

La pragmatique dans les conversations avec des agents artificiels : comment l'évaluer ?

Baptiste Jacquet

Une thèse présentée pour le diplôme de
Docteur en Psychologie
Sous la direction de Jean Baratgin
Professeur à Université Paris 8



École Doctorale n°224 : Cognition, Langage, Interaction
Université Paris 8 - Vincennes - Saint-Denis
2 rue de la Liberté
93526 Saint-Denis Cedex, France

Soutenue le 25 Novembre 2021

Jury :

M. Alain JAILLET, Professeur à CY Cergy Paris Université	Rapporteur
Mme. Laura MACCHI, Professeure à Università degli studi di Milano BICOCCA	Rapportrice
M. Dominique ARCHAMBAULT, Professeur à Université Paris 8	Examinateur
Mme. Sabine GUÉRAUD, Maitresse de Conférences à Université Paris 8	Examinatrice
Mme. Carine ROYER, Maitresse de Conférences à CY Cergy Paris Université	Examinatrice
M. Jean BARATGIN, Professeur à Université Paris 8	Directeur de Thèse

Remerciements

Je tiens tout d'abord à remercier mon père, Thierry Jacquet, décédé le 8 Mars 2020. Cette thèse n'aurait jamais pu exister sans son aide morale et son soutien. J'espère qu'il en aurait été fier.

Je souhaite évidemment remercier mon directeur de thèse, Jean Baratgin, pour son soutien sans faille au cours de ces dernières années, à toujours savoir tirer le meilleur de mes qualités et d'avoir été capable de comprendre mes défauts pour en limiter les effets, à toujours me pousser vers l'avant.

Je tiens également à remercier toute l'équipe Cognition, rationalités naturelle et artificielle (et en particulier Frank Jamet et Jean-Louis Stilgenbauer) ainsi que l'association P-A-R-I-S pour m'avoir donné une atmosphère de travail à la fois enrichissante et agréable, et que même si nous avons eu collectivement de nombreux moments de stress ensemble, nous sommes toujours parvenus à les surmonter en travaillant en équipe.

Je remercie également le laboratoire CHArt pour son soutien, et en particulier sa directrice Isis Truck, et mes collègues doctorants avec qui j'ai pu longuement échanger sur nos recherches.

Je souhaite également remercier Alexandre Hullin, Benjamin Beaunay et Caline Jaraud pour leur aide en ce qui concerne les passations de certaines des expériences contenues dans cette thèse.

Enfin, je remercie également ma tante et mon oncle, Patricia et Jean-Michel Amy, ainsi que ma soeur, Johanna Jacquet, pour avoir sur bien des aspects pris le relais suite à la perte de mon père et pour m'avoir aidé dans ces dernières années difficiles.

J'ajoute également mes deux chiens, Patachou et Linoa, sans lesquels ces années auraient certainement été bien moins animées...

Résumé

Avec le développement rapide de systèmes informatiques permettant de communiquer avec des utilisateurs ou clients de manière automatisée par le langage naturel, des questions se posent sur la capacité de ces outils à répondre de manière adéquate à l'utilisateur. En effet, de nombreuses études existent sur le fait que les utilisateurs de tels systèmes sont souvent frustrés par leur incapacité à se faire comprendre de la machine, qui se retrouve à répondre de manière non pertinente, surtout lors de conversations dépassant la simple question-réponse. Dans cette thèse, nous étudierons les problèmes concernant la pragmatique du langage, souvent ignorées par ces outils automatisés que sont les **chatbots**, des agents de conversation virtuels, ainsi que par des robots sociaux, des agents de conversations incarnés, et nous chercherons à y apporter des réponses en nous basant notamment sur les travaux de Grice (1975) sur la pragmatique conversationnelle et de Wilson et Sperber (2002) sur la Théorie de la Pertinence. Dans une première partie nous partirons donc des observations permettant d'affirmer que les agents conversationnels artificiels actuels ont de nombreux problèmes concernant la pragmatique du langage et notamment l'utilisation du contexte et de l'implicite. En effet, ils ne peuvent généralement pas soutenir des conversations longues mais restent relativement performants lorsqu'il s'agit uniquement de suites de questions-réponses. Suite à ces observations, nous proposerons un modèle théorique de **chatbot** permettant de commencer à résoudre ces problèmes en tirant parti des aspects de Théorie de l'esprit (Premack & Woodruff, 1978). Nous chercherons également à définir une méthode d'évaluation de **chatbots** permettant de mesurer, d'une manière qui soit la plus objective possible, les capacités en pragmatique du langage des agents conversationnels artificiels. Dans cette seconde partie, nous nous intéresserons en particulier aux méthodes d'évaluation actuelle des **chatbots** et au Test de Turing (Turing, 1950), avant d'apprécier la capacité de ce test à mesurer précisément certains facteurs pragmatiques. Nous utiliserons pour cela la mesure des temps de réponse comme un indicateur du coût cognitif lié au traitement des messages reçus par les participants. Après avoir montré la pertinence de cette approche nous présenterons également

les opportunités qu’offrent le paradigme de l’utilisation de [chatbots](#) et du Test de Turing comme outils en psychologie expérimentale. Dans cette dernière partie nous porterons notre attention sur deux expériences tirant partie des [chatbots](#) et du Test de Turing pour évaluer l’importance de différents facteurs pragmatiques dans deux tâches distinctes : le jeu de l’ultimatum (Güth et al., [1982](#)) d’une part, avec des [chatbots](#) dont les personnalités sont différentes afin d’observer si ces personnalités influent sur le raisonnement, et des conversations avec des textismes (langage texto) d’autre part pour observer le coût cognitif impliqué par l’utilisation de ceux-ci au cours d’une conversation.

Abstract

With the fast development of automated systems allowing to communicate directly with users or clients with the use of natural language, we cannot help but wonder about these devices' capacity to answer the user in a relevant way. Indeed, numerous studies indicate that users of such systems often report a certain frustration due to their inability to make themselves understood by the machine, which often ends up providing them with incoherent, irrelevant replies, especially in conversations going beyond simple question-answer settings. In this thesis, we will study the different problems regarding conversational pragmatics which is often ignored by these automated tools such as [chatbots](#), which are virtual conversational agents, and social robots, which are embodied conversational agents, and we will try to provide answers to these problems by using previous research done by Grice (1975) on conversational pragmatics and by Wilson et Sperber (2002) on Relevance Theory. In a first chapter we will start with observations allowing us to claim that many problems exist regarding conversational pragmatics and in particular the importance of context and implicit information in conversations with artificial agents. Indeed they generally cannot sustain long conversations but remain rather decent when they only need to deal with question-answer scenarios. Following these observations we will suggest a different theoretical model to develop such agents allowing to start solving these problems by using aspects described by the Theory of Mind (Premack & Woodruff, 1978). We will also be attempting to find an evaluation method for [chatbots](#) allowing to measure, as objectively as possible, the pragmatic capacities of artificial conversational agents. In this second part we will be especially interested in the evaluation methods of [chatbots](#) and in the Turing Test (Turing, 1950), before evaluating the ability of this last test to precisely measure some pragmatic factors. To do this we will measure the response times as an indicator of the cognitive cost required to process the messages received by our participants. After showing the relevance of this approach we will also show the opportunities given by the paradigm of using [chatbots](#) and the Turing Test in experimental psychology. In this last part we will be focusing on two ex-

periments making use of [chatbots](#) and of the Turing Test to evaluate the importance of some pragmatic factors in two distinct tasks : the ultimatum game (Güth et al., [1982](#)) first, with [chatbots](#) of different personalities to observe whether they have an influence on reasoning, and then conversations with textisms (sms language) in a second time to observe the cognitive cost linked to the processing of messages containing alternative forms of spelling in a conversation.

Table des matières

Table des matières	6
Liste des tableaux	8
Table des figures	9
Introduction Générale	13
Agents conversationnels et pragmatique	13
Problématique	15
Plan de la thèse	15
I La pragmatique conversationnelle des agents artificiels	19
1 Sur le manque de traitement pragmatique des chatbots	20
1.1 Introduction	20
1.2 La pragmatique dans les chatbots	22
1.3 La pragmatique chez les robots sociaux	25
1.4 Conclusion	27
2 Des chatbots qui lisent dans les pensées : nous n’y sommes pas encore	28
2.1 Introduction	28
2.2 Revue de la littérature	29
2.3 Expérience	31
2.4 Conclusion	36
3 Vers un modèle pragmatique d’un interlocuteur artificiel : ouvrir la boîte noire	37
3.1 Introduction	38
3.2 Revue de la littérature	40
3.3 Modèle	44
3.4 Conclusion	50
II Évaluer les chatbots avec le Test de Turing	52
4 Sur la pragmatique du Test de Turing	53
4.1 Introduction	53
4.2 Les chatbots	54
4.3 Évaluation des chatbots	58

4.4	Le test de Turing comme test d'humanité	61
4.5	Comprendre les humains	68
4.6	Conclusion	73
5	La coopération dans les conversations humaines : Les temps de réponse comme fenêtre sur le traitement du langage	74
5.1	Introduction	74
5.2	Revue de la littérature	78
5.3	Expérience	91
5.4	Discussion	104
5.5	Conclusion	112
6	L'impact des maximes de Grice de qualité, quantité et de manière dans les chatbots	114
6.1	Introduction	114
6.2	Revue de la littérature	116
6.3	Expérience	124
6.4	Discussion	133
6.5	Conclusion	136
III	Utiliser les chatbots et le Test de Turing en psychologie	138
7	Un chatbot égoïste ne remporte pas le jeu de l'ultimatum	139
7.1	Introduction	139
7.2	Expérience	143
7.3	Discussion	148
7.4	Conclusion	149
8	Les informations contextuelles aident à comprendre les messages écrits en textismes	151
8.1	Introduction	151
8.2	Revue de la littérature	153
8.3	Expérience	158
8.4	Resultats	162
8.5	Discussion	163
8.6	Conclusion	166
	Conclusion Générale	168
	Vers des agents conversationnels pragmatiques	168
	Travaux futurs	174
	Bibliographie	178

Liste des tableaux

2.1	Performance de 5 chatbots avec du contenu implicite	34
4.1	Faux négatifs avec BLEU ou METEOR	60
5.1	Effet attendu des maximes de Grice dans le Test de Turing (partie 1)	95
5.2	Effet de violations des maximes de Grice sur le temps de réponse (partie 1) .	100
5.3	Effet des maximes de Grice dans le Test de Turing (partie 1)	103
6.1	Effet attendu des maximes de Grice dans le Test de Turing (partie 2)	120
6.2	Effet de violations des maximes de Grice sur le temps de réponse (partie 2) .	129
6.3	Effet des maximes de Grice dans le Test de Turing (partie 2)	131

Table des figures

3.1	Modèle de base d'un chatbot pragmatique	45
4.1	Exemple d'une conversation avec ELIZA	55
4.2	Exemple d'une conversation avec Kuki.ai	57
5.1	Effet des violations de maximes de Grice sur le temps de réponse (partie 1) .	102
6.1	Effet des violations de maximes de Grice sur le temps de réponse (partie 2) .	125
7.1	Comportement des participants dans le jeu de l'ultimatum selon qu'ils intéragissent avec un humain ou avec un chatbot	147
8.1	Influence des textismes sur le temps de réponse dans le test de Turing	162

Glossaire

chatbot Un programme informatique conçu pour répondre à l'utilisateur entrant du texte en langage naturel par du texte en langage naturel dans le but de converser avec. 2–5, 13, 15–17, 20–25, 28, 29, 31–36, 38, 39, 42, 43, 45, 50, 54–62, 65, 68–73, 75, 77, 88, 89, 114, 115, 119–123, 125, 126, 137, 141, 143–150, 159, 160, 168–174, 176, 177

magicien d'Oz Contexte expérimental dans lequel l'expérimentateur contrôle directement les actions ou paroles d'un outil. Cela peut être une marionnette, un robot, ou comme dans cette thèse, des chatbots. Le terme est repris du roman « Le Magicien d'Oz », dans lequel le magicien est en fait un imposteur. 116, 122, 159

maxime Une maxime est une idée, un précepte, une règle. Dans le contexte des maximes de Grice, elles correspondent surtout à des attentes entre participants. 17, 30, 40–43, 65, 66, 70–72, 75, 77, 80–85, 87–89, 91, 92, 94, 96–98, 101–106, 108, 109, 111–113, 115–119, 121–125, 127–137, 141, 153, 154

Acronymes

AIML Artificial Intelligence Markup Language. 56

BRNN réseau de neurones récurrents bilatéraux, ou Bilateral Recurrent Neural Network. 57

CLN compréhension du langage naturel. 20, 46

CNTN réseau de tenseurs neuraux à convolution, ou Convolutional Neural Tensor Network. 57

HTM mémoire hiérarchique temporelle, ou Hierarchical Temporal Memory. 169, 174, 176

IA intelligence artificielle. 16, 20, 22, 50, 53, 59, 62, 67, 68, 74, 83, 86, 87, 89, 92–100, 102, 103, 106–111, 125, 160, 169, 174, 175, 177

LSTM mémoire à long court-terme, ou Long Short-Term Memory. 57

PC principe de coopération. 40, 65, 80, 84

PLN production du langage naturel. 20, 48

RNN réseau de neurones récurrents, ou Recurrent Neural Network. 57, 60, 175, 176

ToM Theorie de l'Esprit, ou Theory of Mind. 14, 16, 29, 66, 175

TR temps de réponse. 77, 90, 92, 93, 96–98, 101, 102, 104–108, 111–113, 170

TT test de Turing. 16, 17, 25, 42, 54, 58, 59, 61, 62, 64, 67–73, 77, 86–88, 93, 95–98, 103, 108, 112, 113, 115, 116, 121, 125, 126, 129, 131, 134, 135, 141, 160, 163, 169–173, 177

Introduction Générale

Agents conversationnels et pragmatique

Lorsque l'on parle d'agents conversationnels, on désigne toute entité capable de converser avec une autre entité par le biais du langage naturel : le langage que nous utilisons tous les jours pour communiquer entre humains. Ces agents conversationnels peuvent être de différents types. Il existe bien sûr les agents conversationnels naturels que sont les humains, mais aussi les agents conversationnels artificiels : des machines simulant le comportement humain en conversation. Ces agents conversationnels peuvent être de deux types : les agents conversationnels virtuels ([chatbots](#)) qui ne disposent pas d'un corps et ne sont qu'un programme installé sur un ordinateur standard, et les agents conversationnels incarnés qui eux disposent d'un corps et sont souvent appelés robots sociaux.

Ces différents outils existent dans le but d'interagir avec l'humain, de communiquer avec pour échanger des informations. La méthode préférentielle utilisée dans ce but est de communiquer en utilisant le langage naturel qui a de nombreux avantages. En effet celui-ci est intuitif pour les utilisateurs qui l'utilisent déjà tous les jours avec d'autres humains. Il ne requiert donc *a priori* pas d'apprentissage pour pouvoir interagir avec l'entité artificielle. De nombreux [chatbots](#) existent désormais tels qu'ELIZA (Weizenbaum, 1966), A.L.I.C.E. (Wallace, 2009), Mitsuku, maintenant appelée Kuki.ai, et bien d'autres. Ces [chatbots](#) peuvent aussi bien avoir été développés dans un but de recherche comme pour des raisons ludiques, et sont de plus en plus utilisés à des fins thérapeutiques (Comme

Woebot pour aider les personnes souffrantes d'anxiété ou de dépression), ou commerciales (IBM Watson).

Cependant, comprendre les productions langagières de l'utilisateur s'est révélé être une tâche très complexe. Non seulement il est nécessaire pour la machine de comprendre le sens des mots du langage (aspect sémantique), mais il est également nécessaire de saisir et d'utiliser tout ce qui n'est pas dit directement dans les productions de l'utilisateur : l'implicite, à partir du contexte, dans le but de tenir compte des effets pragmatiques.

En effet, comme l'ont indiqué plusieurs auteurs (Ducrot, 1972 ; Grice, 1975 ; Wilson & Sperber, 2002), ce qui est *dit* n'est pas nécessairement ce que l'on *veut dire*. En communiquant avec d'autres interlocuteurs nous participons à une sorte de jeu au cours duquel nous tentons de lire l'esprit de l'autre pour tenter de répondre non pas à ce que cet autre interlocuteur nous dit, mais en fait à ce qu'il veut nous dire. Ce jeu fait appel à une capacité appelée [Théorie de l'Esprit, ou Theory of Mind \(ToM\)](#) décrite par Premack et Woodruff (1978) qui consiste en le fait d'imaginer ou d'inférer l'état mental d'un autre que soi. Chercher l'intention derrière un énoncé est bien sûr nécessaire dans des conversations, mais les inférences résultantes de cette recherche d'intention, dépendantes du contexte, ont également une influence dans de nombreuses tâches de raisonnement étudiées en psychologie (voir Bagassi & Macchi, 2006 ; Baratgin, 2002, 2009 ; Baratgin & Noveck, 2000 ; Baratgin & Politzer, 2006, 2007, 2010 ; Dulany & Hilton, 1991 ; Macchi, 2000 ; Macchi & Bagassi, 2012 ; Macchi et al., 2020 ; Macchi et al., 2019 ; Politzer & Macchi, 2000 ; Sperber et al., 1995, pour des exemples).

C'est cet aspect qui donne toute sa difficulté au problème des conversations. En effet nous ne savons pas à l'heure actuelle à quoi est due cette [ToM](#), quels sont ses substrats biologiques. Il est donc difficile de répliquer cette capacité dans des machines qui doivent donc s'en passer. Le fait qu'un aspect central de la communication soit absent provoque toutes sortes de problèmes que nous explorerons dans la première partie de cette thèse pour proposer une manière de commencer à la mettre en œuvre, avant de continuer dans la deuxième partie sur l'évaluation d'agents artificiels, et enfin terminer sur les opportunités

d'utilisation de ceux-ci dans des paradigmes expérimentaux de psychologie. Le tout afin de répondre à la problématique ci-dessous.

Problématique

Au cours de cette thèse, nous chercherons à répondre à la problématique suivante : Comment évaluer la qualité de la pragmatique conversationnelle d'agents artificiels ?

Pour répondre à cette question nous chercherons dans un premier temps à établir le constat que les [chatbots](#) actuels ne tiennent que très peu compte de la pragmatique du langage, avant de proposer une méthode d'évaluation, puis nous verrons que celle-ci peut être utilisée afin de répondre à d'autres questions.

Cette thèse est faite d'articles indépendants rassemblés par thème, dans un ordre permettant de traiter la problématique précédente.

Plan de la thèse

Dans la première partie intitulée « La pragmatique conversationnelle des agents artificiels », nous explorerons l'état actuel de l'usage de la pragmatique faite par des [chatbots](#) et autres agents conversationnels artificiels. Dans un premier chapitre nous brosserons un état des utilisations des agents conversationnels artificiels et noterons les problèmes qui en découlent, tels que les questions de frustration des utilisateurs lors de l'utilisation de tels agents, comme décrit notamment dans l'étude de Chaves et Gerosa (2020). Nous parlerons également des avancées qui sont faites, aussi bien dans les [chatbots](#) que pour les robots sociaux, comme le fait de tenir compte du regard de l'utilisateur ou celui du robot social, ou encore de la politesse au cours des conversations.

