581	0,114

Tableau 8. *Résultats des analyses statistiques (test de Student) sur les variables linguistiques du participant, en gras les effets significatifs à $p < 0{,}001$ et en italique à $p < 0{,}05$,*

3.3.1. *Variables temporelles*

Nous observons un effet significatif de la variable de l'interlocuteur, du facteur *Interlocuteur* et de l'interaction entre les deux facteurs sur le temps de parole des participants. Cela signifie que le temps de parole des participants est influencé par le temps de parole des interlocuteurs et par leur nature. On observe, en l'occurrence, que plus l'interlocuteur parle, moins le participant parle (et inversement), ce qui s'explique de manière triviale par le fait que les locuteurs se partagent un temps limité de 1 minute. Cette anticorrélation est plus faible pour le robot que pour l'humain, ce qui peut s'expliquer soit par des particularités du système de magicien d'Oz (le robot présente des latences entre la sélection et la production de la réponse, cela conduit à des délais plus longs entre les prises de parole pour l'interlocuteur robot), soit par une différence de posture intentionnelle envers le robot qui réduit la propension à discuter. Il est possible que les deux effets interviennent. Concernant la durée moyenne des UIP, nous n'observons qu'un effet de l'*Interlocuteur* : les participants produisent des UIP plus longues avec l'humain plutôt qu'avec le robot, ce qui va dans le sens de l'adoption d'une posture différente selon la nature de l'interlocuteur avec qui ils discutent, en l'occurrence en simplifiant leur discours avec l'agent artificiel.

3.3.2. *Variables d'oralité*

Les seuls effets observables pour ces variables sont un effet de l'*Interlocuteur* pour les feedbacks et pour les pauses remplies, indiquant ici aussi que les sujets adaptent leur comportement selon la nature de leur interlocuteur. Les t-values montrent que les participants produisent plus de feedbacks avec l'humain qu'avec le robot, et plus de pauses remplies avec le robot qu'avec l'humain. Nous n'observons pas d'effet de la variable de l'interlocuteur sur la variable du participant, et pas non plus d'effet de l'interaction. La production de marqueurs linguistiques du participant n'est donc pas influencée par la production de feedbacks, de pauses remplies et de marqueurs

discursifs de l'interlocuteur. D'un point de vue linguistique, ces résultats ne sont pas étonnants :

– un locuteur ne produit pas des feedbacks parce que son interlocuteur en produit, il produit des feedbacks pour signifier que l'énoncé de l'interlocuteur a été perçu et compris,

– il ne produit pas des pauses remplies parce que l'interlocuteur en produit, mais parce que le discours de l'interlocuteur est décousu et imprévisible (ou simplement parce qu'il cherche ses mots mais souhaite faire comprendre qu'il ne veut pas perdre le tour de parole),

– et un locuteur ne produit pas des marqueurs discursifs parce que l'interlocuteur en produit, il produit des marqueurs discursifs pour établir des relations entre les parties de son discours afin que sa parole ait du sens.

3.3.3. *Variables de complexité*

Nous observons un effet de la variable des interlocuteurs pour la complexité syntaxique et la complexité descriptive, mais pas d'effet de l'*Interlocuteur* ou d'interaction. Cela indique que la quantité de connecteurs structurants (complexité syntaxique) et la quantité d'adjectifs et d'adverbes (complexité descriptive) des participants et des interlocuteurs sont corrélées indépendamment de la nature des interlocuteurs, en faveur d'une interprétation de type convergence ou alignement entre interlocuteurs. Pour la complexité lexicale, nous n'observons aucun effet de la variable de l'interlocuteur, du facteur *Interlocuteur*, ou de l'interaction. Les participants n'adaptent pas la quantité de mots de contenu qu'ils prononcent en fonction de la quantité de mots de contenu prononcés par les interlocuteurs, et produisent autant de mots de contenu avec l'interlocuteur naturel qu'avec l'interlocuteur artificiel. Ainsi, contrairement à nos hypothèses, il n'y aurait pas d'influence de l'interlocuteur sur le participant quant au ratio de mots de contenu prononcés. Cependant, et bien que ces résultats n'indiquent pas de corrélation entre la quantité de mots de contenu prononcés par les participants et par les interlocuteurs, nous décidons de nous focaliser sur le lexique en présentant dans la section suivante une variable d'alignement lexical. Cette variable nous permet d'explorer plus en détail le vocabulaire des locuteurs, en s'interrogeant spécifiquement sur les mots de contenu employés par un locuteur et repris par l'autre.

4. Variables d'alignement

4.1. *Description*

Puisque nous nous intéressons à l'alignement conversationnel présent dans notre corpus, et en particulier à l'alignement lexical, nous travaillons sur une variable d'alignement lexical à partir de laquelle nous étudions le lexique commun entre les locuteurs. Nous nous inspirons de la formule LILLA *(lexical indiscriminate local linguistic alignment)* proposée par Fusaroli *et al.* (2012) et reprise par Xu et Reitter (2015). Afin qu'on puisse calculer l'alignement lexical entre les locuteurs en interaction, cette

formule repose sur le principe de l'effet d'amorçage, un principe décrit par Pickering et Garrod (2004) dans leur modèle d'alignement interactif selon lequel la production langagière des individus est directement influencée par les stimulations auditives auxquelles ils sont exposés. Concrètement, cela signifie que sur le plan lexical, les participants devraient tendre à utiliser le vocabulaire des interlocuteurs après que ces derniers ont introduit de nouveaux mots dans la conversation. LILLA étant à l'origine une mesure normalisée du nombre de mots qui apparaissent dans un texte (*prime*) et qui sont repris dans une réponse (*cible*), la formule ne correspond pas exactement à notre situation. D'une part, nous souhaitons étudier l'alignement lexical des participants sur les interlocuteurs mais également l'alignement lexical des interlocuteurs sur les participants. Les locuteurs alternant les prises de parole au cours du dialogue, il se peut qu'un mot présent dans la conversation du *prime* ait été introduit précédemment par une *cible*. Dans ce cas-là, nous adaptons alors le vocabulaire utilisé comme référence pour le *prime* (P) afin ne pas prendre en compte les mots précédemment introduits par la *cible* (T). D'autre part, nous ne nous intéressons qu'aux mots de contenu (w) plutôt qu'à tout le vocabulaire, et nous utilisons ainsi les mêmes mots que ceux analysés dans la variable de complexité lexicale.

En incluant ces réflexions, on obtient une formule comptabilisant, sur toutes les interventions d'un *prime*, le nombre des mots qu'il introduit dans la conversation et qui sont réutilisés par la *cible* dans la suite de l'échange. Afin d'obtenir une mesure entre 0 et 1, ce chiffre est normalisé par le produit du nombre de mots distincts prononcés respectivement par *cible* et *prime* (ne sont pas considérés les mots introduits par *cible* et répétés par *prime*). On peut donc formaliser le calcul de LILLA de manière mathématique :

$$LILLA(T,P) \;=\; \frac{\sum_{i=0}^{L} \sum_{w \in P_i \setminus \bigcup_{j>i} T_j} \delta(w, \bigcup_{j>i} T_j)}{\#(\bigcup_{i=0}^{L}(P_i \setminus \bigcup_{j\,<\,i} T_j)) * \#(\bigcup_{j=0}^{L} T_j)} \qquad [11]$$

avec $\delta(w,X) \;=\; \begin{cases} 1\ si\ w \in X \\ 0\ sinon \end{cases}$.

P_i (respectivement T_j) désigne l'ensemble des mots de la $i^{ème}$ intervention du *prime* (respectivement *cible*). Si i n'est pas une intervention du prime, alors $P_i = \{\}$. $\bigcup T_j$ permet de regarder l'ensemble des mots prononcés par un participant, soit sur toute la conversation ($\bigcup_{j=0}^{L} T_j$), soit de manière plus ciblée, en ne regardant que les mots prononcés avant ($\bigcup_{j\,<\,i} T_j$) une intervention i donnée, soit prononcés après ($\bigcup_{j>i} T_j$). Le tour de parole 0 désignant celui d'un participant (qu'il soit ou non pris) ; L dénote le dernier tour de parole. Pour permettre le calcul, les UIP consécutives d'un même locuteur doivent être concaténées pour n'en faire qu'une (voir tableau 9 pour un exemple). Après cette transformation, les tours pairs désignent ceux du participant, ceux impairs ceux de l'interlocuteur. L'alignement du participant sur l'interlocuteur est donc obtenu en ne considérant que les j pairs et i impairs ; l'alignement de l'interlocuteur sur le participant est obtenu pour i pair et j impair.