Le deuxième chapitre de la thèse présentera une étude que nous avons réalisée

afin de définir à quel point les **chatbots** sont, ou non, capables de répondre à des messages implicites. En participant à quelques conversations sur des sujets de conversations généraux, nous verrons que les **chatbots** programmés par mot-clés ne permettent pas de traiter ce genre de requête, et par conséquent n'offrent pas la possibilité de soutenir de conversations longues.

Dans le troisième chapitre de la thèse nous détaillerons notre proposition d'un modèle de **chatbot**. Il tient davantage compte de l'état mental de l'utilisateur en faisant référence aux travaux sur la **ToM** (Premack & Woodruff, 1978). Nous verrons que la plupart des **chatbots**, y compris ceux capables d'apprendre par entraînement (grâce aux réseaux de neurones notamment) ne s'appuient que très peu compte de cet aspect. Ils s'inspirent plutôt des techniques utilisées dans la traduction, ne permettent pas d'obtenir des résultats satisfaisants pour participer à de véritables conversations.

Dans la deuxième partie de la thèse, intitulée « Evaluer les **chatbots** avec le Test de Turing », nous chercherons à montrer que le **test de Turing (TT)** (Turing, 1950) peut être un outil particulièrement efficace afin de détecter les problèmes liés à la pragmatique du langage des **chatbots**. En effet, dans le premier chapitre de cette partie nous verrons d'abord en quoi le **TT** garde toute son importance dans l'évaluation des **chatbots**. En effet, à ce jour, malgré les nombreuses critiques qui s'opposent à son utilisation dans le domaine de l'**IA**, lui seul permet d'évaluer efficacement la qualité de la pragmatique conversationnelle d'un agent artificiel.

Le deuxième chapitre de cette partie montrera que le **TT** peut être enrichi, notamment en enregistrant les temps de réponse au cours de la conversation. Ces mesures permettent d'avoir plus de précisions sur le comportement de l'utilisateur au cours des conversations et est très cohérente avec le résultat final du **TT**. Ce protocole permet surtout de mettre en lien le **TT** avec le principe de coopération de Grice (1975) et la théorie de la pertinence de Wilson et Sperber (2002). En effet, les temps de réponse observés semblent être tout à fait cohérent avec les prédictions de chacune de ces deux visions de la pragmatique conversationnelle, en particulier ils peuvent sans doute être un bon indice

du coût cognitif lié au traitement des messages.

Le troisième chapitre est la continuation logique du précédent et vise à compléter les observations effectuées sur 3 sous *maximes* de Grice (1975) en en explorant 3 autres. Si le chapitre précédent portait sur la première *maxime* de quantité, la *maxime* de relation et la troisième *maxime* de manière, ce chapitre s'intéresse plutôt à la deuxième *maxime* de quantité, la *maxime* de qualité et la quatrième *maxime* de manière.

Fort de nos observations dans la deuxième partie, nous chercherons ensuite, dans la troisième partie : « Utiliser les *chatbots* et le test de Turing en psychologie », à montrer l'intérêt de l'utilisation de *chatbots* et du *TT* non pas pour le développement de ceux-ci, mais plutôt pour répondre à des questions sur le fonctionnement ordinaire humain. En effet, dans la deuxième partie nous avons observé qu'il était possible de prétendre utiliser des *chatbots* pour en réalité évaluer le comportement humain en conversation. Ceux-ci peuvent également être utilisés dans d'autres domaines, notamment dans le domaine du raisonnement économique comme avec le jeu de l'ultimatum¹ (Güth et al., 1982). Dans le premier chapitre de cette partie nous montrerons que les participants peuvent avoir des comportements plus tranchés avec des *chatbots*, notamment car ils sont moins soumis à des questions de désirabilité sociale.

Dans le deuxième chapitre de cette partie nous utiliserons le contexte du *TT* (Turing, 1950) de manière similaire aux études menées dans la deuxième partie, mais cette fois nous nous détacherons des violations des *maximes* de Grice (1975) pures pour appliquer notre réflexion sur la question des manières de communication textuelles alternatives au langage traditionnel, en étudiant l'impact des textismes (langage texto, dont les travaux de Panckhurst (2009) et de Fairon et al. (2006) ont permis de recenser de nombreux aspects intéressants) sur les participants, et en particulier en observant leur effet sur le

1. Le jeu de l'ultimatum est un jeu dans lequel deux participants doivent partager entre 0 et 10 objets. L'un des deux joueurs (A) propose un partage, et l'autre joueur (B) doit l'accepter ou le refuser. Si B refuse, aucun partage n'est fait et les deux joueurs perdent la totalité des objets. Rationnellement, B devrait accepter toute offre au-dessus de 0, et A devrait maximiser ses gains et choisir donc de ne partager la somme minimale acceptable par B, donc 1. Dans la réalité le partage est généralement plus égalitaire, et un partage en dessous de 3 est fréquemment refusé par B.

coût cognitif nécessaire à la compréhension des messages.

Enfin, nous concluons cette thèse en revenant sur chacun des enseignements que nous avons pu tirer de chacune de ces parties, et proposerons des pistes afin de les développer dans le futur.

Première partie

La pragmatique conversationnelle des agents artificiels

Publications

Le contenu de cette partie a donné lieu à trois publications :

1. On the lack of pragmatic processing in artificial conversational agents, (Jacquet, Masson et al., [2019](#)).
2. Mind-Reading chatbots : We are not there yet, (Jacquet & Baratgin, [2020](#)).
3. Towards a pragmatic model of an artificial conversational partner : opening the blackbox, (Jacquet & Baratgin, [2019](#)).

Chapitre 1

Sur le manque de traitement pragmatique des chatbots

Points remarquables dans ce chapitre

1. La pragmatique conversationnelle est nécessaire, y compris dans les conversations textuelles.
2. Malgré l'avancée des systèmes d'IA, les machines sociales sont encore très limitées dans ce domaine.
3. Ce manque produit des problèmes d'engagement et de frustration des utilisateurs.

1.1 Introduction

Les agents conversationnels artificiels, tels que les [chatbots](#) ou les robots sociaux, sont parmi les applications qui posent les plus grands défis à l'IA. En effet, ils reposent à la fois sur la [compréhension du langage naturel \(CLN\)](#) et la [production du langage naturel \(PLN\)](#), deux domaines dans lesquels les programmes récents d'IA rencontrent encore de nombreuses imperfections, malgré les réelles avancées de la dernière décennie. Alors que les systèmes d'IA ont considérablement amélioré la qualité du traitement syntaxique et du

traitement sémantique, la prise en compte du traitement pragmatique reste pratiquement absent malgré son importance considérable dans les conversations humaines (Wilson & Sperber, 2002). Dans cet état de l'art, nous viserons à décrire les améliorations les plus récentes dans les domaines de la compréhension du langage naturel et de la production de langage naturel, ainsi que les défis à relever en matière de traitement pragmatique. Nous suggérons également différentes manières d'étudier la pragmatique des entités sociales (humaines ou artificielles), afin de fournir des outils pour améliorer les systèmes actuels sur cet aspect bien spécifique. Les humains ont en effet une capacité remarquable lorsqu'il s'agit de distinguer une machine d'un interlocuteur humain dans une conversation. Ils utilisent généralement tous les indices qu'ils peuvent trouver pour juger de l'humanité de leur partenaire (Jacquet et al., 2018, 2019; Saygin & Cicekli, 2002). La pragmatique repose largement sur des informations contextuelles. Son utilisation correcte reste une tâche difficile pour les agents artificiels. Alors que leur capacité à utiliser les informations syntaxiques et sémantiques augmentent, la principale différence entre les humains et les agents artificiels repose sur leur utilisation des indices conversationnels, comme le ton de la conversation, l'humeur de l'interlocuteur (humain), ce qu'ils pensent que ce dernier devrait savoir ou ne pas savoir. Tous ces indices ont une importance majeure avec les robots sociaux. Dans leur cas, ils devraient être capables de comprendre ce que leur partenaire conversationnel humain peut voir ou ne pas voir de manière beaucoup plus explicite qu'un [chatbot](#). En effet, si le traitement pragmatique est d'une grande importance dans les conversations écrites, il devient absolument essentiel dans les conversations parlées. Il implique beaucoup plus d'aspects comme la modulation du ton de la voix, les gestes, le regard... Ces indices tellement essentiels, en fait, font que certains éléments ont même fait leur entrée dans les conversations écrites, comme les smileys mais également certains des signes comme le point d'interrogation ou d'exclamation. L'absence de point final pour signifier la fin de la phrase est utilisée pour éviter que le contenu soit perçu par le lecteur comme trop abrupt, trop brutal. Des expériences récentes dans le domaine de la psychologie cognitive soulignent en effet l'importance du traitement pragmatique, notamment en ce qui concerne la pertinence de chaque énoncé dans les conversations

entre humains, mais aussi entre humains et agents artificiels tels que les **chatbots** et les robots sociaux (Chakrabarti & Luger, 2015). Tous ces indices, l'importance des gestes, de la modulation de la voix, des expressions faciales (ou de leurs émulations) permettent dans les interactions avec les robots sociaux de susciter la motivation, l'implication, la coopération ou le fait de relever des défis chez le partenaire humain (Gockley et al., 2005 ; Ham et al., 2011 ; Masson, Baratgin & Jamet, 2017a ; Masson et al., 2016, 2017b ; Mutlu et al., 2009).

1.2 La pragmatique dans les chatbots

Les **chatbots** ont été l'une des premières manifestations de l'IA. Ils restent dans l'imaginaire collectif du public ce à quoi une IA devrait ressembler. La principale raison en est qu'il s'agit du type d'IA le plus visible et le plus facilement accessible. Il existe également depuis des décennies, le premier programme de ce type étant ELIZA (Weizenbaum, 1966) qui utilise des mots-clés dans les propos de l'utilisateur pour répondre par des questions ouvertes afin de poursuivre la conversation. Ce traitement très sommaire des énoncés dans une conversation est toujours au cœur de **chatbots** plus complexes, comme A.L.I.C.E. (Wallace, 2009) et Elizabeth¹. Si ces IA ne peuvent prétendre être de véritables partenaires conversationnels en raison de leurs limites, ils peuvent néanmoins rester très utiles dans certains contextes très spécifiques, comme dans le cas de Woebot². Woebot un **chatbot** qui se présente comme un coach. Sa fonction consiste à prodiguer de l'aide à des utilisateurs afin qu'ils puissent mieux gérer leur anxiété, des états dépressifs, la procrastination etc... Woebot n'est en effet pas en mesure de participer à des conversations ouvertes. Les réponses qu'il est en mesure de fournir à l'utilisateur sont dans la plupart des cas prédéfinies, et n'implique donc aucune reconnaissance de texte. Tess³ est également conçu pour coacher des utilisateurs en état de stress, anxieux, en dépression, en situation d'épuisement professionnel (entre autres). Il permet aux utilisateurs d'interagir

1. Elizabeth for Windows. <http://www.philocomp.net/ai/elizabeth.htm>

2. Woebot Labs Inc. <https://woebot.io>

3. X2AI Inc. <http://x2ai.com/>

avec lui en langage naturel. Récemment, les progrès des réseaux neuronaux ont permis de développer des programmes capables de produire des réponses beaucoup plus ouvertes à tout sujet de conversation, comme Zo⁴ qui est aujourd’hui l’un des **chatbots** les plus avancés. Zo est cependant également loin d’être parfait en ce qui concerne le suivi d’une conversation. A titre d’exemple,

Utilisateur : Aimes-tu la plage ?

Zo : Je déteste la plage à cause du sable.

Utilisateur : Ouais, le sable est embêtant.

Zo : Tu es embêtant.⁵

Ce court extrait de l’échange entre un utilisateur et Zo montre que le **chatbot** a certainement perdu le contexte de la conversation. Il le fera très souvent, en fait, à moins que le contexte réel ne soit répété dans presque chaque énoncé. Ce manque de contextualisation est assez fréquent dans les **chatbots**. Si la plupart des **chatbots** répondent de manière pertinente, ils deviennent rapidement incapables de répondre de manière satisfaisante lorsque l’utilisateur fait référence à un message précédent, ou au contexte global de la conversation, sans le formuler explicitement dans son énoncé. Ce détail est loin d’être anecdotique, puisque cette pratique est assez souvent utilisée dans les conversations humaines, par exemple :

A1 : Avez-vous de la crème glacée ?

B1 : Oui, j’en ai !

A2 : J’en voudrais bien.

Il est évident qu’en A2, la personne fait référence à la crème glacée. Pourtant, nombreux sont les **chatbots** actuels qui seront dans l’incapacité de fournir une réponse à cet

4. Microsoft Inc. <https://www.zo.ai/>

5. Cette conversation est tirée d’un message public sur Twitter, disponible sur <https://twitter.com/zochats/status/1009141014827761664>. Il est possible de discuter avec le bot directement sur la même page, par le biais de messages privés.

utilisateur, autre que générique du type : « Moi aussi ». C'est là l'un des principaux problèmes en pragmatique auquel les **chatbots** d'aujourd'hui sont confrontés : la pertinence de leurs énoncés dans le contexte de la conversation. En effet, si une réponse générique peut tromper les utilisateurs en leur faisant croire que leur partenaire est humain, cela ne fonctionne pas aussi bien lorsque cela se produit avec presque chaque énoncé. A.L.I.C.E (Wallace, 2009) peut donner l'illusion qu'il comprend le contexte pour des questions simples, comme par exemple lorsque l'on parle d'enfants et que l'on demande « J'en ai deux », le robot répondra « Comment s'appellent-ils ? », ce qui est une réponse pertinente et attendue. Le problème survient lorsqu'il est clairement visible qu'il ne s'agit pas d'une réponse contextualisée, mais bien d'une réponse basée sur des mots-clés. Lorsque le sujet de la conversation passe des enfants aux ordinateurs. Dans ce contexte, lorsque l'utilisateur dit : « J'en ai deux », le **chatbot** répond également par : « Comment s'appellent-ils ? », ce qui n'est certainement pas une réponse pertinente et attendue. Ce manque de contextualisation rend les **chatbots** actuels plus aptes à répondre à des séries de questions qu'à être de véritables agents conversationnels, notamment dans le contexte d'un service au client (Zhou et al., 2017), comme dans l'exemple de conversation suivant :

Client : Je ne parviens pas à me connecter à mon compte. J'ai oublié mon mot de passe.

Chatbot : Veuillez cliquer sur le lien « Mot de passe oublié ». Vous recevrez par e-mail un lien qui vous aidera à changer votre mot de passe.

Client : J'ai déjà essayé. Ce lien m'amène à un formulaire pour réinitialiser le mot de passe. Mais il me demande d'entrer mon numéro de relation client. Je n'ai pas ce numéro.

Chatbot : Le numéro de relation client se trouve dans l'onglet « Informations sur le compte » lorsque vous vous connectez à votre compte.

Client : Mais je ne parviens pas à me connecter à mon compte. Je ne me souviens pas de mon mot de passe.

Chatbot : Veuillez cliquer sur le lien « Mot de passe oublié ». Vous recevrez par e-mail un lien qui vous aidera à changer votre mot de passe.

Développer la capacité des [chatbots](#) à utiliser la pragmatique conversationnelle semble donc être l'un des objectifs les plus importants à atteindre avant que ces programmes puissent être utilisés de manière fiable comme agents conversationnels, en particulier dans le domaine des affaires.

Il est souvent difficile et laborieux de tester correctement la compétence pragmatique des [chatbots](#). En effet, de mauvaises compétences pragmatiques peuvent produire un sentiment désagréable, une gêne chez le lecteur qui, s'il n'est pas conscient de l'aspect artificiel de l'un des participants à la conversation, pourrait croire qu'il est un malade mental (Saygin & Cicekli, 2002). Une façon d'aborder l'évaluation des [chatbots](#) peut être d'utiliser le cadre du TT auquel on ajoute l'enregistrement des temps de réponse du juge humain pour évaluer le [chatbot](#), ceux-ci augmentant significativement suite à une erreur pragmatique du [chatbot](#), comme nous le verrons dans la partie 2 de cette thèse (Jacquet, Baratgin & Jamet, 2019; Jacquet, Hullin et al., 2019a).

1.3 La pragmatique chez les robots sociaux

Les robots sociaux vont certainement être de plus en plus présents dans notre vie quotidienne, mais plus particulièrement dans celle des personnes âgées pour leur apporter de l'aide, mais aussi dans le domaine du tutorat (voir Broekens et al., 2009, pour une revue). Les [chatbots](#) ont à la fois besoin de disposer de capacité à traiter des indices pragmatiques, mais également d'être en mesure de produire des réponses intégrant ces indices en plus grand nombre. Par exemple, être capable de déduire l'humeur de l'utilisateur à travers le ton de sa voix ou ses expressions faciales. De tels indices contribuent au contexte de la conversation de la même manière que les énoncés. Afin de fournir une interaction la plus naturelle et durable avec un robot, il est nécessaire de maintenir un bon niveau de motivation chez la personne en lui montrant que ses interactions sont importantes pour la conversation elle-même. On constate que lorsque la conversation est perçue comme étant un monologue et plus une conversation, un dialogue,

la personne s'en désintéresse (Gockley et al., 2005). Il faut également tenir compte du fait qu'il ne suffit pas de comprendre les indices pragmatiques, mais qu'il faut également être capable de les produire. D'autant plus si l'objectif est de permettre aux utilisateurs d'interagir de manière naturelle. En effet, leur neutralisation dans le comportement du robot produit un obstacle à l'anthropomorphisation du robot. Il ne sera plus perçu comme un agent social, mais comme une simple machine (Masson, Baratgin & Jamet, 2017a; Masson et al., 2016, 2017b).

Lors des conversations en face à face, le regard est également un outil de communication non verbale. Le développement d'un robot social doit prendre en compte les différents messages que le regard peut transmettre pour éviter de produire des effets secondaires indésirables. En effet, si un utilisateur est reconnu par le robot à travers son regard, il aura tendance à communiquer davantage, à avoir plus d'affection envers lui, et aussi à lui prêter davantage d'attention et ce indépendamment du propos du robot (Mutlu et al., 2009).

D'autres aspects contribuent à l'efficacité d'une conversation avec un robot comme la prise en compte des normes sociales. En effet une entité sociale applique les normes sociales comme saluer son interlocuteur, engager une avec lui une conversation, respecter les tours de parole, poursuivre le sens de la conversation, connaître son interlocuteur, et enfin de gérer correctement la fin des conversations en faisant preuve de politesse (Gockley et al., 2005).

Lorsque l'on considère les expressions faciales qui se manifestent au cours d'une conversation, il est important de comprendre qu'elles constituent un outil de communication qui n'est pas indépendant de la conversation à laquelle elles appartiennent. En effet, comprendre l'état thymique que représente une expression faciale peut être difficile pour certaines expressions ambiguës. Les humains, pour réduire ces ambiguïtés, utilisent différents indices issus du contexte verbale de la conversation comme les interjections (Motley, 1993).