Participant : euh c'est un **citron vert**
Interlocuteur : ouais et un **citron vert** avec aussi un un masque dég- euh dé- découpé euh sur le sur le zeste là en enlevant le le zeste autour des yeux
Participant : euh moi j'ai pas compris
Interlocuteur : j'ai l'impression qu'il y avait encore deux yeux
Participant : c'est le c'est genre **Tortue Ninja** ou quoi cette fois-ci
Interlocuteur : ouais et ben ouais c'est ce que j'allais dire c'est ce que j'allais dire c'était le bandeau là ils ont pas découpé les yeux ils ont découpé un truc autour des yeux
Participant : ouais
Interlocuteur : comme un bandeau
Participant : euh ça fait **Tortue Ninja**
Interlocuteur : donc après les **Tortues Ninja** donc l'aubergine Batman et le et le citron **Tortue Ninja** euh ça fait un truc euh *légumes* et super héros
Participant : c'est ça *légumes* et fruits en effet donc plus pour les **enfants** obligé
Interlocuteur : ah ouais et pour les **enfants** ouais ouais ça ça me paraît pour les **enfants** ça brille bien
Participant : ou pour les grands **enfants** ouais
Interlocuteur : ben ouais c'est vrai que ça c'est un peu intergénérationnel parce que les **Tortue Ninja** euh

Tableau 9. *Transcription condensée d'un échange pour faciliter le calcul de LILLA : les interventions participant et interlocuteur sont alternées. Sont surlignés en gras les mots de contenu introduits par le participant et répétés, en italique ceux introduits par l'interlocuteur et répétés. Au total, l'algorithme compte 17 mots de contenu introduits par l'interlocuteur parmi lesquels un seul token est repris par le participant ; 10 mots de contenu sont introduits par le participant, 5 de ces tokens sont repris par l'interlocuteur. L'alignement LILLA du participant sur l'interlocuteur donne donc 0,00588 (assez faible), tandis que celui de l'interlocuteur sur le participant est de 0,02066.*

En accord avec les travaux de Brennan (1996) et Branigan *et al.* (2011) selon lesquels l'alignement est plus fort lorsque l'on s'adresse à un robot, notre hypothèse initiale est celle d'un alignement lexical plus fort du participant sur l'interlocuteur robot que sur l'interlocuteur humain. Ensuite, bien que la conversation avec le robot soit bidirectionnelle, elle n'est dans les faits qu'unidirectionnelle pour l'alignement lexical. En effet, le robot ayant un discours pré-enregistré et un lexique fini, il peut converser avec le participant mais il ne peut pas aligner son lexique si les mots de contenu employés par le participant ne font pas partie du vocabulaire du système de magicien d'Oz. Ainsi, dans la direction d'alignement lexical de l'interlocuteur sur le participant, l'hypothèse est celle d'un alignement lexical plus fort de l'interlocuteur humain que de l'interlocuteur robot sur le participant.

4.2. *Analyse statistique*

Le tableau 10 indique les résultats des analyses statistiques de LILLA selon l'équation 9. Sur la première ligne, le *prime* est l'interlocuteur et la *cible* le participant, c'est-à-dire que la variable mesure l'alignement lexical du participant sur l'interlocuteur. On observe un effet significatif du facteur *Interlocuteur*, et la valeur positive du test t indique qu'elle est plus grande lorsque l'interlocuteur est le robot. Comme attendu, le participant s'aligne donc plus sur le lexique du robot que sur le lexique de l'humain.

La seconde ligne indique les résultats lorsque le *prime* est le participant et la *cible* l'interlocuteur, c'est-à-dire l'alignement lexical de l'interlocuteur sur le participant. On observe aussi un effet significatif du facteur *Interlocuteur*, et la valeur positive du test t indique qu'elle est plus grande lorsque l'interlocuteur est le robot que l'humain. Contrairement à nos hypothèses, l'interlocuteur robot s'aligne plus sur le lexique du participant que l'interlocuteur humain.

	Interlocuteur		**Essai**		**Interlocuteur × Essai**	
Prime	**t**	***p***	**t**	***p***	**t**	***p***
Interlocuteur	3.661	**0,000**	– 1,846	0,065	1,400	0,162
Participant	2,835	*0,005*	– 2,097	*0,036*	– 1,679	0,093

Tableau 10. *Résultats de l'analyse statistique sur LILLA, en gras les effets significatifs à p < 0,001, en italique les effets significatifs à p < 0,05.*

5. Variables neurophysiologiques

L'objectif principal de cet article est l'application des outils d'analyse de traitement automatique à la description de notre corpus. Ce traitement automatique produit des variables numériques qui caractérisent différents aspects de chaque conversation. Elles peuvent être utilisées pour identifier leurs corrélats cérébraux. En pratique, le cerveau a été parcellisé en 247 régions, dont nous avons extrait l'activité moyenne pour chaque conversation. La formule suivante est utilisée pour identifier les régions dont l'activité est modulée par l'interlocuteur, corrélée à la variable, ou significativement associée à l'interaction entre la variable et l'interlocuteur :

$$région_n \sim variable * Interlocuteur + Essai + (1 + Essai|Participant) \quad [12]$$

À noter que, comme précédemment, la variable *Essai* est introduite pour capturer une éventuelle évolution temporelle du signal, mais ne sera pas décrite dans les résultats. L'important est d'identifier les corrélats cérébraux de variables comporte-

mentales et, éventuellement, comment ils sont affectés par la nature de l'interlocuteur (interaction entre chaque variable comportementale et *Interlocuteur*)[4].

5.1. *Analyse du temps de parole*

L'analyse du temps de parole du participant et de l'interlocuteur correspond, respectivement, à la quantité de paroles produite et à la quantité de paroles perçue par le cerveau. L'objectif de cette première analyse est de valider l'approche utilisée avec des variables dont les corrélats cérébraux sont bien connus. On s'attend donc à une corrélation avec les systèmes de production langagière pour le premier (cortex moteur et gyrus frontal inférieur gauche) et de perception pour le second (lobe temporal).

Les résultats sont donnés à la figure 3 pour le temps de parole du participant et à la figure 4 pour le temps de parole de l'interlocuteur. La couleur correspond à la direction des corrélations, négatives en bleu et positives en rouge. On remarque donc la prédominance du lobe temporal (centré sur les gyri temporaux moyens et supérieurs dans les deux hémisphères), la principale région du cerveau humain pour la compréhension du langage, qui corrèle négativement avec le temps de parole du participant (en bleu dans 3) et positivement avec celui de l'interlocuteur (en rouge dans 4).

Un autre résultat est la corrélation entre le temps de parole de l'interlocuteur et l'activité dans l'aire de Broca, au fond du sillon frontal inférieur dans les deux hémisphères. Ce résultat va dans le sens de l'implication de l'aire de Broca, connue pour son rôle dans la production verbale, dans la perception du langage.

5.2. *Analyse LILLA*

Les résultats précédents valident l'approche utilisée pour identifier les corrélats cérébraux des variables issues des analyses précédentes. Nous l'avons donc utilisée pour valider la variable développée dans le cadre de cette étude, LILLA. L'identification de corrélats cérébraux associés à cette variable suggère qu'elle représente un aspect pertinent du comportement langagier. Une région du gyrus parahippocampique gauche, mais surtout trois régions adjacentes au niveau du gyrus cingulaire central gauche sont identifiées comme ayant une activité significativement négativement corrélée à la variable LILLA. À noter que l'interaction entre *LILLA* et *Interlocuteur* n'identifie aucune région au seuil utilisé. Ces régions du système limbique ne sont pas connues pour leur implication dans le langage, mais la continuité anatomique implique donc

4. L'effet principal du facteur *Interlocuteur* est difficile à interpréter : étant donné que pour chaque variable comportementale, l'activité de chaque région et l'étiquetage humain ou robot de chaque essai restent identiques, nous devrions toujours obtenir le même résultat. Or, ce n'est pas le cas. En effet, une partie de la différence de réponse cérébrale entre interlocuteurs humain et robot s'explique par des différences de comportements entre ces deux interlocuteurs. Ainsi, si la variable décrit un comportement très différent entre humain et robot, elle capture une partie des différences associées au facteur *Interlocuteur*, rendant difficile l'interprétation des résultats.

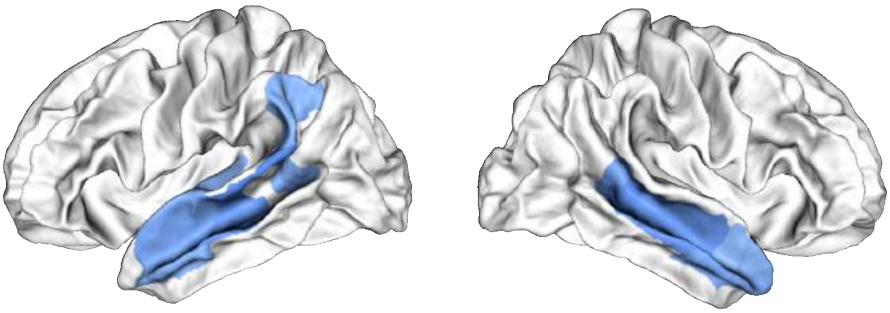


Figure 3. *Régions corticales corrélées avec le temps de parole total du participant*

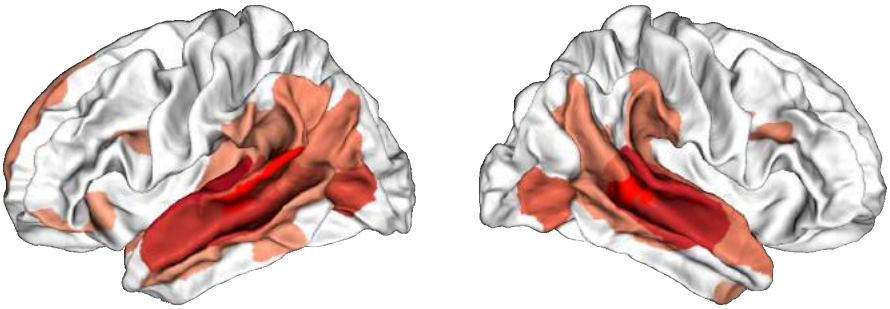


Figure 4. *Régions corticales corrélées avec le temps de parole total de l'interlocuteur*

que leur corrélation significative conjointe représente une implication fonctionnelle de cette région du cerveau. L'hippocampe est surtout impliqué dans les processus mnésiques, le gyrus cingulaire central est impliqué dans le contrôle des actions. L'augmentation de LILLA signifiant une utilisation plus importante des mots de l'interlocuteur, la corrélation négative avec l'activité dans ces régions suggère une utilisation réduite des ressources cognitives propres des participants (contrôle des actions et mémoire) lorsque les mots utilisés sont fournis par les interlocuteurs.

6. Discussion

Cet article présente l'étude, avec des outils de TAL, d'un corpus unique combinant des conversations comparables avec un humain ou un robot, synchronisées avec l'enregistrement de l'activité cérébrale en IRM fonctionnelle. Trois séries de résultats ont été décrits : (i) caractériser les différences, connues, entre les productions linguistiques des interlocuteurs humain et robot, ainsi que leur modification au cours du temps, (ii) mettre en évidence des relations entre les productions des participants et

des interlocuteurs, (iii) développer une variable d'alignement lexical pour ce corpus et valider sa pertinence en étudiant ses corrélats cérébraux.

Pour le premier point, il s'agit de vérifier et de quantifier des différences connues dans les productions des interlocuteurs humain et robot, ainsi que leur éventuelle évolution au cours du temps. Il est important de noter que si ces différences trouvent leur origine dans des considérations techniques, elles ne sont pas pour autant des défauts de l'expérience à proprement parler. En montrant les capacités limitées du robot, elles servent à conforter la croyance du sujet dans son autonomie, alors qu'il est contrôlé par l'humain avec lequel il interagit vraiment. Les différences se retrouvent à différents niveaux : le robot parle moins et produit des UIP plus courtes que l'humain. Le robot a moins de marqueurs d'oralité tels que les pauses remplies, les feedbacks et les marqueurs discursifs que l'humain, ce qui souligne le manque de spontanéité de son discours. Quant aux marqueurs de complexité, il est significatif que leur évolution au cours du temps amplifie les différences observées dans l'effet principal. Concernant le lexique (complexité lexicale et descriptive), les réponses scriptées du robot suppriment les disfluences. La même raison explique une complexité syntaxique plus importante pour l'humain, qui peut produire à volonté des structures grammaticales emboîtées complexes que le robot n'a pas à disposition, et son augmentation au cours du temps peut être associée au passage d'une conversation décrivant les images à une conversation argumentant sur le message de la campagne de publicité.

Le deuxième objectif de ces analyses est de caractériser les relations entre participants et interlocuteurs. En particulier, nous voulons utiliser le fait que les interlocuteurs ont des productions différentes pour distinguer les effets automatiques de convergences de ceux liés à la nature de l'interlocuteur artificiel ou humain. Nous nous plaçons dans le cadre théorique de la posture intentionnelle du philosophe Dennett (1987), qui postule que nous modifions notre comportement en fonction de la posture intentionnelle que nous adoptons selon la nature de l'interlocuteur. Les modèles linéaires évaluent quel facteur, soit la production, soit la nature de l'interlocuteur, explique la production du participant. Ils permettent aussi de mettre en évidence d'éventuelles interactions entre les deux. Après avoir vérifié la validité de l'approche avec une variable aux résultats connus, le temps de parole, les variables d'oralité et de complexité sont analysées à leur tour. Pour les variables d'oralité, c'est le facteur *Interlocuteur* qui a des effets significatifs (sur le ratio de feedbacks et de pauses remplies), indiquant que les participants adaptent leur comportement oral quand ils interagissent avec un robot, en accord avec le cadre théorique de la posture intentionnelle. Cette adaptation prend la forme d'une réduction des feedbacks et d'une augmentation des pauses remplies avec la machine. Mais la différence peut aussi s'expliquer par une augmentation des disfluences chez les participants dues aux limites des productions verbales du robot. Pour les complexités syntaxiques et descriptives, on a au contraire un effet de la production de l'interlocuteur et pas d'effet du facteur *Interlocuteur*. Ceci indique que pour ces variables, les alignements entre les interlocuteurs se font localement, au niveau de l'essai, et ne dépendent pas de la nature intentionnelle de l'interlocuteur.

L'absence de corrélation entre la quantité de mots signifiants prononcés par le participant et l'interlocuteur nous a conduits au troisième objectif, développer une nouvelle variable basée sur des études antérieures d'alignement lexical. La variable LILLA indique une différence significative d'alignement entre le participant et l'interlocuteur, avec un alignement plus important avec le robot qu'avec l'humain. Ainsi, le participant emploie plus de mots signifiants introduits par le robot que par l'humain. En accord avec la littérature existante, cela confirme que les humains s'alignent davantage avec les interlocuteurs artificiels sur le plan lexical pour prendre en compte leurs capacités limitées : utiliser le même mot que le robot permet de s'assurer qu'il connaît ce mot et est donc capable de le comprendre une contrainte pratiquement inversée avec l'humain, où des champs sémantiques permettent d'élargir le socle lexical commun.

La variable LILLA met également en avant une différence significative d'alignement entre l'interlocuteur et le participant, avec un alignement plus important du robot sur le participant. Alors que l'alignement lexical a été décrit par la littérature comme un phénomène robuste dans les interactions humaines, ces résultats vont à l'encontre de notre hypothèse : bien que l'interlocuteur humain ait une parole libre tout au long de l'expérience, au contraire du robot qui est limité par les phrases pré-enregistrées dans le système de magicien d'Oz, ce premier a repris significativement moins de vocabulaire du participant que ne l'a fait l'interlocuteur artificiel. Nous pensons qu'il peut s'agir d'un artefact sous la forme d'un effet de report entre les essais avec l'interlocuteur robot. En effet, le résultat précédent suggère que le participant aligne son lexique sur celui du robot, il se peut que le lexique acquis du robot au cours d'un essai soit utilisé dans un essai suivant avec le robot. S'il s'agit effectivement d'un mot du robot mais qu'il est prononcé pour la première fois par le participant dans un nouvel essai et repris par le robot, on observe effectivement une augmentation de la mesure d'alignement du robot sur l'humain, mais qui ne représente pas un alignement au cours de l'essai, mais plutôt un alignement au cours de l'expérience dans la direction attendue, c'est-à-dire de l'humain vers le robot. Cette possibilité nécessite de développer une nouvelle approche pour prendre en compte les effets d'alignement lexical au cours des différents essais. Enfin, nous avons utilisé les données d'activité cérébrale pour valider la pertinence de la variable d'alignement lexical. Nous avons d'abord étudié les corrélats du temps de parole pour valider l'approche utilisée. Nous avons identifié des régions du cortex limbique gauche négativement corrélées avec LILLA. Alors que ces régions sont impliquées dans des processus mnésiques et dans le contrôle de l'action, nous proposons que plus le participant s'appuie sur le vocabulaire introduit par l'interlocuteur, moins il utilise ses ressources propres pour choisir les mots qui seront utilisés dans la discussion : il s'agit d'une réduction du contrôle cognitif lorsque l'on s'aligne sur le lexique de l'interlocuteur, indépendamment de la nature de l'interlocuteur. Ces résultats suggèrent que la variable d'alignement lexical LILLA est pertinente pour caractériser un aspect des comportements langagiers.