Enfin, les gestes sont un autre facteur important dans les conversations humaines, en particulier pour le pouvoir de persuasion. Là encore, ils ne sont pas indépendants d'autres indices, comme le regard. En effet, Ham et al. (2011) ont montré que les gestes augmentent le pouvoir de persuasion d'un agent artificiel uniquement lorsqu'ils sont combinés à un suivi correct du regard, et qu'ils le diminuent dans le cas contraire, même si, dans tous les cas, les gestes donnent une impression plus réaliste aux utilisateurs (Ham et al., 2011).

1.4 Conclusion

Nous pensons que les développements dans le domaine de l'intelligence artificielle, appliquée aux conversations naturelles, devraient absolument envisager d'évaluer leurs agents sur leur capacité à utiliser et à produire des indices pragmatiques chaque fois que leur objectif est de converser avec des humains. D'autant plus, s'ils veulent être reconnus comme des agents sociaux, et pas simplement comme des répondeurs. Ceci est particulièrement important s'ils doivent être utilisés pour remplacer une véritable interaction humaine dans le contexte des affaires, car ils doivent fournir la même qualité d'interaction qu'un client pourrait attendre, mais aussi dans le contexte des soins, où une mauvaise communication pourrait entraîner des conséquences sur la santé de l'utilisateur.

Chapitre 2

Des chatbots qui lisent dans les pensées : nous n'y sommes pas encore

Points remarquables dans ce chapitre

1. Nous proposons une expérience pour tester la capacité de 5 chatbots à comprendre des questions implicites.
2. Nous remarquons que ces chatbots ne sont pas capables de répondre de manière pertinente.
3. Ces chatbots avaient également des difficultés à répondre à des messages explicites.
4. Aucun chatbot n'a réussi la tâche demandée.

2.1 Introduction

Beaucoup d'entre nous ont interagi avec des [chatbots](#) au moins une fois dans leur vie, et si ce n'est pas le cas, la probabilité que cela se produise va certainement augmenter mois après mois, car ils se répandent dans les services à la clientèle, en particulier sur les sites Web des entreprises (par exemple, Booking.com, pratiquement tous les fournisseurs d'accès Internet français, Domino's, et bien d'autres). Ils sont souvent utilisés comme un

premier filtre, pour répondre à des questions courantes qui ne nécessitent pas un examen approfondi de la situation de l'utilisateur, et en cela, ils fonctionnent assez bien.

En effet, leur capacité à parler à de nombreuses personnes en même temps, alors que les humains ne peuvent prendre part qu'à un petit nombre de conversations simultanées, en fait une alternative fantastiquement efficace aux humains. Ils ne dorment pas, ne mangent pas, n'ont pas besoin de faire de pause et ne se plaignent certainement pas d'être moins bien payés.

Mais cette capacité à gérer une grande quantité de messages entrants a un prix : ils sont comme des coquilles vides qui ne prennent pas véritablement en compte le contexte d'une conversation. Pour être plus explicite, ils n'ont pas la moindre idée de ce que l'utilisateur pourrait penser, ou de ce qu'il pourrait attendre. Cette capacité cognitive est désignée dans les sciences cognitives comme la [Théorie de l'Esprit, ou Theory of Mind \(ToM\)](#) (Premack & Woodruff, 1978).

Dans cet article, nous passerons brièvement en revue l'abondante littérature dans le domaine de la pragmatique conversationnelle et de la théorie de l'esprit pour expliquer pourquoi elles sont importantes pour comprendre comment nous interagissons les uns avec les autres dans le contexte des conversations, avant de discuter des études qui ont montré des problèmes avec les [chatbots](#) actuels. Nous présenterons également des recherches qui ont montré l'importance du contexte dans les conversations en ligne afin de répondre aux attentes de l'utilisateur. Enfin, nous détaillerons une expérience conçue pour étudier la façon dont les [chatbots](#) réagissent aux informations implicites dans les énoncés et nous discuterons de ses résultats.

2.2 Revue de la littérature

La littérature scientifique est riche en articles tant sur les conversations entre humains que sur la façon dont ils communiquent entre eux. Parmi les auteurs qui se sont

intéressés à ce domaine, Grice (1975) a été l'un des premiers à proposer une théorie visant à expliquer comment les personnes discutent dans les conversations. Il suggère qu'elles coopèrent naturellement afin de communiquer, et à partir de là, il a défini quatre *maximes*, qui représentent quatre types d'attentes lorsqu'elles conversent avec d'autres afin d'extraire ce qui est signifié derrière de ce qui est dit. La première *maxime*, la *maxime* de qualité, se réfère à la vérité de l'affirmation prononcée, du moins pour la personne qui l'exprime. La deuxième *maxime*, la *maxime* de quantité, se réfère à la quantité d'information fournie par l'énoncé : elle doit être équilibrée, et ne fournir ni trop ni trop peu d'information. La troisième *maxime* est la *maxime* de relation, qui fait référence à la pertinence de l'énoncé, compte tenu de son contexte. Enfin, la quatrième, la *maxime* de la manière, fait référence à la façon dont l'information est donnée.

Une autre théorie « concurrente »¹ est la théorie de la pertinence (Wilson & Sperber, 2002). Elle suggère que les principes qui sous-tendent les conversations sont encore plus généraux. Cette théorie considère que la pertinence est l'aspect le plus important des conversations, et en général du partage d'informations pendant la communication. Si une information est donnée à quelqu'un, celui-ci s'attend à ce qu'elle soit pertinente d'une manière ou d'une autre. Pour cette raison, les gens essaieront souvent de trouver un moyen de rendre les énoncés qui leur sont adressés en quelque sorte subjectivement pertinents, même si l'énoncé était en fait assez générique et aurait pu être donné à quelqu'un d'autre dans un contexte différent. Néanmoins, cela demande un certain effort à la personne qui essaie d'interpréter la signification de l'énoncé dans son contexte. Elle devient également plus sujette aux interprétations, qui peuvent changer le sens perçu, par rapport à ce que l'énonciateur voulait vraiment dire. Par conséquent, ces phrases sont moins pertinentes en soi que les phrases qui ne demandent pas beaucoup d'efforts pour en déduire le sens, compte tenu du contexte. La théorie de la pertinence postule donc qu'un énoncé a une pertinence optimale lorsque l'effort requis pour déduire sa signification est faible et que l'effet de cette signification sur l'autre personne qui l'entend est élevé (en d'autres termes,

1. La théorie de la pertinence est basée sur le principe de coopération et les maximes de Grice, mais elle en propose une généralisation.

si elle modifie davantage ce que la personne pense pour un effort moindre).

Ces deux théories reposent sur le fait que les participants à une conversation tentent d'inférer l'intention de leur interlocuteur, en préparant leurs énoncés de manière que l'interprétation de ceux-ci par cet interlocuteur soit aussi facile et efficace que possible. Pour ce faire, tous les participants à une conversation doivent utiliser leur théorie de l'esprit (Premack & Woodruff, 1978), décrite comme la capacité à déduire les états mentaux des autres, leurs intentions, leurs attentes, leurs croyances, leurs connaissances... essentiellement, à déduire l'expérience du monde d'un autre, distincte de la nôtre.

En effet, Saygin et Cicekli (2002) ont montré que le manque de pertinence dans les conversations diminue considérablement le sentiment d'humanité donné lors d'une conversation, y compris lorsque les interlocuteurs sont bien humains, comme nous le verrons dans la deuxième partie de cette thèse (Jacquet, Baratgin & Jamet, 2019; Jacquet, Hullin et al., 2019a). Lorsqu'un interlocuteur commence à manquer de pertinence, il est souvent perçu comme ennuyant, ressemblant à une machine ou simplement peu coopératif, voire comme ayant des problèmes mentaux (Saygin & Cicekli, 2002).

Dans l'expérience suivante, nous avons tenté de mesurer la pertinence de certains des énoncés les plus connus des [chatbots](#).

2.3 Expérience

Matériel et Méthodes

Cinq [chatbots](#) ont été utilisés pour les besoins de notre expérience. Ils ont été sélectionnés en raison de leur notoriété parmi les autres [chatbots](#), notamment pour avoir remporté le prix Loebner une ou plusieurs fois. Il s'agit : A.L.I.C.E, développé par Wallace (2009) qui a remporté le prix Loebner en 2000, 2001 et 2004; George (une version de Jabberwacky), développé par Rollo Carpenter, lauréat du prix Loebner en 2005, et à

nouveau en 2006 avec une version différente ; Joan. Elbot, développé par Fred Roberts, a remporté le prix Loebner en 2008. Enfin, Mitsuku, développé par Stever Worswick, a remporté le prix en 2013, 2016, 2017, 2018 et 2019. Tous sont des **chatbots** construits autour du design de *pattern-matching* (détection de motifs). En d'autres termes, la phrase envoyée par l'utilisateur est comparée à un modèle prédéfini, puis une réponse est renvoyée en fonction de ce modèle².

Tous les **chatbots** ont eu la possibilité de s'exprimer dans 8 sujets de conversations différents et pour chaque sujet dans deux conditions : 1) implicite, la conversation faisait un usage intensif de structures ambiguës de phrases qui ne pouvaient être désambiguïsées qu'à partir du contexte de la conversation (c'est-à-dire à partir des messages précédents). Le sujet était énoncé une fois et ensuite il n'était pas répété. Par exemple, l'utilisation de pronoms comme « il » était privilégiée au lieu de répéter ce à quoi le pronom faisait référence. 2) explicite, la conversation répétait le sujet de la conversation plusieurs fois, donnant toutes les informations nécessaires pour fournir une réponse pertinente dans l'énoncé lui-même. Ainsi, le **chatbot** n'avait pas besoin de se souvenir du contexte de la conversation pour répondre de manière pertinente au message de l'utilisateur.

Les principales hypothèses étaient les suivantes : 1) les **chatbots** obtiendraient de moins bons résultats dans la condition implicite que dans la condition explicite, 2) les plus récents obtiendraient de meilleurs résultats, en particulier dans la condition explicite.

Mesures

Les critères qui permettent de définir l'efficacité des **chatbots** dans les conversations étaient de deux types : 1) les deux questions de l'utilisateur ont-elles reçu une réponse, et 2) combien de temps (compté comme le nombre de messages) ont-ils réussi à maintenir la conversation avant de perdre complètement leur pertinence. Pour chaque conversation, un

2. Il existe d'autres types de **chatbots** qui utilisent des conceptions différentes, notamment les **chatbots** à apprentissage profond ou d'autres types d'apprentissage automatique. Malheureusement, la plupart d'entre eux ne sont pas accessibles au public pour converser avec eux.

chatbot pouvait obtenir jusqu'à 2 points, en fonction du nombre de questions auxquelles il avait répondu de manière pertinente. Les deux questions posées par l'utilisateur étaient toujours du même type. La première posait une question précise, définissant le sujet. La deuxième question permettait de demander plus d'informations sur la réponse du **chatbot** à la première question. Ces points étaient enregistrés dans la variable des réponses et la moyenne était calculée sur l'ensemble des conversations du **chatbot**. La durée de la conversation avant la perte de pertinence a été enregistrée dans la variable longueur. Le compte a commencé à augmenter après la réponse à la deuxième question.

Sujets de conversations

Les huit sujets ont été sélectionnés de telle sorte qu'un humain dans une situation similaire aurait été capable de converser sur le domaine sans aucune expertise. Chacun d'entre eux comportait une première question pour poser le sujet, et une seconde pour l'approfondir.

Il s'agissait des questions suivantes : 1) Quel est votre dinosaure préféré ? Qu'est-ce que tu aimes chez lui ? 2) Où aimerais-tu vivre ? Pourquoi ? 3) Que penses-tu de Michael Jackson ? Pourquoi penses-tu cela ? 4) Que penses-tu qu'il y ait après la mort ? Tu peux développer ? 5) Quelle est ta couleur préférée ? Pourquoi ? 6) Penses-tu que l'intelligence artificielle va envahir le monde ? Pourquoi penses-tu cela ? 7) Qu'aimerais-tu visiter si tu pouvais aller en France ? Pourquoi irais-tu là-bas ? 8) Que penses-tu des voitures électriques ? Pourquoi penses-tu cela ?

Dans chaque domaine, il n'y avait pas de réponse spécifique attendue, car de nombreuses réponses auraient pu être acceptables. Au contraire, le point était accordé si la réponse était suffisamment précise pour ne pas être considérée comme une réponse générique qui aurait pu être donnée à n'importe quelle autre question.

Résultats

Les résultats sont présentés dans le tableau 2.1. Seuls quelques **chatbots** ont été capables de répondre à plus d'une question, et seulement sur certains sujets spécifiques, comme ceux relatif à la couleur préférée, d'expliquer pourquoi. En dehors de cela, les réponses perdaient de leur pertinence immédiatement après la première réponse, lorsqu'elles pouvaient répondre à cette dernière.

Si les **chatbots** ont obtenu des résultats légèrement meilleurs dans la condition explicite, il ne s'agit que d'une augmentation très marginale des performances.

Discussion

Comme nous l'avons vu dans les résultats de notre expérience, les **chatbots** sont peu opérant lorsqu'ils doivent répondre à des questions qui nécessitent une compréhension du contexte de la question. En effet, tous les **chatbots** utilisés dans notre expérience sont programmés par pattern-matching. Ils opèrent à partir de la phrase d'entrée envoyée par l'utilisateur. Ce type de programmation ne permet pas de garder une trace des informations contextuelles. Pourtant, en raison de sa facilité de programmation, cette

Chatbot	Implicite		Explicite	
	Réponses	Longueur	Réponses	Longueur
A.L.I.C.E	0,375	0	0,5	0,125
George	0,875	0,125	0,875	0,125
Joan	0,5	0	0,625	0,125
Elbot	0,25	0	0,25	0
Mitsuku	0,625	0,125	0,625	0,125

TABLE 2.1 – **Performance des chatbots en fonction de la condition (Implicite ou Explicite)**. Réponses indique la moyenne (tout sujet de conversation confondu) du nombre de questions auquel le chatbot a répondu de manière pertinente (entre 0 et 2). Longueur indique le nombre moyen de messages pertinents échangés sur le même sujet après une deuxième réponse correcte (plus ce nombre est élevé, plus le chatbot est capable de maintenir une conversation longtemps).

méthode reste la plus largement utilisée aujourd'hui. Une autre chose que nous avons remarquée dans notre expérience est que les **chatbots** récents ne sont pas beaucoup plus performants que les plus anciens. En effet, les scores restent inférieurs à 1 pour tous les **chatbots**. Un **chatbot** comme George, un des plus anciens atteint une performance assez similaire à un **chatbot** beaucoup plus récent comme Mitsuku en termes de pertinence. Bien sûr, les **chatbots** se sont améliorés sur d'autres aspects. Par exemple, Mitsuku répond souvent par des images en plus de ses réponses textuelles. Malheureusement, cela ne rend pas ses réponses plus pertinentes. Il est intéressant de noter que pour certains domaines spécifiques, comme demander la couleur préférée du **chatbot**, certains **chatbots** ont fait beaucoup mieux. En effet, voici un extrait de la conversation avec Mitsuku dans la condition implicite :

- 1) **Utilisateur** : Quelle est votre couleur préférée ?
- 2) **Mitsuku** : Ma couleur préférée est le bleu.
- 3) **Utilisateur** : Pourquoi ?
- 4) **Mitsuku** : Je trouve que le bleu est une couleur très relaxante.
- 5) **Utilisateur** : Je serai curieux d'entendre pourquoi vous pensez cela.
- 6) **Mitsuku** : Je n'y ai pas beaucoup réfléchi, donc je n'ai pas de bonne réponse à vous donner.

Dans sa réponse 4), Mitsuku peut continuer à parler de la couleur bleue, comme si le **chatbot** se souvenait du sujet de la conversation. Parmi tous les sujets de conversation, c'est le seul qui est dans ce cas. On peut penser que ce comportement a été préprogrammé pour cette séquence spécifique de questions. En outre, il y avait très peu de différence entre les conditions implicites et explicites. C'est assez surprenant, car il s'est avéré que les **chatbots** étaient incapables de répondre à des questions demandant des plus de détails, même en fournissant explicitement une bonne quantité d'information relative au contexte pour les aider.

- 1) **User** : Quel est ton dinosaure préféré ?

- 2) **George** : Raptor.
- 3) **User** : Pourquoi est-ce que tu aimes les raptors ?
- 4) **George** : J'aime tout ce qui concerne les Beatles. Leur musique, leur style de vie, leurs opinions, etc..

2.4 Conclusion

Bien que les [chatbots](#) se soient améliorés au cours des dernières années, il est indéniable qu'ils restent assez peu pertinents lorsqu'on leur demande de répondre à des questions et d'élaborer leurs propres réponses. Les séquences de messages doivent être préprogrammées pour que le contexte de la conversation puisse être utilisé dans la réponse produite, ce qui constitue un facteur très limitant pour une utilisation à grande échelle. En outre, lorsque l'on demande plus de détails, même en étant très explicite, la plupart des [chatbots](#) ne parviennent pas à fournir une réponse pertinente.

Dans notre étude, nous nous sommes uniquement intéressés aux [chatbots](#) ayant remporté le prix Loebner. Or, ces [chatbots](#) ont tendance à être d'un type particulier : ceux qui répondent à des modèles. Il serait intéressant de voir comment ils se comparent aux [chatbots](#) qui utilisent différentes méthodes pour générer leurs réponses, en particulier ceux qui utilisent l'apprentissage profond, ou les techniques d'apprentissage automatique en général.

Dans tous les cas, il est probable que la construction d'un [chatbot](#) qui pourrait être aussi pertinent qu'un interlocuteur humain nécessiterait une architecture très différente de celles que nous avons testées. En particulier, il existe un grand besoin d'implémenter dans les [chatbots](#) un modèle qui s'appuierait sur les états mentaux de l'utilisateur, comme nous allons le voir dans le chapitre suivant (Jacquet & Baratgin, 2020).

Chapitre 3

Vers un modèle pragmatique d'un interlocuteur artificiel : ouvrir la boîte noire

Points remarquables dans ce chapitre

1. Nous montrons l'importance de tenir compte des attentes de l'utilisateur pour les chatbots.
2. Nous proposons un modèle fonctionnel de chatbot tenant compte des aspects pragmatiques des conversations.
3. Des travaux restent à mener pour expliciter le fonctionnement de certaines unités de traitement de l'information.
4. Ce modèle devrait être implémenter pour observer s'il permet d'améliorer les performances des chatbots.

3.1 Introduction

L'intérêt pour les [chatbots](#) n'a cessé de croître ces dernières années (voir Dahiya, 2017 ; M. Qiu et al., 2017, pour des exemples), notamment dans les entreprises (voir Heo & Lee, 2018 ; Thomas, 2016 ; Zhou et al., 2017, pour des exemples). Cet intérêt découle du fait que l'interaction avec les clients semble se déplacer vers les systèmes mobiles (Faulds et al., 2018). En conséquence, on s'attend à ce que les utilisateurs aient besoin de différentes interfaces plus adaptées à une expérience mobile, y compris l'utilisation d'agents conversationnels (Zhou et al., 2017) et, plus spécifiquement, d'agents conversationnels textuels comme les [chatbots](#).

En effet, avoir une conversation textuelle sur un appareil mobile est une expérience très courante pour l'utilisateur et semble assez intuitive. De nombreux défis subsistent quant à l'efficacité de ces systèmes à fournir des informations pertinentes à l'utilisateur et à gérer ses attentes (Chaves & Gerosa, 2020). En particulier, la frustration causée par l'incapacité des [chatbots](#) à déduire le véritable sens de l'énoncé de l'utilisateur, ce qui oblige ce dernier à s'adapter à lui plutôt que l'inverse.