7. Conclusion

Dans cet article nous décrivons l'analyse d'un corpus d'interaction de participants humains qui discutent avec un humain ou avec un robot. Cette analyse nous a permis de vérifier et de quantifier les différences de comportement verbal entre les interlocuteurs humain et robot. Nous avons aussi étudié les relations en termes d'alignement entre les productions du participant et celles des interlocuteurs (humain ou robot) afin de dissocier les phénomènes automatiques de type alignement d'autres phénomènes liés à la nature humaine ou robotique de l'interlocuteur. Enfin, l'analyse neurophysiologique suggère que cette variable d'alignement décrit un vrai phénomène cognitif.

Remerciements

Cette recherche est soutenue par les subventions ANR-16-CONV-0002 (ILCB), ANR-11-LABX-0036 (BLRI) et par l'initiative d'excellence Aix-Marseille Université (AAP-ID-17-46-170301-11.1), un programme français « d'investissement futur » (A*MIDEX, ANR-11-IDEX-0001-02).

8. Bibliographie

Al Moubayed S., Beskow J., Skantze G., Granström B., « Furhat : A Back-Projected Human-Like Robot Head for Multiparty Human-Machine Interaction », *in* A. e. a. Esposito (ed.), *Cognitive Behavioural Systems*, Lecture Notes in Computer Science, Springer Berlin Heidelberg, p. 114-130, 2012.

Allwood J., Nivre J., Ahlsén E., « On the semantics and pragmatics of linguistic feedback », *Journal of semantics*, vol. 9, n° 1, p. 1-26, 1992.

Baiocchi L., Pauses remplies en interaction, mémoire de Master Théorie Linguistiques : terrain et expérimentation, Master thesis, Université d'Aix-Marseille, Aix-en-Provence, 2015.

Bigi B., « SPPAS - Multi-lingual approaches to the automatic annotation of speech », *The Phonetician*, vol. 111-112, p. 54-69, 2015.

Blache P., Bertrand R., Ferré G., *Creating and Exploiting Multimodal Annotated Corpora : The ToMA Project*, Springer-Verlag, Berlin, Heidelberg, p. 38–53, 2009.

Boersma P., « Praat, a system for doing phonetics by computer », *Glot. Int.*, vol. 5, n° 9, p. 341-345, 2001.

Branigan H. P., Pickering M. J., Cleland A. A., « Syntactic co-ordination in dialogue », *Cognition*, vol. 75, n° 2, p. B13-B25, May, 2000.

Branigan H. P., Pickering M. J., McLean J. F., Cleland A. A., « Syntactic alignment and participant role in dialogue », *Cognition*, vol. 104, n° 2, p. 163-197, August, 2007.

Branigan H. P., Pickering M. J., Pearson J., McLean J. F., Brown A., « The role of beliefs in lexical alignment : Evidence from dialogs with humans and computers », *Cognition*, vol. 121, n° 1, p. 41-57, October, 2011.

Brennan S. E., « Lexical entrainment in spontaneous dialog », *Proceedings of ISSD 96*, 1996.

Brett M., Anton J., Valabrgue R., Poline J.-B., « Region of interest analysis using an SPM toolbox. Presented at the 8th International Conference on Functional Mapping of the Human Brain, June 2-6, 2002, Sendai, Japan », *Neuroimage*, vol. 13, p. 210-217, 01, 2002.

Bunt H., « Context and dialogue control », *Think Quarterly*, vol. 3, n° 1, p. 19-31, 1994.

Chaminade T., « An experimental approach to study the physiology of natural social interactions », *Interaction Studies*, vol. 18, n° 2, p. 254-275, December, 2017.

Chaminade T., Prévot L., Ochs M., Rauchbauer B., « Challenges in Linking Physiological Measures and Linguistic Productions in Conversations », *1st Workshop on Linguistic and Neuro-Cognitive Resources*, Miyazaki, Japan, 2018.

Clark H. H., Schreuder R., Buttrick S., « Common ground at the understanding of demonstrative reference », *Journal of verbal learning and verbal behavior*, vol. 22, n° 2, p. 245-258, 1983.

Dennett D. C., *The intentional stance*, MIT Press, Cambridge, Mass, 1987.

Fan L., Li H., Zhuo J., Zhang Y., Wang J., Chen L., Yang Z., Chu C., Xie S., Laird A. R., Fox P. T., Eickhoff S. B., Yu C., Jiang T., « The Human Brainnetome Atlas : A New Brain Atlas Based on Connectional Architecture », *Cerebral Cortex*, vol. 26, n° 8, p. 3508-3526, May, 2016.

Fusaroli R., Bahrami B., Olsen K., Roepstorff A., Rees G., Frith C., Tylén K., « Coming to Terms », *Psychological Science*, vol. 23, n° 8, p. 931-939, July, 2012.

Henry S., Pallaud B., « Word fragments and repeats in spontaneous spoken French », *ISCA Tutorial and Research Workshop on Disfluency in Spontaneous Speech*, 2003.

Ochs M., Jain S., Blache P., « Toward an automatic prediction of the sense of presence in virtual reality environment », *Proceedings of the 6th International Conference on Human-Agent Interaction*, ACM, p. 161-166, 2018.

Pickering M. J., Garrod S., « The interactive-alignment model : Developments and refinements », *Behavioral and Brain Sciences*, April, 2004.

Price C. J., « The anatomy of language : a review of 100 fMRI studies published in 2009 », *Annals of the New York Academy of Sciences*, vol. 1191, n° 1, p. 62-88, March, 2010.

R Core Team, *R : A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. 2013.

Rauchbauer B., Nazarian B., Bourhis M., Ochs M., Prévot L., Chaminade T., « Brain activity during reciprocal social interaction investigated using conversational robots as control condition », *Philosophical Transactions of the Royal Society B : Biological Sciences*, vol. 374, n° 1771, p. 20180033, April, 2019.

Rauzy S., Montcheuil G., Blache P., « MarsaTag, a tagger for French written texts and speech transcriptions », *Proceedings of the Second Asian Pacific Corpus linguistics Conference*, 2014.

Roze C., Base Lexicale des Connecteurs discursifs du français, mémoire de DEA, Master thesis, Université de Paris Diderot, Paris, 2009.

Schiffrin D., *Discourse markers*, Cambridge University Press, Cambridge, New York, 1987.

Shriberg E. E., Preliminaries to a theory of speech disfluencies, PhD thesis, Citeseer, 1994.

Xu Y., Reitter D., « An Evaluation and Comparison of Linguistic Alignment Measures », *Proceedings of CMCL 2015*, p. 58-67, 2015.

Notes de lecture

Rubrique préparée par Denis Maurel

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

Rotem DROR, Lotem PELED-COHEN, Segev SHLOMOV, Roi REICHART. Statistical Significance Testing for Natural Language Processing. *Morgan & Claypool publishers*. 2020. 98 pages. ISBN 978-1-68173-795-9.

Lu par **François LÉVY**

Université Sorbonne Paris-Nord / LIPN

J'avais, en choisissant ce livre, l'envie de comprendre ce que signifient vraiment les différents tests qui sont employés pour tirer des enseignements de ces tableaux de résultats sur lesquels nous appuyons une grande part de nos articles. C'était un peu trop optimiste : le livre est construit autour de trois articles parus dans TACL et à ACL et les explications ajoutées sont un peu succinctes pour aider le novice à comprendre en détail la méthodologie statistique. Il contient néanmoins une réflexion intéressante sur l'évaluation des expériences multiples qui jouent un grand rôle dans le travail actuel, et propose des méthodes pour cela. En cent pages, ce n'est pas si mal.

Le plan du texte découle de sa genèse. Trois chapitres présentent la problématique générale du test d'hypothèse, qui décide de la signification statistique. La suite est consacrée à la comparaison d'algorithmes. Le chapitre 4 fait une revue des principales tâches de traitement automatique des langues, des indicateurs servant à évaluer les résultats et des tests de signification utilisables selon le type d'indicateur pour comparer deux algorithmes à partir d'un unique corpus. La troisième partie traite également de la comparaison de deux algorithmes sur le même corpus, dans le cas particulier où ils sont tels qu'il faut un ensemble d'exécutions pour rendre compte de leur performance. Ce choix est motivé par les réseaux neuronaux profonds (RNP) dont les résultats varient avec l'initialisation. Une quatrième partie envisage la situation où la comparaison est faite sur plusieurs corpus : la problématique est très liée à la portabilité et à l'adaptation des logiciels à d'autres domaines ou d'autres langues.

La description générale du test d'hypothèse est indispensable à la compréhension. On a un ensemble d'éventualités d'une taille inaccessible à l'exploration, par exemple une série infinie de lancers d'une pièce de monnaie. Une variable aléatoire v évalue chaque éventualité ; disons -1 pour pile et 1 pour face. On a aussi un jeu de données v_{obs} observé sur une population et décrit par une statistique S_{obs}. On aimerait savoir

si ce jeu est significatif : si la moyenne S_{obs} des mille tirages observés est *0,074*, peut-on en déduire que la pièce est déséquilibrée vers face ? Le raisonnement ressemble au raisonnement par l'absurde. On considère deux hypothèses, H_1 affirmant que l'observation est significative, et H_0 supposant le contraire. La première étape est de modéliser S sous l'hypothèse H_0 par une distribution de probabilité – si la pièce est équilibrée, la distribution de la moyenne de mille tirages s'approxime par une loi normale. Ensuite, on considère la probabilité que S soit dans l'intervalle des valeurs au moins aussi significatives que l'exemple (ici $P(S \geq 0{,}074)$). Cette probabilité est la p-valeur de l'observation : elle mesure la probabilité qu'un résultat apparemment significatif soit seulement l'effet du hasard. Dernière étape, on choisit un seuil de confiance α : si la p-valeur est plus petite que $1\text{-}\alpha$, l'observation permet de rejeter H_0.

On voit bien, dans cette présentation abstraite, les conditions de rigueur derrière la technique. En premier, il faut justifier la distribution supposée de S dans les jeux de données sous l'hypothèse H_0, en particulier contrôler que les données vérifient les conditions de validité de la distribution (par exemple, l'indépendance). En second, la p-valeur ne dit pas tout de la qualité de la décision : c'est la probabilité de S_{obs} *connaissant* H_0 et on ne peut en déduire la probabilité de H_0 *connaissant* S_{obs}. Augmenter le seuil de confiance réduit le nombre de cas où l'on rejette H_0 à tort, mais augmente ceux où l'on ne détecte pas que H_0 devrait être rejeté ; minimiser l'erreur est plus compliqué qu'il n'y paraît. D'un point de vue opérationnel, on appelle tests paramétriques ceux qui utilisent la distribution de S; quand celle-ci est inconnue, on peut se rabattre sur des tests dits non paramétriques : soit on utilise un indicateur dérivé moins précis, soit on évalue empiriquement la distribution. Quelques exemples de tests constituent l'essentiel du chapitre 3 ; ils sont trop succinctement expliqués pour apprendre grand-chose à qui ne les connaît pas déjà, ce qui fait de ce chapitre pour l'essentiel une indirection vers d'autres sources.

Le premier thème abordé ensuite est donc une revue des tests de signification servant à comparer deux algorithmes dans les différentes tâches de traitement automatique des langues. Les éventualités sont implicites ; on peut pour certaines tâches identifier dans ce rôle l'ensemble des phrases ou l'ensemble de textes possibles. Pour un jeu de données, la statistique est la différence δ des indicateurs mesurant la performance de chacun des algorithmes à comparer (par exemple, leur précision) et, sans perte de généralité, l'hypothèse H_1 est que δ est positif. Les valeurs au moins aussi significatives que l'exemple vérifient alors $\delta \geq \delta_{obs}$. L'essentiel est un tableau d'une vingtaine d'indicateurs utilisés dans de nombreuses tâches de TAL, avec les tests paramétriques (si possible) et non paramétriques qui conviennent pour cet indicateur. À noter les indications de rigueur qui figurent en renvoi, par exemple, que les tests paramétriques supposent une distribution normale de l'indicateur. Et la dernière notation du chapitre vaut la peine d'être signalée : quand la différence de performance entre les algorithmes est faible, on peut changer la réponse du test en augmentant la taille du jeu de données observé ! Cela s'appelle du « *p-hacking* ».

Le thème de la comparaison des réseaux neuronaux profonds est plus récent. Le constat de départ est simple : même à ensemble d'apprentissage et de développement constants, une seule exécution du réseau ne suffit pas à déterminer un indice de performance fiable sur l'ensemble de test. Ceci est dû à ce que les auteurs appellent le non-déterminisme des RNP : le grand nombre de paramètres de structure, la variabilité de leurs méthodes d'optimisation, les choix aléatoires (par exemple, d'initialisation des poids ou d'ordre de présentation des exemples) peuvent, toutes choses égales par ailleurs, produire des variations de performance importantes. Il est donc nécessaire pour comparer deux algorithmes de les caractériser chacun par un ensemble d'exécutions. En passant, la définition implicite des éventualités a changé par rapport au premier thème : ce sont les exécutions possibles sur un seul jeu de données. Les auteurs critiquent les méthodes de comparaison simples. Ils proposent à la place un affaiblissement de l'ordre stochastique naturel du premier ordre[1], dont ils estiment le critère trop strict. Ils mesurent pour cela à quelle distance εW_2 (entre *0* et *1*) une des variables peut dominer l'autre. On peut alors calculer le meilleur majorant de εW_2 pour les ensembles d'exécutions et un seuil de confiance donné. L'analyse empirique comporte des expériences avec un bruit contrôlé et cinq cent dix comparaisons extraites de cinq tâches différentes. Elle incite à poursuivre cette piste.

Le dernier thème concerne la comparaison de deux algorithmes sur plusieurs jeux de données. Une remarque liminaire pose le problème : si dix jeux rejettent chacun H_0 avec un seuil de confiance de *0,95*, le seuil de confiance de l'appréciation globale « A est dans les dix cas meilleur que B » est un peu en dessous de *0,6*. Ici, les éventualités sont donc les types de jeux de données (type de texte, langue, etc.). Les auteurs utilisent des méthodes développées pour le traitement d'images médicales. S'il y a N jeux de données, on considère une famille d'hypothèses nulles ${H_0}^{u/N}$ considérant que moins de u hypothèses nulles parmi celles attachées à chaque jeu de données individuel sont fausses. Comme on connaît les p-valeurs associées aux hypothèses nulles individuelles, les auteurs proposent deux formules pour calculer celles associées aux ${H_0}^{u/N}$ selon que les jeux de données sont garantis d'être indépendants ou pas. Il est alors possible de calculer le nombre d'hypothèses nulles que l'on peut rejeter avec le seuil de confiance global α. Dans un deuxième temps, une procédure fournit une liste de jeux de données pour lesquels on peut rejeter l'hypothèse nulle avec le même seuil de confiance global (dans le cas d'indépendance, moins que ce que l'on avait calculé précédemment). Ici aussi, une série d'expériences complète l'argumentation. Les trois solutions comparées donnent des résultats différents, mais les critères d'indépendance qui permettent d'en choisir un sont très elliptiques.

En conclusion, ce livre traite d'un problème important pour le traitement automatique des langues : quand on compare des algorithmes, comment interpréter rigoureusement les résultats obtenus sur un corpus particulier. Et il propose des

1 Pour les variables aléatoires réelles X et Y, $X \geq_{1a} Y$ si, et seulement si, pour tout a, $P(X > a) \geq P(Y > a)$.

solutions pour des configurations dont l'importance est récente, les algorithmes non déterministes et les évaluations multicorpus. Il signale aussi les points encore obscurs dans l'utilisation du test d'hypothèses par le traitement automatique des langues, au premier rang desquels la difficulté à décider de l'indépendance des données qui joue pourtant un rôle crucial. Il me semble qu'une explicitation plus détaillée de l'ensemble des éventualités sur lequel on raisonne aiderait à clarifier la question.

Shashi NARAYAN, Claire GARDENT. Deep Learning Approaches to Text Production. *Morgan & Claypool publishers*. 2020. 176 pages. ISBN 978-1-68173-758-4.

Lu par **Caio Filippo CORRO**

Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique (LISN)

Cet ouvrage propose une vue d'ensemble sur la recherche en génération automatique de textes fondée sur des réseaux de neurones. Il est découpé en trois parties couvrant (1) les modèles fondamentaux, (2) les méthodes neuronales avancées et enfin (3) les jeux de données disponibles. L'ouvrage est agréable à lire et contient de nombreuses illustrations et exemples. Il ne nécessite pas de prérequis sur le domaine et conviendra donc parfaitement à un stagiaire de master ou un doctorant voulant découvrir les approches modernes de génération automatique de textes.

Génération automatique de textes et réseaux de neurones

Le domaine de la recherche sur la génération automatique de textes s'intéresse à un ensemble très hétérogène de tâches qui ont toutes en commun de produire des énoncés en langage naturel, mais qui diffèrent à la fois sur le type de source d'information que l'on utilise en entrée du modèle et les objectifs attendus pour la sortie. L'entrée peut être une phrase, un document complet, une représentation sémantique ou encore des données brutes. Les sorties attendues peuvent être soit dans la même langue que l'entrée, soit dans une langue différente (traduction automatique). Certaines tâches portent beaucoup d'importance sur les propriétés du texte généré. Par exemple, pour la simplification de textes, l'objectif est de réécrire une ou plusieurs phrases afin de les rendre plus accessibles. Bien que l'idée du « modèle unique » qui s'appliquerait à tous les cas de figure soit séduisante, en pratique cela ne fonctionne pas. L'ouvrage propose un panorama des architectures neuronales et des méthodes d'entraînement permettant de prendre en compte ces différentes spécificités.