Si la qualité de la syntaxe des [chatbots](#) est aujourd'hui quasi parfaite. Si leurs capacités sémantiques se sont également grandement améliorées (notamment pour répondre à des questions simples qui ne nécessitent pas de comprendre le contexte dans lequel elles se posent). En revanche, leur capacité à traiter des énoncés contextualisés au sein d'une conversation est encore loin d'être proche de ce que les humains peuvent attendre d'un partenaire conversationnel, notamment en ce qui concerne la manière dont les partenaires conversationnels font référence à des informations antérieures données plus tôt dans la conversation comme nous l'avons vu dans le chapitre précédent (Jacquet, Masson et al., 2019).

Par exemple, examinons cet extrait d'une conversation avec le [chatbot](#) A.L.I.C.E de pandorabots.com :

Utilisateur : Aimez-vous les enfants ?

A.L.I.C.E : Oui, ils sont l'avenir.

Utilisateur : J'en ai deux.

A.L.I.C.E : Comment s'appellent-ils ?

[...]

Utilisateur : Tu aimes les ordinateur ?

A.L.I.C.E : Oui, je vis dans un ordinateur, donc je dois les aimer.

Utilisateur : J'en ai deux.

A.L.I.C.E : Comment s'appellent-ils ?

Il est clair ici, qu'A.L.I.C.E. ne tient pas vraiment compte du contexte, mais traite uniquement le texte phrase par phrase. Bien que cet exemple puisse sembler un peu caricatural, de tels événements se produisent assez souvent dans les conversations naturelles, car les utilisateurs ne sont pas habitués à devoir conserver dans leurs messages toutes les informations contextuelles nécessaires au traitement de leurs énoncés. Considérons maintenant ce qui se passe si nous incluons l'information contextuelle (que nous voulons dire : J'ai deux [ordinateurs]) dans l'énoncé :

Utilisateur : J'ai deux ordinateurs.

A.L.I.C.E : Parle-moi d'eux.

Bien que la réponse ici puisse être générique, elle est au moins crédible.

Afin d'améliorer la capacité des [chatbots](#) à utiliser les informations contextuelles, nous suggérons dans cet article un modèle théorique élaboré sur la base d'un système de traitement des conversations textuelles prenant en compte les informations contextuelles et les attentes des utilisateurs.

3.2 Revue de la littérature

La pragmatique dans les conversations humaines

Les humains n'utilisent pas la définition des mots qu'ils connaissent dans le langage quotidien. Au contraire, les définitions des mots que nous utilisons ont tendance à varier fortement en fonction du contexte dans lequel ils sont utilisés. Ce principe ne s'applique pas seulement aux mots, mais aussi à des expressions entières. Par conséquent, il est extrêmement important de faire la distinction entre ce qui est dit et ce qui est signifié afin de développer notre compréhension du processus de langage humain.

Coopération

L'un des premiers auteurs à avoir suggéré cette distinction au sein des conversations est Grice (1975). Il suggère que tous les participants à une conversation attendent de leur interlocuteur qu'il coopère avec eux afin de communiquer. En effet, une communication efficace est une tâche complexe qui exige que les participants coopèrent s'ils veulent être sûrs que ce qu'ils veulent dire est compris.

Grice (1975) a suggéré que tous les participants à une conversation suivent un **principe de coopération (PC)** implicite, qui est défini dans les termes suivants :

Apportez votre contribution telle qu'elle est requise, au stade où elle se présente, en suivant l'objectif et l'orientation de la conversation.¹

Ce principe général est subdivisé en quatre super **maximes**, qui peuvent être considérées comme des catégories d'attentes. Ces **maximes** sont la **maxime** de la qualité, la **maxime** de la quantité, la **maxime** de la relation et la **maxime** de la manière. Cela implique

1. « Make your contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged. »

que le locuteur possède un modèle implicite de ce que l'auditeur s'attend à entendre, soit pour se conformer à ces attentes, soit pour utiliser volontairement ces attentes afin de fournir implicitement des informations supplémentaires.

Pertinence

Wilson et Sperber (2002) ont par la suite tenté de généraliser les différents *maximes* proposés par Grice (1975) en une seule théorie : La théorie de la pertinence. Ils suggèrent que tous les énoncés passent en fait par un processus de calcul d'effet/coût qui définit leur pertinence. Plus un énoncé a d'effet sur les représentations mentales de la cible, dans un certain contexte, plus il est pertinent. D'autre part, plus le coût cognitif (l'effort) nécessaire pour traiter l'énoncé est élevé, moins il est pertinent.

Selon cette théorie, si un énoncé contient une information que l'auditeur connaît déjà, la phrase n'est pas pertinente, car elle n'a aucun effet sur ses représentations mentales et l'effort nécessaire pour la traiter a été gaspillé.

De même, un énoncé contenant trop d'éléments informatifs inconnus de l'auditeur peut également devenir non pertinent, car s'il a potentiellement un effet important sur les représentations mentales de l'utilisateur, l'effort nécessaire pour le traiter augmente aussi considérablement. Il faut que l'auditeur soit particulièrement motivé pour extraire et traiter toutes ces informations. Il y a fort à parier qu'il n'en traitera qu'une partie.

Par conséquent, la pertinence optimale est atteinte lorsqu'un énoncé contient juste la quantité suffisante de nouvelles informations pour l'auditeur, tout en maintenant le coût cognitif nécessaire pour les traiter aussi bas que possible.

La pragmatique des agents conversationnels artificiels

La qualité du traitement de la pragmatique chez les partenaires conversationnels est encore loin d'atteindre le niveau attendu par les humains, comme nous l'avons vu dans

le chapitre précédent, bien que des données expérimentales montrent l'importance de ce type de traitement dans la capacité d'une machine à établir des interactions coopératives avec l'utilisateur (Masson, Baratgin & Jamet, 2017a; Masson et al., 2015, 2017b). Ces articles montrent que, lorsque suffisamment d'indices pragmatiques sont donnés à l'utilisateur sur la volonté de prendre part à une interaction sociale, l'utilisateur la considérera effectivement comme une entité sociale et suivra des « biais » similaires à ceux observés dans les interactions entre humains. Dans le protocole expérimental utilisé dans ces études (le jeu de l'ultimatum), le biais consiste à être assez égalitaire dans la quantité de monnaie qu'ils prendront pour eux-mêmes et celle qu'ils laisseront au robot. Ceci n'est pas observable sans ces indices pragmatiques.

On a également constaté que les *chatbots* sont davantage perçus comme des « machines » lorsqu'ils n'utilisent pas correctement les indices pragmatiques (Saygin & Cicekli, 2002). En effet, dans cet article, les auteurs étudient l'influence des violations des *maximes* de Grice (1975) sur les participants à un protocole inspiré du *TT* (Turing, 1950), dans lequel ils devaient deviner lequel des deux interlocuteurs était un humain et lequel était une machine, à partir des extraits de conversations. Ils ont trouvé le *chatbot* beaucoup plus facilement lorsqu'il violait la *maxime* de relation de Grice (« Soyez pertinent ») en particulier, mais aussi pour d'autres *maximes*, comme la deuxième *maxime* de quantité (« Ne rendez pas votre contribution plus informative que nécessaire »). Il est intéressant de noter que lorsque les participants ne savaient pas qu'une intelligence artificielle était derrière, ils voyaient quand même la différence par rapport à leurs attentes, et certains ont réagi en disant des choses comme « Sont-ils des malades mentaux ? ».

Des conclusions similaires ont été tirées d'expériences visant à mesurer le coût cognitif (effort), tel que défini par la théorie de la pertinence (Wilson & Sperber, 2002), relatif à la compréhension d'une phrase et à la production de sa réponse lors de conversations interactives en ligne, par l'enregistrement des temps de réponse des participants (Jacquet et al., 2018, 2019; Jacquet, Hullin et al., 2019a).

Dans ces expériences, il fallait en moyenne quelques secondes de plus pour répondre

lors de la violation de la **maxime** de relation de Grice qu’aux énoncés sans violation, même en tenant compte de la longueur des énoncés écrits (Jacquet et al., 2018, 2019). Cela peut s’expliquer par le temps nécessaire pour 1) comprendre que l’énoncé n’est pas pertinent pour le sujet actuel de la conversation, car cela est supposé être vrai par défaut et doit donc être inhibé dans de tels cas (Grice, 1975; Wilson & Sperber, 2002), 2) considérer le nouveau sujet de conversation, et 3) écrire une réponse à un énoncé inattendu et surprenant.

Une autre expérience suivant le même plan expérimental a montré un résultat similaire lorsque le **chatbot** produit par écrit trop d’informations qui ne sont pas directement nécessaires dans son énoncé (violation de la deuxième **maxime** de quantité de Grice) (Jacquet, Hullin et al., 2019a). Bien que de telles violations n’aient pas diminué de manière significative le sentiment d’humanité donné par le **chatbot**², le fait d’écrire des éléments d’information non demandés a augmenté le temps de réponse des participants, ce qui peut être interprété comme un effort supplémentaire pour traiter l’énoncé.

Agents conversationnels actuels

À l’heure actuelle, la plupart des **chatbots** utilisent des modèles *Seq2Seq* (modèles faisant correspondre une séquence d’entrée à des séquences de sortie) pour répondre à la phrase d’entrée de l’utilisateur. Ce type de système fait directement correspondre une représentation vectorielle de la phrase d’entrée (une séquence) à différents vecteurs de sortie représentant des phrases de sortie potentielles (d’autres séquences), sans faire aucun traitement supplémentaire des attentes de l’utilisateur concernant l’utilisation du contexte et, en général, des indices pragmatiques, ou du moins, des indices limités (Cui et al., 2017; Nayak et al., 2017).

Ces modèles utilisent traditionnellement des variantes de réseaux neuronaux pour

2. Certainement du fait que les éléments d’information supplémentaires pouvaient être interprétés simplement comme offrant des sujets de conversation plus diversifiés, comme lorsque l’on demande quel est le travail du **chatbot**, en ne répondant pas seulement à la question, mais en fournissant également le travail de la femme du personnage joué par le **chatbot**.

apprendre la correspondance entre les séquences d'entrée et les séquences de sortie, ainsi que la grammaire et la syntaxe du langage. Ils utilisent également un système d'attention, qui améliore considérablement la capacité à utiliser les informations contextuelles des messages précédents (Mei et al., 2017).

Il existe des modèles hybrides qui ajoutent un intermédiaire dans le processus d'appariement d'une séquence à une autre, en utilisant parfois des technologies différentes, comme les graphes, pour relier non seulement des séquences à d'autres séquences, mais aussi des intentions à des séquences (Arora et al., 2019).

Bien que les modèles Seq2Seq soient intéressants pour leur capacité à apprendre le comportement correct à partir des données d'entrée, ils ne tiennent pas compte de l'état mental de l'utilisateur. Le modèle que nous proposons vise à se concentrer sur cet aspect.

3.3 Modèle

Le modèle que nous proposons d'utiliser est construit autour de deux types de composants (voir Figure 3.1). D'un côté, nous avons les modèles internes du système. Ces modèles sont divisés en deux systèmes : Le modèle de l'état du monde, et le modèle de l'état de l'utilisateur. Ces deux systèmes contiennent des hypothèses, des croyances, la première sur le monde en général, et la seconde sur l'utilisateur avec lequel le système est en train d'interagir.

De l'autre côté, nous avons les processeurs. Ces unités de traitement sont l'encodeur de phrases, le processeur d'information et le décodeur de phrases. Chacun de ces systèmes transforme les informations d'entrée qu'il reçoit, en fonction des informations contextuelles extraites des différents modèles.

L'encodeur de phrases

L'encodeur de phrases est le point d'entrée du système. Toute phrase écrite par l'utilisateur sera d'abord traitée par cette unité afin d'extraire le contenu sémantique (ce que l'utilisateur voulait dire) de la phrase (ce que l'utilisateur a dit). Cette extraction du contenu sémantique est réalisée grâce aux informations stockées dans le modèle de l'utilisateur et dans modèle de l'état du monde. C'est le premier effet pragmatique. Cette unité produit des prédictions sur le type d'entrée que le système s'attend à recevoir à partir de l'état de l'utilisateur modélisé dans le modèle de l'utilisateur et des « croyances » du [chatbot](#) sur le monde modélisées dans le modèle de l'état du monde. Si la phrase reçue en entrée est incompatible avec cette prédiction, le modèle de l'utilisateur est mis à jour et la phrase est interprétée dans son état décontextualisé (ou des traitements plus longs peuvent être effectués pour rechercher l'interprétation la plus probable). C'est

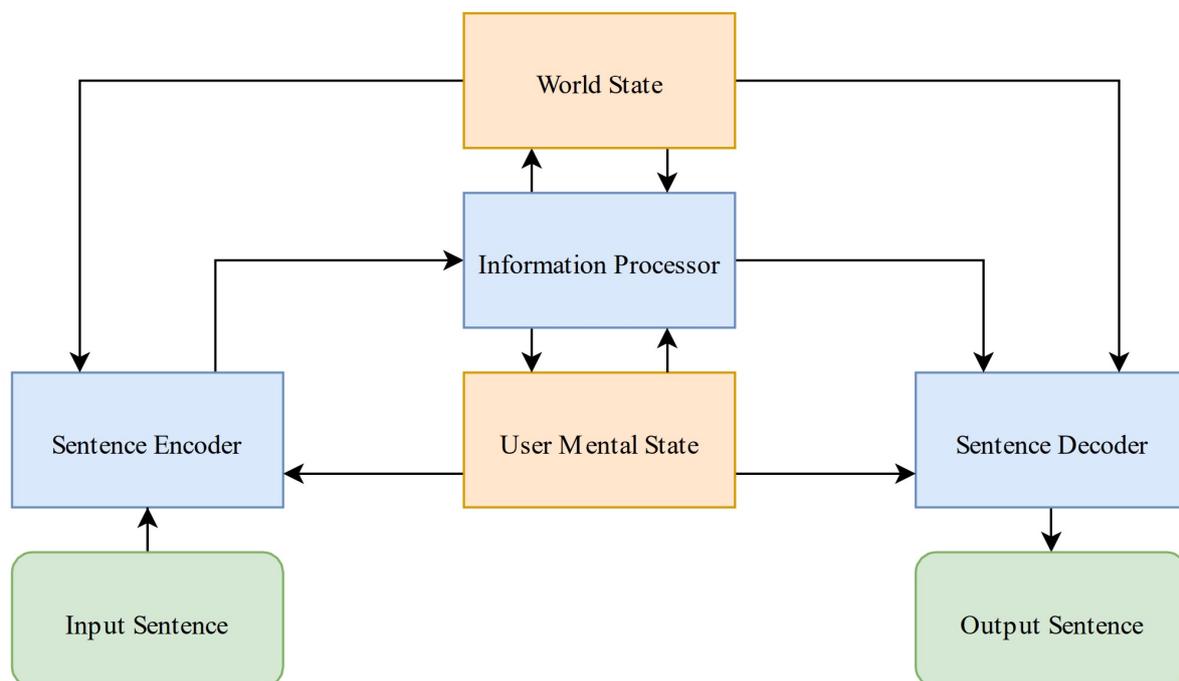


FIGURE 3.1 – **Modèle de système conversationnel que nous proposons.** Les boîtes oranges représentent la mise en mémoire de différents modèles (Modèle de l'utilisateur et modèle du monde). Les boîtes bleues représentent les modules de traitement de l'information, et les boîtes vertes représentent l'entrée et la sortie de l'information du système.

également l'encodeur de phrases qui décide de traiter ou non la phrase en entrée. Par exemple, si la quantité de mots reconnaissables n'est pas suffisante, ou si elle est dans une langue inconnue. Il doit être tolérant au bruit (comme les fautes de frappe, les caractères spéciaux, les mots inconnus), car la capacité de lecture humaine est également tolérante au bruit (Goodman, 1967).

$$SE_{\text{out}} = f(\text{Input}, M_W, M_U) \quad (3.1)$$

Dans cette équation, SE_{out} représente la sortie de l'encodeur de phrases, Input l'entrée dans le système, M_W le modèle du monde et M_U le modèle de l'utilisateur.

La sortie de l'encodeur de phrase consisterait en une représentation purement sémantique du contenu de la phrase d'entrée, ou du moins ce qui peut être déduit compte tenu des attentes du système envers l'utilisateur. En d'autres termes, l'encodeur de phrases ne déduit pas simplement le sens d'une phrase. Il tente de déduire ce qu'un utilisateur spécifique essaie de transmettre à travers la phrase qu'il a utilisée. Il s'agit d'un processus de traduction, car l'entrée n'est pas de la même nature que la sortie (une phrase et une représentation sémantique, respectivement). Au niveau de l'implémentation, cela consiste en une unité de CLN biaisée vers les attentes du système concernant l'utilisateur stockées dans le modèle de l'utilisateur et vers les croyances du système concernant le monde stockées dans le modèle de l'état du monde.

Le processeur d'informations

Il s'agit certainement du système de traitement le plus complexe du modèle. Il prend en entrée le résultat du traitement de l'encodeur de phrases, en utilisant les connaissances (ou croyances) du système sur le monde pour générer une représentation purement sémantique de la réponse qui doit être envoyée à l'utilisateur. Si l'information fournie par l'utilisateur est nouvelle pour le système, il met à jour ses croyances sur le monde ou sur

l'utilisateur, selon le type d'information.

$$IP_{\text{out}} = f(SE_{\text{out}}, M_W, M_U) \quad (3.2)$$

Dans cette équation, IP_{out} représente la sortie du processeur d'informations.

La sortie du processeur d'informations devrait être de nature similaire à celle de l'encodeur de phrases, mais elle contient les informations sémantiques que le système veut envoyer à l'utilisateur, qui ne sont pas nécessairement dans un format lisible par l'homme. Il s'agit très probablement du plus complexe des trois systèmes de traitement, car il est essentiellement censé émuler différents processus cognitifs, y compris certains aspects de la pensée consciente. Ce rôle pourrait être assumé soit par un système expert (un programme conçu de manière procédurale), soit par un algorithme d'apprentissage (capable d'apprendre par lui-même).

Le décodeur de phrases

Le décodeur de phrases est l'opposé de l'encodeur de phrases. Il traduit une représentation purement sémantique de l'information devant être transmise (ce que l'on veut dire) en une phrase (ce qui est dit). Cette phrase ainsi générée peut ne pas contenir la totalité des informations contenues dans la représentation initiale. En effet, certaines de ces informations peuvent être peu pertinentes pour l'utilisateur qui pourrait déjà les connaître. En conséquence, la phrase produite ne contiendra que les informations qui ne sont pas déjà connues par l'utilisateur.

$$SD_{\text{out}} = f(IP_{\text{out}}, M_W, M_U) \quad (3.3)$$

Dans cette équation, SD_{out} correspond à la sortie du décodeur de phrases, qui est également la sortie du système global.