Contenu de l'ouvrage

La première partie de l'ouvrage comprend deux chapitres introduisant les grandes généralités sur la génération automatique de textes. Le chapitre 2 est un résumé très bref (une dizaine de pages) sur les méthodes non neuronales. Le

chapitre 3 détaille l'architecture encodeur décodeur standard de l'approche neuronale :

- l'encodeur transforme l'entrée, par exemple une phrase ou une représentation sémantique, en un vecteur de taille fixe aussi appelé plongement ;
- le décodeur utilise ce vecteur pour conditionner la génération d'une phrase avec un modèle autorégressif.

Ce chapitre décrit brièvement, mais efficacement, le fonctionnement des réseaux de neurones récurrents, les plongements lexicaux et les méthodes d'entraînement. Cela peut tout à fait faire office d'introduction à l'apprentissage profond pour un apprenti chercheur dans ce domaine. Le chapitre se termine par une brève comparaison avec les méthodes présentées dans le chapitre 2.

La seconde partie de l'ouvrage, certainement la plus importante, présente les différentes modifications pouvant être apportées à l'architecture encodeur décodeur. Le chapitre 4 s'intéresse aux mécanismes qui permettent d'améliorer les interactions entre l'encodeur et le décodeur : attention, copie et contraintes de couverture. Ces trois mécanismes sont très liés et sont fondés sur l'idée d'introduire un lien direct entre les mots de l'entrée et la distribution sur le vocabulaire de sortie à chaque position. L'attention permet de créer « un raccourci » entre l'entrée et la sortie, ce qui est particulièrement important, car le plongement produit par l'encodeur est de taille fixe, ce qui induit des difficultés pour les entrées longues et complexes. La copie permet de faire un copier-coller des termes de l'entrée, par exemple pour générer des noms propres qui ne sont pas vus lors de l'entraînement et donc sont non présents dans le vocabulaire du décodeur. Enfin, la couverture permet d'empêcher le décodeur de produire des contenus redondants ou d'oublier des parties importantes de l'entrée. Le chapitre est suffisamment détaillé pour permettre à celui qui le souhaite de travailler directement sur une implémentation de ces trois mécanismes.

Le chapitre 5 présente les différents types d'architectures neuronales pouvant être utilisés au niveau de l'encodeur pour s'adapter à la structure de l'entrée. Bien que de nombreux travaux dans la littérature se focalisent uniquement sur les réseaux de neurones récurrents, peu importe la structure de l'entrée, ces derniers rencontrent des difficultés en pratique. Dans le cas de textes longs, une approche naïve consiste à simplement concaténer l'ensemble des phrases en une unique séquence. Les réseaux de neurones récurrents se heurtent alors au problème de la prise en compte des dépendances à longue distance. En effet, ceux-ci doivent encoder un historique arbitrairement long dans une représentation de taille fixe. L'ouvrage présente des approches de modélisation fondées sur des architectures neuronales hiérarchiques où la structure du document (découpage en phrases, en paragraphes) est prise en compte pour contrer ce problème. Dans le cadre d'entrées qui ne sont pas des énoncés en langage naturel, comme dans le cas des représentations sémantiques qui ont une forme de graphe, l'approche dite de linéarisation consiste à représenter ce graphe sous forme d'une séquence. Ceci pose des problèmes de pertes d'informations (certaines relations ne sont pas représentées dans la linéarisation) et de localité (des nœuds proches dans le graphe peuvent se retrouver éloignés dans la

linéarisation). L'ouvrage décrit les approches fondées sur des réseaux de convolution de graphes pour encoder directement les représentations sémantiques sans linéarisation artificielle.

Enfin, le chapitre 6 s'intéresse à la prise en compte des objectifs attendus sur le texte produit. Par exemple, dans le cas de la génération de résumés, il est important de considérer des systèmes en deux étapes qui commencent par repérer les contenus saillants avant de s'attaquer à la génération proprement dite. Ce chapitre s'intéresse aussi à l'utilisation des métriques d'évaluation comme objectif à optimiser lors de l'apprentissage. Ces métriques sont pour la plupart non différentiables. L'ouvrage se concentre sur les approches d'apprentissage par renforcement qui permettent de contourner ce problème en utilisant directement le score donné par la métrique comme signal d'apprentissage. Les autres méthodes comme la construction de substituts différentiables ne sont pas abordées.

La troisième et dernière partie de l'ouvrage présente en un unique chapitre les jeux de données disponibles pour les différentes tâches de génération automatique de textes.

Conclusion

Cet ouvrage propose une vue d'ensemble sur la recherche en génération automatique de textes fondée sur des réseaux de neurones. Un soin particulier a été apporté au niveau de la pédagogie : très peu de prérequis sont nécessaires, de nombreuses figures permettent de comprendre rapidement les concepts décrits dans le texte (le chapitre 3 est d'une exemplarité rare dans ce domaine !) et les équations sont clairement détaillées. Cet ouvrage pourrait être un très bon support pour un cours niveau master sur le sujet.

Cependant, il décevra probablement les chercheurs aguerris. Les différents thèmes ne sont couverts que de façon sommaire. Le livre ne contient ni analyse théorique ni analyse expérimentale. Par exemple, la section sur l'apprentissage par renforcement contient une explication claire qui permet de bien comprendre pourquoi cette approche n'est pas dépendante directement du gradient des métriques d'évaluation pour entraîner un modèle. Cependant, la forte variance de cet estimateur n'est pas expliquée formellement et n'est que brièvement discutée. Aucune analyse des méthodes de réduction de la variance et de leur impact sur le temps d'entraînement et les résultats expérimentaux n'est proposée.

Jacques MOESCHLER. Pourquoi le langage ? Des Inuits à Google. *Armand Colin*. 2020. 286 pages. ISBN 978-2-200-62855-0.

Lu par **Alain MILLE**

Université Lyon 1 / LIRIS UMR CNRS 5205

L'auteur s'appuie sur les théories linguistiques, cognitives et pragmatiques pour s'interroger sur ce que l'on sait et ce que l'on ne sait pas sur le langage. Il ne répond pas à la question du

titre « Pourquoi le langage ? », mais brosse un tableau de toutes les questions qu'il faut étudier pour répondre à cette simple interrogation. L'ouvrage a une facette pédagogique, avec beaucoup d'exemples et de contre-exemples pour illustrer la complexité de la tâche. Il manipule souvent la démonstration par l'absurde pour démontrer que les théories standard ne suffisent pas et il introduit progressivement la théorie pragmatique qu'il défend comme une approche permettant de comprendre le langage en contexte. Certaines approches, essentiellement les approches en neurolinguistique, ne sont pas abordées dans l'ouvrage qui donne toutefois une bonne vision de l'état des connaissances actuelles pour répondre à la question « Pourquoi le langage ? ».

Notice d'ouvrage

En avant-propos, l'auteur livre ses intentions : 1) parce que le langage est important et certainement la propriété unique qui définit notre espèce, 2) parce que le langage est généralement considéré comme allant de soi, et que pourtant si les connaissances sont maintenant importantes elles sont encore très lacunaires. Il y a donc d'une part un discours de pédagogie pour montrer l'importance du langage et de ce qu'il révèle de la nature humaine et, d'autre part, une sorte d'état de l'art des connaissances en la matière, en assumant le fait de ne pas aborder les travaux en neurosciences cognitives ou en linguistique informatique.

L'introduction reprend ces objectifs en précisant le public cible : toute personne s'intéressant au langage qu'il soit scientifique ou non. L'auteur annonce qu'il présentera surtout ce qu'il connaît bien et fera un focus sur ses propres travaux en pragmatique. Il donne plusieurs raisons qui font qu'il est difficile de traiter du langage dans un ouvrage unique : 1) chaque locuteur se sent expert de sa propre langue, 2) une approche scientifique du langage est considérée avec scepticisme par certaines disciplines, 3) l'idée que la culture et la langue française seraient supérieures aux autres, d'autant que les autres doivent penser la même chose de leur propre langue… Une bonne partie du reste de l'introduction est consacrée à la démonstration de la valeur et de l'importance de la linguistique comme discipline scientifique. Il pose que l'hypothèse relativiste (pas d'invariants linguistiques universels) l'emporte sur l'hypothèse que toutes les langues seraient des variations d'un même patron, une grammaire universelle. Il annonce deux thèmes majeurs qu'il abordera de manière plus approfondie : la différence entre étude du langage et étude de la communication, et la valeur d'une approche pragmatique qu'il défend en tant que chercheur engagé dans cette voie.

Partie I

La partie I traite du lien entre langage et communication.

Dans le chapitre 1, l'auteur énonce ce qui constitue à son avis huit idées fausses sur le langage : 1) les langues non écrites ne seraient pas de « vraies » langues, 2) il y aurait des langues plus « importantes » que d'autres, 3) le français serait une langue logique, claire et belle, 4) les langues souffriraient de l'influence d'autres langues, 5) il faudrait protéger le français de l'influence des autres langues, 6) les enfants apprendraient leur langue maternelle par imitation des paroles de leurs parents, 7) seuls les mots du dictionnaire appartiendraient à la langue et 8) le

linguiste s'intéresserait à l'origine des mots. Il conclut par les propositions alternatives consistant à nier les idées fausses.

Remarque sur ce chapitre : la réfutation de l'idée fausse 6 est l'occasion d'introduire un principe qui sera repris largement dans le reste de l'ouvrage : « *les enfants n'apprennent pas seulement par la langue maternelle par imitation de leurs parents, mais naturellement parce qu'ils sont programmés pour cela* ». On peut s'interroger sur le terme de *programmation.*

Le chapitre 2 se présente avec une volonté démonstrative de l'affirmation que le langage n'est pas la communication et la communication n'est pas le langage. Après avoir fait une présentation pédagogique de la communication, l'auteur introduit deux modèles : un modèle *codique* basé sur les codes linguistiques et les codes sociaux et un modèle *inférentiel* où les informations non formulées sont déduites du contexte. De la même façon la notion de langage est présentée laconiquement : « *Un langage est constitué d'une phonologie, d'une sémantique et d'une syntaxe* ». Deux fonctions du langage sont affirmées : communication et cognition. Si l'aspect communication a été déjà anticipé, l'aspect cognition est démontré par la propriété que le langage a d'enchâsser une structure, quelle qu'elle soit, dans une structure du même type. La communication ne serait qu'un effet de bord d'une émergence du langage comme support du développement de la cognition : le langage comme support de la pensée. Cette externalisation de la pensée permet la réflexion et la récursivité qui expliqueraient le lien entre la pensée, le langage et le raisonnement.

Le chapitre 3 défend l'idée que ce n'est pas la structure du langage qui en définit l'usage. C'est ici le thème de prédilection de l'auteur : la pragmatique comme étude de l'usage du langage dans la communication. L'apprentissage de la langue ne se fait pas à partir des règles du langage. La conversation obéit à des règles qui ne sont pas liées spécifiquement au langage. À l'instar de Noam Chomsky, la distinction est faite entre compétence et performance du langage. Une distinction est ainsi introduite entre langue interne (li) et langue externe (le). L'auteur interroge l'agenda de la recherche sur la question en se demandant s'il faut être compétent (connaître finement la langue) pour être performant (utiliser finement la langue). À la suite de Chomsky, il est admis que le système cognitif (*Faculty of Language in the Narrow sense*, FLN) interagit avec une faculté (*Faculty of Languange in the Broad sense*, FLB) s'appuyant sur deux systèmes « externes » : le système articulatoire perceptuel et le système conceptuel intentionnel. FLB serait une évolution d'une faculté partagée avec beaucoup d'animaux tandis que FLN serait une faculté apparue plus récemment et spécifique de l'humain. Il y aurait donc eu un protolangage (langage sans syntaxe) avant le langage moderne. L'auteur décrit comment la compréhension de la communication verbale a été modifiée par l'hypothèse d'un principe de coopération et de maximes de conversation, qui sont liées à la cognition. De nombreux exemples émaillent cette partie pédagogique de l'ouvrage. Cette section permet de montrer qu'il faut ajouter un principe de pertinence pour expliquer la communication verbale. Cette théorie de la pertinence est traitée largement, en particulier pour le cas de la communication implicite.

Partie II

La partie II s'intéresse à la partie sociale du langage.

Dans le chapitre 4, l'auteur s'interroge à nouveau sur ce qui serait caractéristique de l'humain, le langage n'étant pas seul candidat. C'est dans ce chapitre que l'auteur développe la théorie qu'il va ensuite affirmer à partir de l'hypothèse SAPIR-WHORF qui articule le langage avec la culture *via* le vocabulaire qui se stabilise. Cette relation met en évidence qu'il existe des mots intraduisibles car en relation avec des éléments culturels disjoints. Dans l'approche pragmatique, la valeur d'un mot ne prend son sens que dans son usage en contexte. Les variations linguistiques sont traitées avec l'exemple du français et du « vernaculaire noir-américain ». Le chapitre se termine par des mises en évidence des formes de politesse, de face et de figuration qui sont autant de formes étudiées dans la communication verbale.

Dans le chapitre 5, l'auteur commence par démontrer qu'il n'y a pas de règles spécifiques au discours par un raisonnement par l'absurde mobilisant des contre-exemples. Il conclut qu'aucune des règles souvent admises n'est générale ni constructive de la notion de discours. Cette démonstration faite, l'auteur propose sa théorie d'une pragmatique du discours. La cohérence est un jugement de l'observateur du discours, pas une propriété intrinsèque du discours : c'est la conséquence du processus de compréhension. Dans le cadre de la théorie de la pertinence, la compréhension est liée à deux effets sur les croyances : effets non propositionnels et effets propositionnels. Les effets non propositionnels sont liés aux émotions et à l'adhésion aux éléments du discours. Les effets propositionnels, au contraire, sont liés à l'argumentation et à la persuasion. Le discours n'est pas une unité linguistique, mais une unité pragmatique.

Dans le chapitre 6, une introduction s'interroge sur la fonction des métonymies, des métaphores, des images mentales dans le langage. L'auteur réfute les fonctions classiques qu'on leur attribue, ce qui lui permet d'introduire le modèle explicatif de la théorie de la pertinence. Cette théorie fournirait des clés nouvelles pour comprendre ces usages. C'est le contexte du discours émis ou observé qui fournit le sens, pas la valeur littérale des expressions utilisées, même si elles sont ancrées dans la culture partagée. La structure de récit est également remise en question comme n'étant pas une structure de communication, mais caractérisée par un ordre des événements dans un style indirect libre encore une fois lié essentiellement à une stratégie linguistique pour prêter des paroles ou des pensées à une troisième personne. L'auteur termine le chapitre en introduisant la notion de super-pragmatique, au-delà de la pragmatique déjà traitée dans les chapitres précédents.

Dans le chapitre 7, en exploitant l'idée de la super-linguistique, l'auteur propose l'idée de super-pragmatique pour utiliser la méthode de la pragmatique pour aller au-delà du domaine d'investigation de la pragmatique. C'est ainsi qu'il se propose d'exploiter deux concepts centraux de la pragmatique, présupposition et implicature, pour comprendre des enjeux de la société et ses crises. La compréhension d'un énoncé mobilise plusieurs mécanismes : implicature conversationnelle, implicature conventionnelle, présupposition et implications logiques. L'auteur propose d'ajouter la notion « d'explicature » qui peut expliciter directement la proposition exprimée

dans l'énoncé, mais aussi, à un niveau supérieur, la valeur d'acte de langage de l'énoncé. Ces mécanismes varient en termes d'accessibilité et de force. Les rôles des présuppositions puis des implicatures sont interrogés. Les présuppositions doivent être partagées et les implicatures peuvent être annulées par des implications qui s'imposent. Pour illustrer l'intérêt de la super-pragmatique, l'auteur s'intéresse à la posture de décryptage adoptée par les journalistes qui formulent des implications sur les informations qu'ils traitent. Le cas d'usage de « *Je suis Charlie* » est utilisé pour montrer que si « *Je suis Charlie* » alors cela implique un point de vue relativement partagé en solidarité avec les victimes de l'attentat, le « *Je ne suis pas Charlie* » n'implique par un point de vue partagé avec les assassins pour autant. L'auteur conclut le chapitre par cette phrase : « *Ce qui manque crucialement aujourd'hui, c'est une réflexion approfondie sur le langage et les implications de son usage par la presse en générale et les journalistes en particulier* ».

En conclusion, l'auteur distingue ce que nous savons et ce que nous ne savons pas encore sur le langage. *Ce que nous savons (imparfaitement) :* de la syntaxe à la pragmatique. La linguistique est complexe, car une part fondamentale repose sur sa nature orale, dynamique et sur le contexte échappant à l'énoncé. *Ce que nous ne savons pas encore :* émotions, origine du langage, traduction automatique, et communication homme-machine.

Résumés de thèses et HDR

Rubrique préparée par Sylvain Pogodalla

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

Vincent CLAVEAU : vincent.claveau@irisa.fr

Titre : Du traitement des langues en recherche d'information et vice versa

Mots-clés : traitement automatique des langues, recherche d'information, intelligence artificielle.

Title: *About Natural Language Processing for Information Retrieval and vice versa*

Keywords: *natural language processing, information retrieval, artificial intelligence.*

Habilitation à diriger des recherches en informatique, IRISA, Université de Rennes 1. Habilitation soutenue le 10/01/2020.