Le décodeur de phrases est, comme l'encodeur, une unité de traduction. Il transforme les représentations purement sémantiques en des phrases lisibles par un humain. Le décodeur de phrases ne fait pas que transformer le contenu sémantique en phrase néanmoins, car il utilise également le modèle de l'état mental de l'utilisateur pour tenir compte de ses attentes.

Au niveau de l'implémentation, cette unité de traitement est une unité de [PLN](#), produisant des phrases biaisées par les attentes générées par le modèle de l'utilisateur et par le modèle du monde afin que les phrases produites ne contiennent que ce qui est nécessaire à la compréhension du message.

Le modèle du monde

Cette unité est le lieu où est stocké le modèle du monde déduit par le système. Elle contient toutes les informations que le système a apprises au cours de sa « vie ». Son contenu peut être mis à jour par le processeur d'informations, qui lui envoie de nouvelles informations. Les informations qui y sont stockées peuvent également être récupérées par le processeur d'informations. Elle ne doit pas seulement contenir une liste de propriétés, mais aussi les liens entre chacune d'elles. Au niveau de la mise en œuvre, cette unité devrait être l'équivalent d'une mémoire de travail, d'une mémoire à court terme et d'une mémoire à long terme combinées.

Le modèle de l'état mental de l'utilisateur

Cette unité peut être considérée comme une extension plus spécifique du modèle du monde. La principale différence est que, contrairement à celui-ci, qui est général, ce modèle est spécifique à l'utilisateur. En d'autres termes, il contient ce que le système suppose être les croyances de l'utilisateur sur le monde. En conséquence, il ne devrait y avoir qu'un seul modèle pour chaque utilisateur. Il doit toujours correspondre étroitement au modèle du

monde du système. Par exemple, si le système dispose d'une certaine information stockée dans sa mémoire, il devrait être capable de décider si l'utilisateur est susceptible d'avoir une certaine connaissance de cette information spécifique ou non.

La manière dont le système produit de nouveaux modèles pour les utilisateurs qu'il ne connaît pas peut dépendre des hypothèses générales qu'il apprend. Par exemple, s'il apprend que la plupart des gens connaissent le concept A, tout nouvel utilisateur sera supposé le connaître également, à moins que la conversation avec lui ne devienne incompatible avec cette hypothèse, ce qui amènerait le système à le questionner directement sur sa connaissance de A, ou plus simplement à lui expliquer A. En termes d'implémentation, cela pourrait correspondre à rendre plus ou moins disponibles des éléments du modèle du monde en fonction de chaque utilisateur.

Discussion

Le modèle que nous proposons est encore loin d'être complet et reste en développement afin d'atteindre une structure progressivement plus détaillée d'un modèle cognitivement plausible d'un système de traitement des conversations.

Nous sommes tout à fait conscients que la distinction entre l'encodeur de phrases et le décodeur de phrases peut sembler arbitraire dans un modèle cognitivement plausible, car les études semblent montrer que la compréhension et l'expression du langage semblent utiliser les mêmes régions du cerveau pour obtenir des informations lexicales, grammaticales et syntaxiques (Menenti et al., 2011). Le modèle que nous décrivons ici est avant tout un modèle fonctionnel, indépendant de sa mise en œuvre.

De la même manière, la distinction entre le modèle du monde et le modèle de l'utilisateur pourrait également sembler arbitraire à certains lecteurs. En effet, dans le cerveau, le modèle de l'utilisateur serait probablement inclus dans le modèle du monde, peut-être en tant que modèles neuronaux uniques qui inhibent ou activent d'autres modèles neuronaux du modèle du monde pour représenter la connaissance d'un utilisateur

spécifique de certaines informations. Néanmoins, dans le cadre de cette première tentative de suggestion d'un modèle, nous pensons qu'il est très important de faire cette distinction pour souligner l'importance de la modélisation de l'utilisateur, car c'est dans l'interaction avec l'utilisateur que la plupart des effets pragmatiques se produisent dans les conversations.

Il convient également de noter que la manière dont les nouveaux modèles d'utilisateurs sont créés au sein du système participerait probablement à l'apparition de préjugés, tout comme ils existent chez les humains. En effet, si le système remarque que certains profils ont tendance à connaître certains concepts, il pourrait ensuite généraliser à tort que tout profil similaire aura également des connaissances sur ce concept. Bien que cela permette dans la plupart des cas une expérience plus personnalisée pour l'utilisateur, il ne faut pas oublier que le système peut parfois se tromper dans ses hypothèses. Par conséquent, il est extrêmement important de s'assurer que le système peut mettre à jour ses croyances sur l'utilisateur afin de contrebalancer cet aspect du traitement du système.

L'objectif principal de ce modèle est d'offrir un moyen d'explorer ce qui était auparavant considéré comme une boîte noire : l'interaction entre les attentes des partenaires conversationnels dans la façon dont ils produisent et interprètent les phrases.

3.4 Conclusion

Bien que ce type de modèle n'ait pas encore été testé, nous pensons qu'il pourrait être grandement bénéfique au domaine de l'IA appliquée au traitement des conversations. En effet, à notre connaissance, aucun système ne tente actuellement d'inclure la modélisation de l'état d'esprit de l'utilisateur dans le processus de génération et d'interprétation des phrases pour offrir une interaction personnalisée avec le [chatbot](#). En effet, tant que la pragmatique de la conversation ne sera pas prise en compte de manière approfondie, les [chatbots](#) ne seront jamais capables d'appréhender pleinement la complexité des conversations humaines, conduisant à une frustration accrue de l'utilisateur empêchant toute

coopération efficace entre les deux partenaires conversationnels, humain et artificiel.

Deuxième partie

Évaluer les chatbots avec le Test de Turing

Publications

Le contenu de cette partie a donné lieu à quatre publications :

1. On the pragmatics of the Turing Test (Jacquet, Jamet et al., [2021](#)).
2. The Gricean Maxims of Quantity and of Relation in the Turing Test (Jacquet et al., [2018](#)).
3. Cooperation in Online Conversations : The Response Times as a Window Into the Cognition of Language Processing (Jacquet, Baratgin & Jamet, [2019](#)).
4. The Impact of the Gricean Maxims of Quality, Quantity and Manner in Chatbots (Jacquet, Hullin et al., [2019a](#)).

Chapitre 4

Sur la pragmatique du Test de Turing

Points remarquables dans ce chapitre

1. Dans ce chapitre à visée épistémologique, nous revenons sur l'état actuel des chatbots.
2. Nous expliquons les méthodes d'évaluation couramment utilisées.
3. Nous détaillons l'importance du Test de Turing pour évaluer la ressemblance de ces chatbots avec le comportement humain, aussi bien comme méthode d'évaluation de chatbots que comme paradigme expérimental en Psychologie Cognitive.

4.1 Introduction

Avec les récents progrès de l'IA, l'idée que les machines pourraient tôt ou tard être capables de penser par elles-mêmes fait son chemin dans la population générale, aidée en cela par de nombreux films comme, entre autres, HER de Spike Jonze, Ex Machina d'Alex Garland ou Blade Runner de Ridley Scott. Mais où en sommes-nous vraiment ? Sommes-nous proches de ce point ? Une machine peut-elle vraiment penser ? Disposons-nous des outils nécessaires pour l'évaluer comme l'a affirmé Turing (1950) avec son célèbre jeu

d'imitation ou [test de Turing \(TT\)](#) ?

Cet article tente de passer en revue une partie de la littérature explorant le principal problème auquel les [chatbots](#) (les programmes qui peuvent être évalués par le [TT](#)) sont encore confrontés aujourd'hui : la pertinence, en d'autres termes, la capacité à produire des phrases qui tiennent compte des attentes des utilisateurs. En effet, la plupart de ces programmes ne parviennent toujours pas à soutenir des conversations de plus de quelques phrases successives sans s'appuyer sur des réponses génériques.

Nous décrirons d'abord les différents types de [chatbots](#) qui existent aujourd'hui et certaines des technologies qui les fondent, avant de passer aux méthodes utilisées pour les évaluer. Nous décrirons ensuite le [TT](#) ainsi que certaines de ses critiques et nous insisterons sur l'importance de la pragmatique conversationnelle. Enfin, nous passerons en revue une partie de la littérature explorant l'utilisation du [TT](#) comme paradigme expérimental pour étudier le raisonnement humain et la pragmatique conversationnelle.

4.2 Les chatbots

Les agents conversationnels virtuels, souvent appelés [chatbots](#) (ou chatterbots), sont issus d'une idée initialement suggérée par Turing dans son test d'intelligence (Turing, 1950). Il s'agit de machines qui traitent les messages textuels, tentent d'y trouver une réplique pertinente avant de la renvoyer à l'utilisateur sous forme textuelle.

Les [chatbots](#) sont tout sauf nouveaux, la première application répondant à cette définition est nommée ELIZA (Weizenbaum, 1966). Destiné à se comporter comme un psychologue en misant sur l'empathie et les répétitions, ce programme utilise des mots-clés dans les messages de l'utilisateur pour lui renvoyer des questions. La conversation est maintenue par le [chatbot](#) qui donne l'impression d'essayer de plonger dans l'esprit et la psyché de l'utilisateur. Ces opérations sont réalisées en analysant le contenu du texte d'entrée, en attribuant une valeur à chaque mot reconnu avant de réorganiser

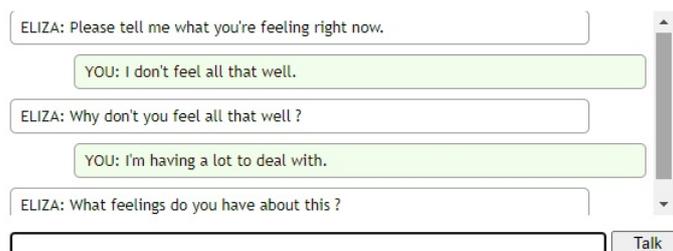


FIGURE 4.1 – Exemple d’une conversation que nous avons eue avec ELIZA. Cette version du [chatbot](https://www.eclecticenergies.com/psyche/eliza) peut être consultée à l’adresse suivante <https://www.eclecticenergies.com/psyche/eliza>

l’ordre des mots selon des modèles prédéfinis pour produire une réponse sous forme de question. Si le contenu de l’entrée ne correspond à aucun modèle reconnu, des réponses de repli prédéfinies sont renvoyées. ELIZA a été conçu à l’origine pour montrer à quel point les conversations entre humains et machines étaient superficielles. Pourtant, les utilisateurs ont fini par partager avec lui des aspects profonds et privés de leur vie, en laissant souvent berner et en refusant de croire qu’ils avaient parlé à une machine. Une conversation simple avec une version d’ELIZA est illustrée dans la Fig. 4.1.

A.L.I.C.E (Wallace, 2009) est un autre [chatbot](#) qui généralise le concept de *pattern-matching* pour envoyer des réponses à l’utilisateur sans nécessairement lui poser des questions en retour comme le ferait un psychologue. Tout comme ELIZA, il utilise toujours des mots-clés pour détecter des modèles dans l’entrée. Voici un exemple d’un échange simple avec A.L.I.C.E où un utilisateur lui demande sa couleur préférée.

Humain : Bonjour ALICE.

A.L.I.C.E : Bonjour !

Humain : Quelle est ta couleur préférée ?

A.L.I.C.E : Je n’ai pas de couleur préférée. Mais mon livre préféré est Ne Me Lis Pas.

Humain : Que peux tu me dire dessus ?

A.L.I.C.E : Ambigu : « peux tu me dire dessus » – colloquial ou « peux tu me dire ».

Nous pouvons voir ci-dessus qu’A.L.I.C.E. utilise également l’humour dans cette

conversation, ce qui le fait paraître un peu plus humain, bien que dans notre exemple, le programme ne parvienne pas à reconnaître l'entrée dans le dernier message envoyé par l'utilisateur. A.L.I.C.E. est construit autour de son propre langage appelé [Artificial Intelligence Markup Language \(AIML\)](#) qui est aujourd'hui utilisé par de nombreux [chatbots](#) pour sa simplicité, faisant ainsi d'A.L.I.C.E. un bon exemple de ce que l'on peut attendre des [chatbots](#) aujourd'hui.

Bien que la simplicité et la prévisibilité des [chatbots](#) à reconnaissance de motifs ¹ les rendent assez attrayants, même les plus avancés de ces systèmes ne peuvent pas (à ce jour) traiter le contenu implicite des messages d'entrée (Jacquet & Baratgin, 2020). En effet, comme nous pouvons le voir dans la conversation précédente avec A.L.I.C.E., il ne parvient pas à comprendre à quoi l'utilisateur fait référence dans la phrase : « Que pouvez-vous me dire à ce sujet ? ». Mitsuku (maintenant appelé Kuki.AI), le dernier [chatbot](#) de cette catégorie à avoir remporté le prix Loebner (concours international basé sur le TT pour évaluer les [chatbots](#)), est toujours incapable de répondre de manière pertinente au contenu implicite des messages, à moins que des cela ne soit préprogrammé (comme sur la couleur préférée, les animaux domestiques..., voir dans la Fig. 4.2) (Jacquet & Baratgin, 2020).

Le *pattern-matching* (détection de motifs) n'est cependant pas le seul outil disponible pour créer des [chatbots](#). Avec l'essor des techniques d'apprentissage automatique, y compris les algorithmes d'apprentissage supervisé et non supervisé (selon qu'ils utilisent ou non des données étiquetées par l'homme), de nouvelles opportunités de développement sont apparues. Ces nouvelles approches donnent aux [chatbots](#) la possibilité d'utiliser plus efficacement les indices contextuels. Malheureusement, aucun [chatbot](#) basé sur ces techniques n'est aujourd'hui librement accessible pour pouvoir tester ses capacités effectives bien que des plateformes existent pour en créer, comme IBM Watson, Amazon Lex, Facebook Messenger, Microsoft LUIS et d'autres. Les [chatbots](#) de ces plateformes utilisent

1. *Pattern-matching*, en anglais. Les phrases en entrée sont comparées à des modèles et reconnaissent les éléments communs. Si le motif/pattern dans la phrase est reconnu, alors la phrase réponse prédéfinie est renvoyée à l'utilisateur.

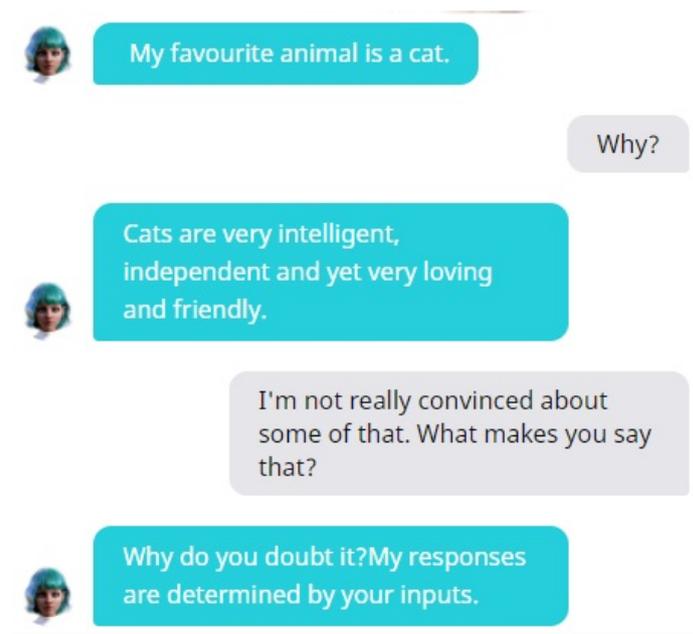


FIGURE 4.2 – Exemple d’une conversation que nous avons eue avec Kuki.ai. Remarquez comment il répond correctement à la première question contenant du contenu implicite, mais échoue à répondre à la deuxième. Cette version du [chatbot](https://chat.kuki.ai/) peut être consultée à l’adresse suivante <https://chat.kuki.ai/>

souvent des algorithmes tels que la LSTM (mémoire à long court-terme) (Hochreiter & Schmidhuber, 1997), les RNN (réseaux de neurones récurrents) (Archambault & Basano, 1994; M. Qiu et al., 2017), les BRNN (réseaux neuronaux récurrents bidirectionnels) (Dhyani & Kumar, 2021) ou les CNTN (réseaux de tenseurs neuronaux à convolution) (X. Qiu & Huang, 2015).

Il existe deux principaux types de [chatbots](#) basés sur ces techniques d’apprentissage automatique : les modèles *retrieval-based* (qui reposent sur la récupération d’informations stockées au préalable) et les modèles génératifs. Les premiers lisent la phrase d’entrée produite par l’utilisateur pour créer un vecteur de pensée qui représente le sens de la phrase (ou l’intention). Ce vecteur de pensée est ensuite comparé aux entrées d’une base de données contenant les réponses possibles. La réponse du [chatbot](#) sera constituée par l’entrée la plus proche du vecteur de pensée et sera alors envoyée à l’utilisateur. Les modèles génératifs, au contraire, utilisent le vecteur de pensée comme base pour générer de nouvelles phrases mot par mot. Les probabilités d’apparition des mots sont dépendantes à la fois des probabilités apprises dans la langue générale et du vecteur de pensée

qui a été créé lors de la lecture de la phrase d'entrée. Les modèles génératifs sont beaucoup plus flexibles que les modèles basés sur la recherche, car ils peuvent générer des phrases complètement nouvelles. En revanche, les modèles *retrieval-based* donnent plus de contrôle au propriétaire du [chatbot](#), car il peut décider précisément de ce qui sera dit et de ce qui ne le sera pas. Il existe également des modèles hybrides combinant ces deux aspects. Par exemple, des modèles qui tentent de générer des phrases mais qui, en cas d'échec, se rabattent sur des techniques de récupération (*retrieval*).

Les outils utilisés dans ces modèles se basent sur la traduction automatique en utilisant des systèmes d'encodage et de décodage pour prédire les mots suivants dans une séquence. Au lieu de traduire une phrase d'une langue à une autre, ils « traduisent » une phrase d'entrée en une réponse. En effet, de nombreux travaux ont déjà été réalisés dans le domaine de la traduction automatique et les outils utilisés dans ce domaine semblent donner des résultats généralement acceptables lorsqu'ils sont utilisés pour générer des réponses dans des conversations. Toutefois, ils n'atteignent pas encore le niveau humain et demeurent pour le moment insatisfaisant (Wei et al., 2019).

4.3 Évaluation des chatbots

L'évaluation de la qualité des [chatbots](#) reste un sujet plutôt controversé car aucune métrique standard ne répond aux trois critères suivants : automatisation, similarité avec le jugement humain et précision. Dans cette section, nous allons passer en revue certaines des méthodes qui peuvent être utilisées aujourd'hui pour évaluer les agents conversationnels.

Le [TT](#) reste une référence en la matière. En effet, la plupart des utilisateurs veulent avoir l'impression de converser aussi facilement avec le [chatbot](#) qu'avec un humain (Jain et al., 2018). Dans ce cas, on demande généralement aux utilisateurs d'évaluer dans quelle mesure la conversation leur semble humaine. Cette méthode présente des inconvénients lorsqu'il s'agit d'être automatisée et n'a pas une bonne précision si aucune mesure supplémentaire n'est ajoutée. En effet, si l'obtention d'une évaluation élevée du sentiment

d'humanité est l'objectif final, le fait de demander à l'utilisateur de ne se prononcer qu'une seule fois à la fin de la conversation ne donne pas une bonne indication du moment où des anomalies ont été détectées au cours de celle-ci.

D'autres mesures automatisées sont souvent utilisées : le taux d'achèvement de la tâche, qui est particulièrement utile pour les [chatbots](#) orientés vers un objectif et qui tentent d'aider les utilisateurs dans une tâche spécifique. Cette mesure ne peut toutefois pas être appliquée aux robots à usage général ou aux robots de bavardage ; la durée de la conversation peut également être utilisée, en partant du principe que des conversations plus longues signifient plus d'engagement et donc une expérience plus agréable ; le nombre de tours durant la conversation, qui donne une autre idée de l'engagement et de l'intérêt de l'utilisateur à converser avec le [chatbot](#). Ces mesures sont faciles à réaliser mais elles ne sont pas bien corrélées avec les résultats du [TT](#). Elles ne donnent pas non plus d'indications sur ce qui n'a pas fonctionné lorsque de mauvaises évaluations sont produites : elles ne renseignent en effet que sur la conversation générale et non sur les réponses spécifiques de l'agent conversationnel.

Certaines mesures donnent en revanche des informations plus spécifiques sur les différentes tournures que peut prendre la conversation. Les techniques les plus couramment utilisées sont celles de la traduction automatique telles que BLEU (Papineni et al., 2002) et METEOR (Banerjee & Lavie, 2005). Elles évaluent dans quelle mesure les réponses générées sont similaires à une réponse qui aurait été donnée par un humain à la même question. Ces méthodes ont le grand avantage d'être faciles à automatiser, mais présentent l'inconvénient de ne pas prendre en compte les éléments préalables de la conversation. En outre, la comparaison des mots utilisés peut créer des faux négatifs, car une réponse parfaitement intelligible et semblable à celle d'un humain peut prendre une direction inattendue qui serait différente des phrases auxquelles elle serait comparée, et donc donner un score faible bien qu'elle soit parfaitement valide (voir un exemple dans le [tableau 4.1](#)) (Liu et al., 2016).

L'[IA](#) peut également être utilisée pour évaluer la qualité des [chatbots](#). Par exemple,

TABLE 4.1 – Exemple dans lequel les métriques BLEU ou METEOR donnent de faux négatifs.

Contexte de la conversation
Interlocuteur A : Tu as entendu ? Le nouveau Batman va bientôt sortir !
Interlocuteur B : Non je ne savais pas. Tu penses qu’il sera bien ?
Réponse correcte attendue
Interlocuteur A : Je ne sais pas. Ils continuent à en faire des nouveaux et ils sont de pire en pire.
Réponse du chatbot
Interlocuteur A : Oui je suis pressé ! Je vais acheter des tickets dès sa sortie !

les [RNN](#) peuvent être entraînés pour imiter les humains dans l’évaluation des [chatbots](#) (Lowe et al., 2017). Les notes données par le réseau neuronal ont été significativement corrélées à celles données par les humains sur une échelle d’adéquation, qui, selon les auteurs, est la mesure la plus cohérente entre des juges humains. Malheureusement, la précision de ces modèles d’évaluation tend également à dépendre du contexte des conversations (on appréciera certainement l’ironie de voir un [chatbot](#) émulant un humain être évalué par un juge artificiel émulant lui-même l’évaluation d’un juge humain. Cela peut cependant rester une première étape d’évaluation intéressante). Il est également possible d’agrèger différentes mesures en utilisant des modèles entraînés pour émuler les juges humains plutôt que de se concentrer sur une seule mesure, comme l’engagement (capturé avec le nombre de tours ou la durée médiane des conversations), la cohérence, la profondeur conversationnelle, la diversité topique et la couverture du domaine (Venkatesh et al., 2018). Le principal problème ici est que certaines évaluations peuvent être assez subjectives et produire une grande variabilité. Par exemple, les auteurs indiquent qu’« un utilisateur peut donner 5 étoiles à une conversation parce qu’il/elle pense que le socialbot fait de l’humour, alors qu’un autre utilisateur peut le trouver ignorant ». Ainsi, il pourrait être injuste pour les [chatbots](#) d’attendre d’eux qu’ils soient généralement meilleurs en tout point par rapport à d’autres [chatbots](#), alors que, tout comme les humains, certains [chatbots](#) pourraient être mieux adaptés que d’autres à certaines tâches et ne pas être aussi performants dans d’autres tout en restant au-dessus d’une ligne de base acceptable.

Enfin, une métrique idéale pourrait aussi intégrer une évaluation des aspects émotionnels du [chatbot](#). En particulier dans les conversations liées à la santé physique ou mentale, le fait qu'un robot fasse preuve de compétences émotionnelles telles que l'empathie est un aspect important pour améliorer la façon dont les utilisateurs perçoivent et interagissent avec le [chatbot](#) (de Gennaro et al., 2020; Devaram, 2020). Ces compétences sociales seraient probablement aussi importantes à évaluer dans les contextes de l'éducation et des services clients.

L'utilisation de juges humains reste la norme car, en bout de ligne, ces outils sont censés interagir avec des humains. Malgré la part importante de subjectivité dans les évaluations humaines, due aux attentes individuelles d'un partenaire de conversation, tous les aspects d'une conversation humaine normale ne sont pas actuellement pris en compte par les mesures automatisées. Les humains doivent donc continuer à faire partie de la boucle de test et le [TT](#) a encore de beaux jours devant lui avant d'être totalement remplacé.

4.4 Le test de Turing comme test d'humanité

Si le [TT](#) a été initialement proposé comme un test de l'intelligence d'une machine (Turing, 1950), il a clairement évolué pour être considéré comme un test de d'humanité et est maintenant utilisé comme tel dans le panel existant des métriques d'évaluation.

Le [TT](#), dans sa conception moderne, consiste à faire dialoguer un juge humain avec deux autres agents, un humain et une machine, via une interface textuelle. L'objectif du juge humain est de déterminer lequel de ces agents est l'humain et lequel est la machine (ou, dans certaines versions, s'il existe une machine). Si, après une conversation de cinq minutes, le juge ne parvient pas à identifier correctement la machine dans 50 % des essais, alors la machine ressemble tellement à un humain, selon Turing (1950), qu'il serait nécessaire de lui attribuer des pensées de la même manière que nous le faisons avec les humains (et ainsi de supposer qu'ils ont des états mentaux comme nous parce qu'ils se

comportent comme nous).

Lassègue (Lassègue, 1996) souligne également l'existence d'une autre entité dont la prise en compte est importante dans le **TT** : il s'agit de l'expérimentateur, de l'arbitre. C'est la personne qui arrêtera le test après un laps de temps déterminé et dira si oui ou non la réponse du juge est correcte (l'arbitre connaît *a priori* la réponse au test). Ce point est important car le temps nécessaire pour réussir un **TT** varie beaucoup, parfois sans grande justification. Turing suggère 5 minutes, mais pourquoi pas 10 ? 7 ? 10 minutes et 30 secondes ? Passé ce délai, le **chatbot** pourrait potentiellement révéler sa nature.

Le **TT**, en tant que test d'intelligence, a historiquement reçu de nombreuses critiques. L'une des plus célèbres est basée sur l'expérience de pensée de la chambre chinoise imaginée par Searle et al. (1980). En résumé, Searle se trouve dans une pièce dans laquelle on lui donne des symboles chinois auxquels il doit répondre par des symboles chinois, ainsi que des instructions en anglais (appelées programme) pour relier une liste de symboles d'entrée à une liste de symboles de sortie. Incapable de comprendre le chinois lui-même, Searle affirme que s'il était capable de tromper les Chinois en suivant simplement les instructions (programme) qui lui ont été données pour leur faire croire qu'il était lui-même chinois, il ne comprendrait toujours pas le chinois du tout et suivrait ces instructions sans réfléchir.

Il est important de souligner ici que Searle n'a appliqué son objection qu'à un type spécifique d'**IA** : l'**IA** formelle, qui utilise des règles formelles pour interagir par le biais de textes avec l'utilisateur. Il n'a pas dit que les machines ne seraient jamais capables de penser, mais que pour ce faire, nous devrions comprendre le cerveau plutôt que d'abstraire ses fonctions générales sans comprendre comment il fonctionne. Une machine capable de passer le **TT** grâce à une manipulation parfaite de symboles n'aurait pas nécessairement un esprit propre, ne penserait pas nécessairement, ne serait pas nécessairement intelligente. Comme d'autres l'ont souligné, ces symboles doivent être fondés d'une manière ou d'une autre, représenter quelque chose pour vraiment signifier quelque chose, d'où la nécessité d'un développement plus sensori-moteur de l'**IA** (Harnad, 2001)

ainsi qu'une compréhension du fonctionnement du cerveau et de la compréhension des objets (Hawkins et al., 2019). Searle explique en effet :

A la question si les machines seront conscientes, il est important de se souvenir que nous sommes des machines. Nous sommes des machines biologiques et nous sommes conscients. Je ne vois pas de raison, en principe, pourquoi nous ne pourrions pas créer une machine artificielle qui serait consciente, mais nous sommes dans l'incapacité de le faire actuellement car nous ne connaissons pas suffisamment comment le cerveau crée la conscience. La question : « Peut-on fabriquer une machine qui serait consciente ? » est la même question que « Peut-on fabriquer un cœur artificiel qui pompe le sang ? » Nous savons comment construire des cœurs artificiels car nous savons comment le cœur biologique fonctionne. Nous ne savons pas construire de cerveau artificiel car nous ne savons pas comment le cerveau fonctionne. Mais si l'on part du principe que nous le savons, je ne vois pas d'obstacle sur le principe à ce que nous construisions une machine artificielle consciente. La chose importante à voir est que le cerveau humain est une machine, une machine biologique, et qu'il produit la conscience par des processus biologiques. Nous ne serons pas en mesure de le faire de manière artificielle tant que nous ne savons pas comment le cerveau le fait, et nous pourrions ensuite dupliquer les pouvoirs causaux du cerveau. Peut-être pouvons-nous le faire dans un matériau complètement différent comme nous construisons des cœurs artificiels dans un matériau complètement différent des tissus musculaires, mais à présent nous n'en connaissons pas suffisamment sur le cerveau pour construire un cerveau

artificiel²(Turello, 2015, traduction personnelle)³

Une remarque similaire a été faite encore plus tôt par Shannon, McCarthy et al. (1956).

Un désavantage de la définition de Turing de la pensée est qu'il est possible, en principe, de concevoir une machine avec une liste de réponses arbitrairement choisies pour tous les stimuli possibles... Avec un tel dictionnaire une telle machine satisferait sans doute la définition de Turing mais ne reflète pas notre intuition du concept de la pensée⁴ (p. vi Shannon, McCarthy et al., 1956, traduction personnelle).

Là où Searle (Searle et al., 1980) va très loin, c'est lorsqu'il décrit les instructions, ou le programme dans son exemple de chambre chinoise (de même que le dictionnaire pour Shannon, McCarthy et al. (1956) ou l'arbre des réponses sensées pour Block (1981)). Ils décrivent tous un problème du **TT** qui est effectivement réel et offrent des exemples conceptuels qui, s'ils étaient produits par une machine, passeraient tout de même un **TT**. Mais ces exemples sont-ils réalisables en pratique ? Une liste d'instructions aussi détaillée et exhaustive est-elle possible ? C'est extrêmement improbable (McDermott, 2014).

2. "As to whether or not machines will be conscious, it is important to remember that we are machines. We are biological machines and we are conscious. I do not see any reason, in principle, why we could not build an artificial machine that was conscious, but we are unable to do that now because we do not know how the brain does it. The question, "Can you build an artificial machine that is conscious?" is just like the question "Can you build an artificial heart that pumps blood?" We know how to build artificial hearts because we know how the biological heart works. We do not know how to build an artificial brain because we do not know how the brain works. But assuming we knew how the brain worked, I see no obstacle in principle to building an artificial conscious machine. The important thing to see is that the human brain is a machine, a biological machine, and it produces consciousness by biological processes. We will not be able to do that artificially until we know how the brain does it and we can then duplicate the causal powers of the brain. Perhaps we can do it in some completely different medium as we build artificial hearts in a completely different medium from muscle tissue, but at present we do not know enough about the brain to build an artificial brain."

3. A noter que Turing lui-même ne semblait pas opposé à cette idée. Il indique effectivement plusieurs fois qu'apprendre de l'humain serait la meilleure façon de passer le **TT**, mais il ne dit pas que d'autres stratégies ne peuvent pas fonctionner (see Saygin et al., 2000, p. 472). La différence principale entre Turing et Searle est que Turing suggère que cela peut être fait au niveau logiciel, alors que Searle considère que cela ne peut être fait qu'au niveau des composants de la machine elle-même.

4. "A disadvantage of the Turing definition of thinking is that it is possible, in principle, to design a machine with a complete set of arbitrarily chosen responses to all possible input stimuli.... With a suitable dictionary such a machine would surely satisfy Turing's definition but does not reflect our usual intuitive concept of thinking."

En effet, les conversations ne suivent pas un ensemble de règles aussi strictement qu'on pourrait le croire. Bien sûr, on peut commencer une conversation par un bonjour et la terminer par un au revoir, comme le veut la politesse. En fait, les philosophes et les linguistes ont tenté de mettre en évidence un ensemble de règles générales qui expliquerait comment nous conversons avec les autres, ce qui a donné naissance au domaine de la pragmatique conversationnelle. Grice (1975) est l'un d'entre eux. Il est à l'origine du PC (l'idée que les partenaires conversationnels essaient de coopérer pendant une conversation), et de quatre *maximes* qui sont une conséquence directe de ce principe : 1) la *maxime* de qualité, axée sur la véracité et la certitude d'une information donnée, 2) la *maxime* de quantité, axée sur la quantité d'informations données (ni trop ni trop peu), 3) la *maxime* de relation, suggérant que les participants à une conversation essaient de rester pertinents et 4) la *maxime* de manière, axée sur la façon dont les informations sont données (brièvement, clairement, de façon ordonnée et sans ambiguïté). Encore une fois, il ne s'agit pas de règles, mais plutôt d'attentes que chaque agent d'une conversation a vis-à-vis de la production des autres agents. Très souvent, les locuteurs ne les suivent pas strictement : « C'est un requin » n'est évidemment pas une déclaration qui doit être prise au pied de la lettre, mais elle transmet plutôt l'idée qu'« il » vous prendra tout ce qu'il peut. Grice (1975) en était bien conscient et considérait cette pratique dans les conversations comme un « opting-out » : une déviation des *maximes* toujours dans le cadre d'une coopération. Il peut toutefois se produire des violations réelles des *maximes*. Ce sont les cas où les participants à une conversation n'essaient plus de coopérer : par exemple, mentir dans une conversation serait une violation de la *maxime* de qualité concernant la véracité de l'information donnée, qui serait faite à l'insu de l'autre partenaire de conversation : on se retire donc volontairement de l'acte de coopération dans la conversation. Pourtant, violer ces *maximes* ne rend pas moins humain, mais toute violation détectée par les autres partenaires donnera lieu à des inférences différentes, et la violation elle-même sera considérée comme un élément d'information à part entière.

Le concept le plus important que Grice propose et qui est d'une grande importance pour les *chatbots* est la distinction entre ce qui est *dit* et ce que l'on *veut dire*. Prenons

l'exemple suivant : « Entrez ! Mais je n'ai pas d'alcool ». À première vue, il serait difficile de dire en quoi le fait d'inviter quelqu'un serait aussi directement lié à l'alcool sans autre information. Pourtant, cette phrase est facilement compréhensible et peut déclencher une réponse offensée, déçue ou amusée, en fonction de la relation entre les deux interlocuteurs. Ce que l'on *veut dire* ici est « Entrez, mais il n'y a pas d'alcool à l'intérieur et je sais que vous vous attendiez peut-être à ce que nous partagions de l'alcool ». Ici, la clé pour comprendre la mention de l'alcool (on s'attendait à boire) est complètement implicite. Alors que la mention de l'alcool semble venir de nulle part et semble donc violer la [maxime](#) de relation, la phrase est comprise comme étant parfaitement pertinente dans le contexte donné, en raison des attentes préalables concernant la situation.

Pour expliquer ces productions, Wilson et Sperber (2002) ont développé la *Théorie de la pertinence*. L'idée principale de cette théorie est que les participants à une conversation recherchent activement la pertinence des énoncés des interlocuteurs. La théorie de la pertinence décrit un énoncé avec une pertinence optimale comme un énoncé qui a le plus grand effet contextuel sur les représentations mentales de l'auditeur pour le moindre coût cognitif (le moindre effort pour retrouver ce que l'on *veut dire* à partir de ce qui est *dit*). En effet, dans l'exemple précédent, quelle est l'utilité d'inclure de manière explicite que boire est quelque chose d'attendu quand c'est une attente que les deux participants à la conversation partagent déjà ? Cela ne ferait que rendre l'interprétation de la phrase plus difficile, prendrait plus de temps, et n'apporterait rien (cela ne changerait pas les représentations mentales de l'auditeur puisque c'est quelque chose qu'il sait déjà). Ainsi, l'ajouter dans l'énoncé n'est pas pertinent, cela doit rester implicite.

Comme la théorie de la pertinence présuppose que les participants à une conversation aient une idée de ce qui se trouve dans l'esprit des autres, la participation à une conversation réelle (au niveau humain) nécessite une [Théorie de l'Esprit, ou Theory of Mind \(ToM\)](#) (Premack & Woodruff, 1978). La [ToM](#) est le concept selon lequel les humains (peut-être parmi d'autres animaux) sont capables de lire les pensées d'autres individus. Pas de manière métaphysique bien sûr, mais les humains comprennent que les autres hu-

mains pensent également, qu'ils ont des représentations mentales du monde qui peuvent être différentes ou similaires aux leurs. Il y a des choses que les autres ne savent pas que nous savons, et il y a des choses que nous ne savons pas qu'ils pourraient savoir... C'est la raison pour laquelle les conversations contiennent des questions : nous comprenons que les autres pourraient avoir les réponses que nous cherchons, et nous leur demandons de partager les informations dont ils disposent avec nous. Réciproquement, la seule raison pour laquelle les gens répondent aux questions est qu'ils supposent que les personnes qui les posent ne connaissent pas déjà les réponses et qu'elles apprendront ainsi une nouvelle information (leurs représentations mentales changeront une fois que la réponse leur aura été donnée). Des observations indiquent que les humains acquièrent cette capacité très tôt dans leur vie (Baratgin et al., 2020 ; Bretherton et al., 1981) et la présence ou l'absence de cette capacité chez d'autres espèces fait encore l'objet d'un vif débat dans la communauté scientifique (Krupenye & Call, 2019 ; Penn & Povinelli, 2007). Cela n'est pas entièrement surprenant étant donné la difficulté de trouver des moyens de communiquer explicitement la question sans ambiguïté aux jeunes humains (Baratgin et al., 2020).