Jury : Mme Adeline Nazarenko (Pr, Université Paris-Nord, présidente), Mme Catherine Berrut (Pr, Université de Grenoble Alpes, rapporteuse), M. Philippe Langlais (Pr, Université de Montréal, Canada, rapporteur), M. Jacques Savoy (Pr, Université de Neuchâtel, Suisse, rapporteur), M. Patrice Bellot (Pr, Université Aix-Marseille, examinateur), M. Olivier Dameron (MC, Université Rennes 1, examinateur).

Résumé : *La recherche d'information (RI) et le traitement automatique des langues (TAL) sont deux domaines de recherche de l'informatique partageant en commun leur matériau premier : la langue. Pourtant, à quelques exceptions notables près, ces deux domaines ont longtemps évolué indépendamment, avec peu d'interactions.*

C'est ce rapprochement entre TAL et RI, pour peu que l'on veuille les distinguer, qui est le fil conducteur principal de ce manuscrit. Au travers de la présentation d'une partie de nos travaux, nous montrons les allers-retours, les convergences, les synergies, qu'il peut y avoir entre ces deux domaines.

Ce document n'offre donc pas une vue exhaustive de nos travaux, ni en largeur (tous n'y sont pas présentés), ni en profondeur (tous les détails techniques n'y sont pas

reportés). Ce document n'est pas non plus une revue de l'état de l'art, mais pour situer les travaux présentés dans un contexte plus large, nous proposons deux brefs panoramas des interactions entre TAL et RI.

Au travers d'une sélection de nos travaux passés, nous montrons ainsi tous les bénéfices à croiser les connaissances acquises dans chacun de ces domaines. Précisément, nous avons articulé ce mémoire en deux parties, l'une dédiée aux apports du TAL pour la RI, et l'autre aux apports de la RI pour le TAL. Nous revisitons ainsi plusieurs de nos contributions sur, d'une part, la morphologie, la translittération, la segmentation thématique, l'analyse fine de termes médicaux dans un contexte de RI, et d'autre part, sur l'utilisation des moteurs de recherche comme classifieurs, les tâches de RI comme techniques d'évaluation de techniques de TAL, la sémantique distributionnelle et les plongements de mots par et pour la RI. Nous discutons également de la pertinence de cette dichotomie entre ces deux domaines à l'heure de l'intelligence artificielle, et de la convergence de leur corpus technique (notamment les approches neuronales). Nous présentons enfin quelques enjeux de recherche à la croisée de ces domaines.

URL où le mémoire peut être téléchargé :
https://hal.archives-ouvertes.fr/tel-03027676

Alice MILLOUR : alice.millour@abtela.eu

Titre : Myriadisation de ressources linguistiques pour le traitement automatique de langues non standardisées

Mots-clés : myriadisation, traitement automatique des langues, langues peu dotées, langues non standardisées, corpus annoté, morphosyntaxe, annotation manuelle.

Title: *Crowdsourcing Linguistic Resources for Non-standardised Languages Processing*

Keywords: *crowdsourcing, natural language processing, less-resourced languages, non-standardized languages, annotated corpora, part-of-speech, manual annotation.*

Thèse de doctorat en informatique, Sens Texte Informatique Histoire, Sociologie et Informatique pour les Sciences Humaines, Sorbonne Université, Paris, sous la direction de Karën Fort (MC, Sorbonne Université) et Claude Montacié (Pr, Sorbonne Université). Thèse soutenue le 14/12/2020.

Jury : Mme Karën Fort (MC, Sorbonne Université, codirectrice), M. Claude Montacié (Pr, Sorbonne Université, codirecteur), M. Laurent Besacier (Pr, Université Grenoble Alpes, Laboratoire d'Informatique de Grenoble, rapporteur), M. Benoît Sagot (DR, Inria, rapporteur), Mme Iris Eshkol-Taravella (Pr, Université Paris Nanterre, présidente), Mme Delyth Prys (Pr, Bangor University, Royaume-Uni, examinatrice).

Résumé : *Les sciences participatives, et en particulier la myriadisation (*crowdsourcing*) bénévole, représentent un moyen peu exploité de créer des*

ressources langagières pour certaines langues encore peu dotées, et ce malgré la présence de locuteurs sur le Web. Or, le développement de technologies du langage est très fortement dépendant de l'existence de ressources pérennes, qu'elles soient brutes ou annotées.

Nous présentons dans ce travail de thèse nos expériences de production participative de ressources et de développement — grâce à ces ressources — d'outils d'annotation automatique en parties du discours. Nous avons appliqué notre méthodologie à trois langues non standardisées, en l'occurrence l'alsacien, le créole guadeloupéen et le créole mauricien. Pour des raisons historiques différentes, de multiples pratiques (ortho)graphiques co-existent en effet pour ces trois langues. Les contextes linguistiques choisis nous ont confrontée à l'adaptabilité des méthodes habituellement employées pour développer des outils en TAL. En particulier, les difficultés posées par l'existence de cette variation nous ont menée à proposer trois tâches de myriadisation permettant respectivement la collecte de corpus bruts, l'annotation en parties du discours de ces corpus, et la production de variantes graphiques.

L'analyse intrinsèque et extrinsèque de ces ressources recueillies auprès de locuteurs montre l'intérêt d'utiliser la myriadisation dans un cadre linguistique non standardisé : les locuteurs ne sont pas considérés dans notre travail comme un ensemble uniforme de contributeurs dont les efforts cumulés permettent d'achever une tâche particulière, mais comme un ensemble de détenteurs de connaissances complémentaires. En outre, la variation graphique observée tend à dégrader les performances des outils reconnus comme performants dans des contextes standardisés. Ainsi, parallèlement à la définition d'une méthodologie de collecte de ressources variées, nous menons une évaluation de l'impact de la variation sur les performances des outils entraînés, puis nous proposons une démarche qui vise à intégrer ces ressources variées au développement d'outils plus robustes.

La qualité des ressources produites au cours de ce travail et les gains observés quant aux performances des outils entraînés nous permettent de conclure au bien-fondé de l'utilisation de la myriadisation pour le développement de ressources langagières dans ces contextes linguistiques particuliers. Les plateformes développées, les ressources langagières, ainsi que les modèles d'outils d'annotation entraînés sont librement disponibles.

URL où le mémoire peut être téléchargé :
https://tel.archives-ouvertes.fr/tel-03083213

Bénédicte PIERREJEAN : benedicte.pierrejean@gmail.com

Titre : Évaluation qualitative des *word embeddings* : étude de l'instabilité dans les modèles neuronaux

Mots-clés : *word embeddings*, sémantique distributionnelle, évaluation qualitative.

Title: *Qualitative Evaluation of Word Embeddings: Investigating the Instability in Neural-Based Models*

Keywords: *word embeddings, distributional semantics, qualitative evaluation.*

Thèse de doctorat en sciences du langage, CLLE-ERSS, UMR 5263, Université Toulouse 2 - Jean Jaurès, sous la direction de Ludovic Tanguy (MC, Université Toulouse 2 - Jean Jaurès). Thèse soutenue le 08/01/2020.

Jury : M. Ludovic Tanguy (MC, Université Toulouse 2 - Jean Jaurès, directeur), M. Olivier Ferret (IR, CEA LIST, rapporteur), M. Alessandro Lenci (Pr, University of Pisa, Pise, Italie, rapporteur), Mme Cécile Fabre (Pr, Université Toulouse 2 - Jean Jaurès, examinatrice), Mme Aurélie Herbelot (*assistant professor*, University of Trento, Trente, Italie, examinatrice).

Résumé : *Distributional semantics has been revolutionized by neural-based word embeddings methods such as word2vec that made semantics models more accessible by providing fast, efficient and easy-to-use training methods. These dense representations of lexical units based on the unsupervised analysis of large corpora are more and more used in various types of applications. They are integrated as the input layer in deep learning models, or they are used to draw qualitative conclusions in corpus linguistics. However, despite their popularity, there still exists no satisfying evaluation method for word embeddings that provides a global yet precise vision of the differences between models. In this PhD thesis, we propose a methodology to qualitatively evaluate word embeddings and provide a comprehensive study of models trained using word2vec. In the first part of this thesis, we give an overview of distributional semantics evolution and review the different methods that are currently used to evaluate word embeddings. We then identify the limits of the existing methods and propose to evaluate word embeddings using a different approach based on the variation of nearest neighbors. We experiment with the proposed method by evaluating models trained with different parameters or on different corpora. Because of the non-deterministic nature of neural-based methods, we acknowledge the limits of this approach and consider the problem of nearest neighbor's instability in word embeddings models. Rather than avoiding this problem we embrace it and use it as a means to better understand word embeddings. We show that the instability problem does not impact all words in the same way and that several linguistic features are correlated. This is a step towards a better understanding of vector-based semantic models.*

URL où le mémoire peut être téléchargé :
https://hal.archives-ouvertes.fr/tel-02628954