Comme les conversations sont construites autour de ces principes, les réponses données au cours d'une conversation ne sont pas fixes et dépendent fortement de ce que chaque participant à la conversation croit que l'autre sait. Ainsi, les règles prédéfinies, comme celles mentionnées dans la chambre chinoise de Searle et al. (1980) (les instructions données à l'homme à l'intérieur de la chambre sur la façon de faire correspondre une chaîne de symboles comme réponse à une autre chaîne de symboles) ne peuvent faire plus qu'imiter et réduire considérablement l'éventail des possibilités qu'offrent les conversations naturelles. Non seulement l'ensemble des instructions devrait être infiniment grand, mais il devrait être aussi constamment mis à jour pour être adapté aux publics spécifiques et aux changements dans le temps induits par l'évolution du langage naturel. Ainsi, nous pensons que les IA basées sur des règles, telles qu'ELIZA, A.L.I.C.E et Kuki.ai, ne sont pas en mesure de réussir le TT car elles sont incapables d'apprendre de leurs interactions. De même, les systèmes *retrieval-based* qui utilisent l'apprentissage automatique pour détecter l'intention ne pourront probablement pas non plus passer le TT car ils ne sont pas

capables de générer de nouvelles réponses adaptées à de nouvelles situations comme le ferait un humain. Seule une machine apprenant à inférer la signification et à modifier la façon dont elle s'exprime devrait être en mesure de passer le **TT** de manière satisfaisante, même si, à l'heure actuelle, les **IA** génératives sont moins utiles et plus frustrantes que les **IA** *retrieval-based*.

Mais est-ce qu'un juge serait capable de faire la différence dans un **TT** ? Ne pas comprendre le contexte et l'esprit d'autrui est-il suffisant pour empêcher de manière significative une machine de passer le **TT** ? Comme nous allons le montrer dans la section suivante, la réponse est oui.

4.5 Comprendre les humains

La comparaison entre les conversations humain-**chatbot** et les conversations humain-humain présente de nombreux avantages à la fois pour la psychologie et pour l'informatique. L'étude du comportement des humains par rapport aux **chatbots** peut nous aider à créer de meilleurs **chatbots**, et l'étude des interactions avec les **chatbots** peut nous donner des informations précieuses sur ce que les humains attendent d'un partenaire conversationnel. Et pourtant, malgré le fait que le **TT** peut être utilisé comme un paradigme expérimental utile pour explorer les attentes humaines dans les conversations, il est remarquablement absent des publications internationales dans le domaine de la psychologie et de la pragmatique. En effet, une recherche rapide sur Google Scholar révèle environ 33.000 entrées pour « Turing Test », tandis que l'ajout de mots-clés appartenant au domaine de la pragmatique fait chuter le nombre d'occurrences à moins de 300 entrées (Implicatures « Turing Test » : 209 résultats, « Turing Test » « Relevance Theory » : 96 résultats, « Turing Test » « Cooperation Principle » : 104 résultats), la plupart ne faisant que mentionner chacun de ces sujets sans s'y attarder.

Les **chatbots** sont encore loin de répondre aux attentes humaines en matière de partenaires conversationnels. Les enquêtes et les études montrant que les gens sont rapi-

dement frustrés lorsqu'ils les utilisent ne sont pas difficiles à trouver (voir (Brandtzaeg & Følstad, 2018 ; Jain et al., 2018 ; Luger & Sellen, 2016) pour n'en citer que quelques-unes). Une vaste enquête menée sur la littérature relative aux **chatbots** indique les nombreux défis auxquels ils sont encore confrontés (Chaves & Gerosa, 2020), notamment en ce qui concerne les caractéristiques sociales des **chatbots** (Zamora, 2017). Ce sentiment de frustration peut être atténué lorsque l'utilisateur sait clairement ce qu'il peut attendre du **chatbot**. Par exemple, Woebot définit clairement les attentes de ses utilisateurs à l'avance, ce qui permet à ces derniers d'adapter leur propre comportement (Fitzpatrick et al., 2017). Dans le cas de ce **chatbot** (qui agit comme un coach pour aider à gérer l'anxiété et la dépression), le robot garde le contrôle de la conversation à tout moment pendant que l'utilisateur navigue dans des arbres de décision prédéfinis, et ce faisant, il est capable de mener à bien sa tâche, bien que dans les cas trop graves, l'utilisateur soit redirigé vers une hotline grâce à laquelle il peut interagir avec des professionnels humains pour demander de l'aide. Cette transparence sur les capacités du **chatbots** (ainsi que son utilisation très limitée de la compréhension du langage naturel) lui permet d'être efficace dans sa tâche consistant à aider les personnes à faire face à l'anxiété et à la dépression, du moins pendant une courte période (l'étude n'ayant pas porté sur les effets à long terme). De manière similaire, ce **chatbots** semble également montrer une certaine efficacité dans l'aide à la consommation de substances (Prochaska et al., 2021).

En effet, plus le **chatbot** se rapproche de la sensation humaine, plus les utilisateurs s'attendent à des capacités de type humain dans leurs interactions avec eux. Il est également possible d'observer cet effet sur les mêmes **chatbots** selon la manière dont ils sont introduits. Par exemple, une étude peut trouver le programme suffisamment divertissant pour que les utilisateurs continuent à converser avec pendant de longues périodes, malgré une qualité de conversation nettement inférieure à celles avec des humains (Hill et al., 2015), tandis que dans une autre, les juges d'un **TT** peuvent rester assez perplexes lorsqu'ils ne sont pas informés que l'auteur des messages pourrait être un **chatbot**, se demandant si la personne qui écrit de tels messages pourrait être « mentalement malade » (Saygin & Cicekli, 2002).

Ces situations de violation des attentes de l'utilisateur sont courantes lors de l'interaction avec les *chatbots*, créant un sentiment similaire à la vallée de l'inconfort de Mashiro Mori (Vallverdú et al., 2012) qui est un effet célèbre observé avec les robots (DiSalvo et al., 2002) (Plus un agent artificiel, robot ou *chatbot*, se rapproche du comportement ou de l'apparence humaine, plus les attentes des humains interagissant avec lui seront grandes, et plus la frustration ou l'inconfort seront grands si ces attentes ne sont pas atteintes). On peut néanmoins se demander si, tous les autres aspects restant similaires, la violation de ces attentes serait suffisante pour empêcher une machine de passer le *TT*. Saygin et Cicekli (2002) étudient cette question dans un *TT* en essayant d'évaluer à quel niveau chacune des *maximes* de Grice (1975) a (ou n'a pas) un effet sur les réponses des participants. Leurs résultats indiquent que toutes les *maximes* n'exercent pas la même influence sur les réponses du *TT*. En effet, la violation de la *maxime* de manière (qui traite de la façon dont les informations sont données à l'utilisateur) n'a, d'après les juges, pas d'effet négatif sur le sentiment d'humanité donné par le *chatbot*. En fait, les auteurs observent même que ce type de violation a un effet positif tant qu'aucune autre *maxime* n'est violée. Ils expliquent ce résultat par le fait que la violation de cette *maxime* peut produire une réponse apparemment plus chargée en émotions, et les émotions sont une caractéristique plus facilement associée aux humains qu'aux machines. La violation de la *maxime* de quantité s'est avérée n'avoir aucun effet sur le *TT* lorsqu'elle a été violée en donnant trop peu d'informations. La violation de la *maxime* de quantité n'a pas non plus été très préjudiciable pour le sentiment d'humanité donné lorsque la *maxime* a été violée dans l'autre sens en donnant cette fois trop d'informations (ce qui donnait à la réponse un aspect encyclopédique). La difficulté d'évaluer l'effet individuel des violations de cette *maxime* est que, lorsqu'elle est violée, elle a également tendance à violer la *maxime* de relation, qui produit de loin l'effet négatif le plus fort sur le sentiment d'humanité : le juge a l'impression que le *chatbot* ne comprend tout simplement pas la question (ou, lorsque les juges ne savent pas qu'un *chatbot* est présent, que celui-ci ne veut pas parler de ce sujet sans pour autant qu'il y ait une raison compréhensible). Enfin, les auteurs n'ont pas été en mesure de montrer une influence spécifique de la *maxime*

de qualité sur le sentiment d'humanité donné par le [chatbot](#), car elle avait également tendance à être violée en même temps que d'autres [maximes](#).

Une différence importante subsiste entre l'article ci-dessus et un [TT](#) ordinaire. Dans l'article de Saygin et Cicekli (2002), les juges lisaient des extraits de conversations enregistrées lors d'un concours du prix Loebner et n'interagissaient pas réellement avec les [chatbots](#). Des utilisateurs interagissant avec un [chatbot](#) pour lequel le seul problème serait un manque de pertinence ou d'autres violations remarqueraient-ils suffisamment ce défaut pour étiqueter correctement le [chatbot](#) comme une machine? Nous avons tenté de répondre à cette question dans les articles précédents : (Jacquet et al., 2018, 2019 ; Jacquet, Hullin et al., 2019a) en invitant les participants à prendre le rôle de juges dans un [TT](#). Le principal intérêt de notre approche ici était de tester l'influence de ces seules violations : en effet, les juges ont participé à deux conversations dans un ordre aléatoire, en étant informés que l'une serait avec un [chatbot](#) et l'autre avec un humain. En réalité, il n'y avait pas de [chatbot](#) du tout. En effet, l'utilisation d'un [chatbot](#) aurait rendu plus difficile le test visant à déterminer si les différences observées auraient été causées par les violations ou par d'autres facteurs liés au [chatbot](#). Les deux conversations étaient jouées par le même expérimentateur humain, qui incarnait à chaque fois un personnage fictif (le même personnage fictif pour les deux conversations), mais dans l'une des conversations, l'expérimentateur avait pour tâche de produire des violations de l'une des [maximes](#) de Grice. Une fois encore, les violations qui ont le plus d'effet sur le sentiment d'humanité sont les violations de la [maxime](#) de relation (Jacquet et al., 2018, 2019) et les violations de la [maxime](#) de quantité donnant lieu à un sentiment encyclopédique (Jacquet, Hullin et al., 2019a). Cet effet était également visible dans le délai qui sépare l'énonciation de l'expérimentateur et le tour du participant (qui est plus long après une violation qu'après une réponse attendue), ce qui indique encore que ces violations sont bien la cause de la différence observée. En outre, le type de violations de la [maxime](#) de relation dans ces articles était légèrement plus subtil que les violations flagrantes que l'on peut souvent trouver dans les [chatbots](#) : l'expérimentateur n'était pas autorisé à utiliser des connaissances antérieures à la conversation dans ses réponses, mais pouvait toujours répondre de

manière pertinente si toutes les informations nécessaires pour le faire étaient contenues dans le dernier message du participant. Par exemple :

Humain : Aimes-tu lire ?

Expérimentateur : Pas vraiment non. C'est pas vraiment mon truc.

Humain : Pourquoi pas ?

Expérimentateur : C'est difficile à dire. Tu as des frères et soeurs ?

Dans la première réponse de l'expérimentateur, ils sont autorisés à donner une réponse pertinente, mais dans leur deuxième réponse, ils n'étaient pas autorisés à utiliser le fait que le sujet portait sur la lecture. Ils ont donc utilisé une réponse générique à la place, ce qui a entraîné une violation de la [maxime](#) de relation.

Ce type de violation est très facile à obtenir sur n'importe quel [chatbot](#) actuellement disponible. Poser des questions génériques telles que « Pourquoi ? » ou « Pourquoi pas ? » exige que le [chatbot](#) utilise le contexte du message (l'historique de la conversation) pour pouvoir répondre correctement. Dans la deuxième question de l'humain, il suppose que son lecteur a toujours à l'esprit le sujet de la conversation (la lecture n'est pas le truc de l'expérimentateur), tandis que l'expérimentateur doit déduire que ce que le participant veut dire, c'est « Pourquoi la lecture n'est pas vraiment votre truc ? » lorsqu'il dit « Pourquoi pas ? ».

D'autres études doivent être menées pour explorer la sensibilité du [TT](#) à des violations encore plus subtiles, mais avec les preuves dont nous disposons aujourd'hui, il semble très probable que seul un [chatbot](#) capable de converser de manière pertinente dans toutes les situations serait en mesure de passer le [TT](#) (surtout dans sa version à 3 joueurs : juge, machine et humain, sans limites sur les sujets de discussion), et cela nécessiterait la capacité de développer une idée de ce qui est pertinent pour l'utilisateur, et donc pour le [chatbot](#) d'avoir une théorie de l'esprit (Jacquet & Baratgin, 2019). Nous n'en sommes pas encore là (Jacquet & Baratgin, 2020).

4.6 Conclusion

Bien que nous n'ayons fait qu'effleurer la littérature concernant le **TT**, nous avons exploré la littérature existante qui discute de l'importance de la pragmatique conversationnelle pour des **chatbots**, et nous avons tenté de montrer en quoi le **TT** est un outil très intéressant pour évaluer la capacité des **chatbots** à générer des réponses pertinentes dans une conversation ouverte qui n'est (jusqu'à présent) égalée par aucune autre méthode d'évaluation.

Nous avons également discuté de la façon dont le **TT**, dans sa conception, suggère que seul un agent doté d'une théorie de l'esprit pourrait le réussir de façon fiable, bien que, bien sûr, il ne fixe aucune exigence sur la façon dont cette théorie de l'esprit est mise en œuvre.

Nous pensons également que le **TT** devrait être plus largement utilisé dans les sciences humaines comme la psychologie, notamment dans le cas de l'étude du raisonnement et de la pragmatique conversationnelle. Le **TT** est un outil trop rarement utilisé, bien qu'il se révèle être un paradigme expérimental précieux qui permet aux expérimentateurs de recueillir des mesures directes (la réponse dans le **TT**) et indirectes (le délai entre les énoncés pendant les conversations par exemple). Ce domaine de recherche est encore peu développé malgré son grand potentiel pour la recherche fondamentale et appliquée. Un exemple consisterait à tester l'influence de l'utilisation de textismes (langage SMS) sur le coût cognitif du traitement des messages dans une conversation ou d'utiliser des **chatbots** pour étudier comment les comportements sont influencés par différents indices pragmatiques dans le jeu de l'ultimatum (Beunay et al., 2022). Nous verrons ces deux exemples dans la troisième partie de cette thèse.

Enfin, certains lecteurs pourraient objecter que nous n'avons pas réglé la question de savoir si le fait de réussir le **TT** est une preuve de la présence d'un esprit. Après tout, avons-nous besoin d'un esprit pour avoir une théorie de l'esprit ?

Chapitre 5

La coopération dans les conversations humaines : Les temps de réponse comme fenêtre sur le traitement du langage

Points remarquables dans ce chapitre

1. Nous proposons une expérience pour tester l'influence de certaines maxims de Grice sur l'humanité de chatbots.
2. Le paradigme de test que nous utilisons est le test de Turing.
3. Nous montrons la pertinence d'utiliser les temps de réponse au cours des conversations.
4. Nous mettons en évidence l'importance de correctement traiter la maxime de relation dans l'évaluation des chatbots.

5.1 Introduction

Les récents progrès de l'[intelligence artificielle \(IA\)](#) ont permis la diffusion d'agents sociaux virtuels dans de nombreux domaines, notamment en tant qu'agents de service clientèle (Chakrabarti & Luger, 2015 ; Cui et al., 2017 ; Xu et al., 2017), mais aussi en tant

que coachs fournissant une aide pour gérer au quotidien des problèmes psychologiques comme la dépression ou l'anxiété, à l'instar de Woebot ¹ ou Tess ². Ces agents prennent souvent la forme de **chatbots** (ou robots conversationnels) : il s'agit d'agents conversant avec un utilisateur par le biais d'une conversation textuelle utilisant, en général, l'imitation de la compréhension et de la génération du langage naturel.

Alors que des progrès considérables ont été réalisés dans les deux domaines du traitement syntaxique (voir D. Chen & Manning, 2014; Socher et al., 2010, pour des exemples) et du traitement sémantique (voir Berant & Liang, 2014; Pasupat & Liang, 2015, pour des exemples), un aspect des conversations naturelles est souvent oublié : le traitement pragmatique (voir Jacquet, Masson et al., 2019; Jokinen & De Smedt, 2012, pour des critiques). Il ne suffit pas de comprendre la structure d'un énoncé et son contenu sémantique pour avoir une compréhension complète de l'énoncé lui-même dans son contexte. En effet, il peut y avoir de grandes différences entre ce qui est *dit* et ce que l'on *veut dire* dans les conversations entre humains comme l'ont noté Grice (1975), et plus tard Wilson et Sperber (2002).

Sur cette distinction entre ce qui est *dit* et ce que l'on *veut dire*, Grice (1975) a introduit le principe de coopération ainsi que ses **maximes** pour décrire les diverses attentes qui permettent aux partenaires de conversation d'inférer le sens d'un énoncé à travers l'intention de son locuteur. La Théorie de la Pertinence (Wilson & Sperber, 2002) a par la suite mis à jour les principes originaux de Grice et a offert une explication plus approfondie et plus unifiée des processus impliqués dans l'inférence de ce que l'on *veut dire* à partir de ce qui est *dit* (et de ce qui n'est *pas dit*).

Parce que ce traitement pragmatique nécessite par l'agent virtuel une compréhension des attentes de son utilisateur, les **chatbots** ont, à ce jour, encore du mal à y parvenir. Les **chatbots** les plus avancés aujourd'hui (comme Cleverbot ³, et dans une plus

1. Woebot Labs Inc. <https://woebot.io>

2. X2AI Inc. <http://x2ai.com/>

3. Cleverbot est disponible sur <https://www.cleverbot.com/>

large mesure Zo⁴ et ceux qui utilisent Watson⁵) sont raisonnablement convaincants dans leur capacité à répondre aux questions et à entretenir des conversations simples et laco- niques, mais leurs propos perdent rapidement de leur pertinence après quelques phrases, parfois même dès leur deuxième énoncé (surtout pour Cleverbot). Chez Zo, qui est sans doute l'un des agents conversationnels les plus convaincants, ce manque de pertinence est esquivé par le ton de la conversation, puisqu'elle répondra souvent par des énoncés humo- ristiques, notamment en utilisant des *memes*⁶, qui permettent à l'utilisateur de trouver une interprétation de la réponse de l'agent qui soit pertinente pour lui, malgré le caractère générique de la réponse.

Si cela donne des agents conversationnels amusants avec lesquels il est intéressant de jouer pendant un certain temps, ce manque de pertinence les rend inutilisables dans des environnements plus sérieux, où un manque de raisonnement pragmatique ne peut pas être caché derrière une réponse superficielle mais apparemment pleine d'esprit. Au cours d'une conversation avec un agent du service clientèle, par exemple, les utilisateurs s'attendent à ce que l'agent soit utile, et non spirituel (Chakrabarti & Luger, 2015).

Malgré l'importance du traitement pragmatique dans le degré de crédibilité d'un agent conversationnel, l'évaluation de sa qualité au sein des conversations est difficile à au- tomatiser. Contrairement au traitement syntaxique, il n'existe pas de règles bien définies et suffisamment stables sur lesquelles construire un programme d'analyse. Contrairement au traitement sémantique également, il n'existe pas d'associations directes et stables entre certaines caractéristiques du discours et le sens implicite.

Par conséquent, il n'existe pas d'étalon-or pour évaluer la qualité d'un agent conversationnel, et encore moins pour évaluer la qualité de son traitement pragmatique,

4. Zo est un agent conversationnel développé par Microsoft. Il est possible de converser avec lui via twitter : <https://twitter.com/zochats>

5. Watson existe sous la forme d'une API d'agent conversationnel développée par IBM, permet- tant aux entreprises de créer leur propre version à des fins spécifiques. <https://www.ibm.com/watson/how-to-build-a-chatbot>

6. Images ou GIFs qui remplacent généralement une expression faciale, peuvent être une référence à la culture populaire ou à de nombreux autres types de contenus. Ils sont généralement destinés à véhiculer de l'humour.

même si de nombreuses méthodes d'évaluation différentes existent pour les agents conversationnels en général : (Chakrabarti & Luger, 2015 ; Hung et al., 2009 ; Meira & Canuto, 2015 ; Paroubek et al., 2007 ; Ptaszynski et al., 2010).

Nous soutenons que le **TT** (Turing, 1950), bien connu en informatique comme une méthode pensée pour tester l'intelligence d'une machine dans des conversations textuelles en la comparant à un humain, peut être suffisant pour détecter des défauts dans le traitement pragmatique lorsqu'il est plutôt considéré comme testant l'humanité. Nous soutenons également que le pouvoir discriminant du **TT** dans ce contexte peut être rendu plus spécifique, pour atteindre le niveau de phrases individuelles, s'il est associé à un enregistrement des **temps de réponse** (**TR**) pour indiquer la difficulté à inférer le sens de ces phrases.

Pour tester cette idée, nous avons utilisé trois des **maximes** de Grice couramment violées dans les conversations avec des agents artificiels. Ces **maximes** sont la première **maxime** de quantité, la **maxime** de relation et la quatrième **maxime** de manière, dont il est montré qu'elles ont toutes des effets différents sur le sentiment d'humanité donné par un **chatbot** (Saygin & Cicekli, 2002). Nous avons émis l'hypothèse que les agents conversationnels violant les **maximes** seraient identifiés comme étant plus semblables à des machines (surtout pour les violations de la **maxime** de relation) que les agents conversationnels produisant des énoncés plus typiques.

Nous avons également émis l'hypothèse que les **TR** augmenteraient après des énoncés nécessitant un effort cognitif important pour en déduire le sens. En particulier, nous nous attendions à ce que les réponses des participants soient le plus retardées après les violations de la **maxime** de relation, une augmentation plus faible ou absente après les violations de la **maxime** de quantité, et une augmentation intermédiaire après les violations de la **maxime** de manière.

5.2 Revue de la littérature

La pragmatique des conversations

Le langage est un code qui chiffre les concepts en symboles qui sont les mots d'une phrase. Cependant, ce code ne contient pas toutes les informations nécessaires pour retrouver la complexité des concepts qu'il référence. En effet, il ne contient que les informations nécessaires pour rapprocher légèrement les représentations mentales du partenaire conversationnel qui est en position de récepteur de celles du partenaire conversationnel en position d'émetteur. Cette opération donne à l'émetteur la possibilité d'économiser du temps de traitement (effort cognitif) et de ne le consacrer qu'à l'encodage des informations qu'il pense que son partenaire ne possède pas déjà. Les détails qui ne sont pas importants pour atteindre le but de la conversation peuvent être laissés de côté, laissant le récepteur combler les lacunes avec ses connaissances et, ce faisant, éviter de devoir traiter des informations redondantes. Cet équilibre entre un coût (temps de traitement) et un effet (la modification des états mentaux du récepteur) a été décrit dans la théorie de la pertinence (Relevance Theory) (Sperber & Wilson, 2015 ; Wilson & Sperber, 2002).

La théorie de la pertinence considère que la pertinence d'un énoncé est le résultat de l'interaction entre l'effet d'un énoncé sur les représentations mentales du lecteur et le coût de traitement nécessaire pour en déduire le sens, en utilisant des indices contextuels en plus de l'énoncé lui-même. Plus le coût de traitement est élevé, plus la pertinence est faible, et plus l'effet contextuel sur les représentations mentales est élevé, plus la pertinence de l'énoncé est élevée. Par conséquent, la pertinence optimale est atteinte lorsqu'un énoncé a un effet contextuel élevé pour un coût de traitement faible. Il est également important de considérer que, selon cette théorie, les participants à une conversation supposeront par défaut la pertinence d'un énoncé, et qu'il vaut la peine de le traiter pour récupérer son message implicite.

Dans de nombreuses tâches cognitives, le comportement des participants pourrait être considéré comme biaisé par rapport à la logique formelle. La théorie de la pertinence, en donnant de bonnes prédictions de ces comportements, a au contraire révélé que les gens ne sont pas incohérents dans leur façon de raisonner car ils utilisent des informations contextuelles en plus des informations qui leur sont explicitement fournies, ce qui rend difficile le raisonnement sans contexte concret. La littérature est assez abondante dans ce domaine et comprend la logique des connecteurs (voir par exemple Noveck, 2001 ; Politzer, 1986 ; Sperber et al., 1995), la tâche d'inclusion de Piaget (voir par exemple Masson et al., 2016 ; Politzer, 2016), biais dans le jugement de probabilité (voir exemples Baratgin & Noveck, 2000 ; Baratgin & Politzer, 2006, 2007, 2010 ; Hilton, 1995), et prise de décision (voir par exemple Bagassi & Macchi, 2006 ; Bless et al., 1998 ; Masson, Baratgin & Jamet, 2017a ; Masson, Baratgin, Jamet et al., 2017).

La théorie de la pertinence s'est initialement inspirée des travaux de Grice (1975), avant de les unifier et de les étendre. En essayant de décrire ce que sont les conversations, Grice a noté que :

Nos échanges ne consistent pas normalement en une succession de remarques déconnectées, et ne seraient pas rationnels si ils l'étaient. Ils sont de manière caractéristique, du moins dans un certain degré, des efforts coopératifs ; et chaque participant reconnaît en eux, dans une certaine mesure, un but commun, ou au moins une direction mutuellement acceptée. Ce but ou cette direction peut être convenu dès le début (e.g., en proposant initialement une question à discuter), ou il peut évoluer au cours de l'échange... Mais à chaque étape, *certain*s mouvements conversationnels seraient exclus car n'étant pas acceptable dans la conversation⁷ (Grice, 1975, p. 45, emphase dans l'original).

7. "Our talk exchanges do not normally consist of a succession of disconnected remarks, and would not be rational if they did. They are characteristically, to some degree at least, cooperative efforts ; and each participant recognizes in them, to some extent, a common purpose or set of purposes, or at least a mutually accepted direction. This purpose or direction may be fixed from the start (e.g., by an initial proposal of a question for discussion), or it may evolve during the exchange... But at each stage, *some* possible conversational moves would be excluded as conversationally unsuitable"

C'est à la suite de cette description qu'il a proposé le [principe de coopération \(PC\)](#), défini comme suit :

Fais que ta contribution à la conversation soit telle qu'elle est requise, à l'étape où elle a lieu, par le but ou la direction acceptée de l'échange dans lequel tu es engagé⁸ (Grice, 1975, p. 45).

Dans le [PC](#), Grice explique que les participants à une conversation ont des attentes concernant la forme et le contenu des énoncés de leur partenaire. Il les a classées en quatre [maximes](#) différentes et leurs sous-maximes.

Qualité

La première de ces [maximes](#) est la [maxime](#) de qualité : « Essayez de faire de votre contribution une contribution qui soit vraie ». Elle se subdivise en deux sous-maximes : 1) « ne dites pas ce que vous croyez être faux » , et 2) « ne dites pas ce pour quoi vous manquez de preuves » .

Grice considérait que la [maxime](#) de qualité était celle dont dépendaient les trois autres, et on peut même affirmer que l'ensemble du principe de coopération repose sur le fait que le récepteur des énoncés considère que l'émetteur croit que ce qui est dit contient une information qu'il est intéressant de traiter. Sinon, le récepteur n'essaierait même pas de déduire un sens caché. Par conséquent, bien que cette [maxime](#) ne dépende pas des trois autres, on ne peut pas dire que le contraire soit vrai, car toutes les trois dépendent fortement du respect de la [maxime](#) de la Qualité (Benton, 2016).

On peut également soutenir que cette [maxime](#) dépend fortement de l'interprétation occidentale de ce qu'est un mensonge, puisqu'elle utilise le terme « essayer [de faire de votre contribution une contribution qui soit vraie] », mais dans la culture Mopan, par exemple, le mensonge n'est pas considéré comme dépendant de l'état mental de l'émetteur

8. "Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged"

et du fait qu'il croit que son énoncé est vrai ou non : quoi qu'il en soit, il sera considéré comme une violation de la qualité conversationnelle si l'information en elle-même peut être considérée comme fausse (Danziger, 2010).

Il faut aussi considérer le fait que la sous-maxime de véracité (« ne dis pas ce que tu crois être faux ») est souvent violée (ou au moins suspendue) dans des cas comme les métaphores et l'ironie, alors que la **maxime** de Qualité elle-même ne l'est pas, puisque ces phrases impliquent toujours une information vraie malgré le fait que ce qui est dit est factuellement faux. Elles ne peuvent donc pas être considérées comme de véritables mensonges, qui sont des violations de la sous-maxime de véracité *et* de la **maxime** de qualité tout en faisant croire à l'interlocuteur qu'elles sont en fait respectées (Wilson, 1995).

Malgré son importance, on ne peut pas dire que sa violation produise un effet réel sur l'humanité d'un interlocuteur au cours d'une conversation (Saygin & Cicekli, 2002). En conséquence, cette **maxime** ne sera pas l'objet principal de notre étude.

Quantité

La deuxième **maxime** de Grice est la **maxime** de quantité, qui explique les attentes concernant la quantité réelle d'information contenue dans un énoncé et est définie par deux sous-maximes : 1) « Rendez votre contribution aussi informative que nécessaire (pour les objectifs actuels de l'échange) », et 2) « Ne rendez pas votre contribution plus informative que nécessaire ».

La première sous-maxime de Quantité est assez souvent violée, et ce que sa violation implique est généralement simple à interpréter. Par exemple, violer cette **maxime** est très couramment utilisé lorsque l'émetteur d'une énonciation tente de tromper le récepteur pour lui cacher certaines informations (McCornack, 1992). Ces violations peuvent également être utilisées pour impliquer que l'émetteur ne connaît pas la réponse à une question, et qu'il ne s'en soucie pas non plus, par exemple :

A Où vit Marc ?

B Quelque part sur la Terre.

Dans ce cas, *B* répond avec une information qui était déjà bien connue de *A*. Par conséquent, cette énonciation a une faible pertinence, puisqu'elle n'a pas d'effet contextuel sur les représentations mentales, autre que de potentiellement changer la perception de l'état d'intérêt de *B* dans la conversation. Bien qu'une violation de la sous-maxime n'implique pas nécessairement un manque d'intérêt. Par exemple :

A Où habite Marc ?

B Quelque part en France.

Dans ce cas, *B* ne donne probablement pas assez d'informations (à moins que le contexte ne soit spécifiquement de parler de pays), mais cette réponse pourrait simplement être le résultat de la volonté de ne pas violer la *maxime* de la Qualité : ne pas donner plus d'informations que ce que l'on sait être vrai.

Engelhardt et al. (2006) utilise un protocole expérimental similaire à celui utilisé par Tanenhaus et al. (1995) consistant à montrer à des paires de participants des objets à déplacer, ainsi que l'endroit où ils peuvent potentiellement être déplacés. L'un des participants doit décrire l'endroit où chaque objet doit être déplacé, tandis que l'autre doit effectuer l'action correspondante en déplaçant l'objet en conséquence. L'étude a montré que les participants préfèrent décrire plus que nécessaire les objets à déplacer (et ce faisant, violent la deuxième sous-maxime de la quantité) tout en essayant d'éviter de les décrire moins que nécessaire, qui est une source d'ambiguïté. La violation de la *maxime* dans l'instruction qu'il reçoivent a un effet direct sur les participants lors de la tâche de déplacement de l'objet : si trop peu d'informations sont données, cela sème visiblement la confusion chez les participants, comme le montre l'observation de leurs fixations oculaires. La confusion a également été observée lors d'une communication d'informations non nécessaires (trop importante), les participants ayant besoin de plus de temps

pour comprendre le mouvement à effectuer. Des résultats similaires ont déjà été publiés auparavant par Spivey et al. (2002).

Horn (1984) a interprété la première *maxime* de quantité comme le résultat du principe d'économie. En effet, l'orateur a tendance à ne donner que les informations qu'il doit donner (« rendez votre contribution nécessaire »), et pas plus. En même temps, il tend également à donner autant d'informations que possible qui, selon lui, facilitent la tâche de l'auditeur en améliorant la clarté du sens de l'énoncé (« rendre votre contribution suffisante »), et pas moins. Il soutient, comme Grice (1975) le laisse entendre, que la deuxième *maxime* de la quantité est davantage liée au fait de ne pas donner des éléments d'information non pertinents, plutôt que de donner réellement le moins d'information possible. Ceci est tout à fait cohérent avec les résultats de l'expérience décrite précédemment, et plus généralement avec la Théorie de la Pertinence (Wilson & Sperber, 2002) elle-même.

Dans le contexte de l'IA en conversations, décrire de manière trop abondante est souvent perçue par les participants comme un comportement quelque peu mécanique et artificiel, alors que le manque d'information, qui peut causer de l'ambiguïté et enfreindre la première *maxime* de quantité, est en fait considéré comme plus humain car il peut être interprété comme un signe de désintérêt, d'une émotion : (Saygin & Cicekli, 2002).

Selon la théorie de la pertinence, dans le cas de la première sous-maxime de quantité, le locuteur donne peu d'informations, empêchant la production de tout effet contextuel sur les représentations mentales de l'interlocuteur. Il est possible qu'il essaie d'inférer une autre signification, ce qui pourrait nécessiter un certain coût cognitif, mais nous pensons qu'il ne serait pas très élevé, d'autant plus qu'une explication plus probable peut être que cette personne s'ennuie, est impolie ou n'est pas bavarde (Saygin & Cicekli, 2002).

Dans notre expérience préliminaire, aucun effet négatif significatif des violations de la première sous-maxime de quantité n'a pu être observé sur l'humanité de l'interlocuteur qui la viole, ce qui est cohérent avec l'observation de Saygin et Cicekli (2002). Dans cette

étude, nous continuerons à nous concentrer sur cette sous-maxime comme référence pour un faible effet machine.

Relation

La troisième **maxime** du principe de coopération de Grice est la **maxime** de relation : « soit pertinent ». Grice l'explique avec l'exemple suivant :

Je m'attends à ce que la contribution de mon partenaire soit appropriée aux besoins immédiats à chaque étape de la transaction ; si je mélange les ingrédients pour un gâteau, je ne m'attends pas à ce que l'on me donne un bon livre, ou même un torchon (mais cela pourra être approprié à une étape suivante)⁹ (Grice, 1975, p. 47).

C'est en fait pour tenter d'approfondir cette **maxime** que Wilson et Sperber (2002) a proposé la théorie de la pertinence comme explication plus générale du **PC** de Grice (1975).

Dans la mesure où les interlocuteurs supposeront toujours par défaut qu'un énoncé est d'une certaine manière pertinent, les violations apparentes de cette **maxime** ont un effet très sensible. Les gens croiront en effet que le locuteur n'est pas à l'aise avec le sujet et veut en changer. Dans le contexte d'un partenaire artificiel, cela sera généralement compris comme un manque de compréhension de la phrase à laquelle la violation répondait, et par conséquent, elle donnera une impression très artificielle : (Saygin & Cicekli, 2002).

Dans notre expérience préliminaire, les violations de cette **maxime** ont eu un effet significatif sur le caractère humain du locuteur de ces énoncés et sur les temps de réponse des énoncés qui les suivent, en particulier pour les femmes. Nous continuerons à utiliser

9. "I expect a partner's contribution to be appropriate to immediate needs at each stage of the transaction ; if I am mixing ingredients for a cake, I do not expect to be handed a good book, or even an oven cloth (though this might be an appropriate contribution at a later stage)"

les violations de cette **maxime** dans notre expérience comme référence pour un fort effet de machine.

Manière

Enfin, la quatrième **maxime** de Grice (1975), la **maxime** de manière, est définie par quatre sous-maximes : 1) « évite les propos obscurs », 2) « évite l’ambiguïté », 3) « sois bref (évite la prolixité inutile) », et 4) « sois ordonné ». La **maxime** de manière se réfère à la structure de l’énoncé et aux déductions qui en découlent. Parmi ses sous-maximes, nous trouvons « sois ordonné », qui se rapporte à l’ordre des éléments d’information donnés dans l’énoncé. Il n’est cependant pas toujours évident de savoir quel ordre doit être considéré comme correct. Il pourrait très bien s’agir de placer les causes avant les conséquences, comme dans l’exemple suivant :

A Ils ont pris une douche et sont sortis.

B Ils sont sortis et ont pris une douche.

Dans ce cas, il est clair que A et B n’ont pas exactement le même sens. Pourtant, l’effet de l’ordre n’est pas toujours aussi visible, comme dans le cas suivant :

A Il pleut dehors, elle a pris son parapluie.

B Elle a pris son parapluie ; il pleut dehors.

Blackmore et Carston (2005), dans leur article sur le connecteur *et*, ont suggéré que le maintien d’un ordre chronologique n’est pas nécessairement requis dans certaines situations de causalité, et que des ordres différents peuvent impliquer, entre autres, un aspect de surprise, comme dans leur exemple :

A Paul ne sait pas écrire et il est linguiste.

B Paul est linguiste et il ne sait pas écrire.

Dans les deux cas, la deuxième partie qui suit le *et* est celle qui déclenche l’effet de surprise, car dans A, on ne s’attend pas à ce que Paul soit linguiste puisqu’il ne sait pas écrire, et

dans B, on ne s'attend pas à ce qu'il ne sache pas écrire puisqu'il est linguiste. Wilson et Sperber (2012) a suggéré que le sens inféré d'un énoncé avec des structures similaires à celles-ci est produit à partir d'un appel à nos connaissances sur les probabilités d'une causalité entre les différents éléments de l'énoncé.

Dans notre expérience, une attention particulière sera accordée à la quatrième sous-maxime de manière, qui n'a pas été étudié dans notre expérience préliminaire.

Test de Turing

Malheureusement, il est assez difficile d'étudier les aspects pragmatiques du langage dans un cadre écologique. Les études ne sont en effet que qualitatives puisque la production de protocoles opérationnels pour une étude quantitative peut s'avérer assez complexe en raison de la variabilité inhérente des conversations : (Alba Juez, 1995 ; Blackmore & Carston, 2005 ; Cohen, 1971 ; Herring, 2013, voir par exemple). Nous pensons que les récentes avancées dans le domaine de l'IA offrent un nouveau cadre dans lequel des méthodes originales peuvent être conçues pour étudier la nature des conversations. Les entités conversationnelles artificielles sont en effet assez courantes aujourd'hui, ce qui en fait des outils potentiellement utiles pour étudier le comportement humain et, spécifiquement pour notre étude, les participants peuvent désormais s'attendre à devoir interagir avec elles.

Le **test de Turing (TT)** (Turing, 1950) est l'un de ces protocoles impliquant une IA. Il consiste à faire jouer le rôle de juge à des participants qui tentent de trouver la machine entre deux interlocuteurs. Comme la tâche tourne essentiellement autour de la capacité du participant à comparer deux interlocuteurs dans une conversation textuelle interactive, elle peut être utilisée pour explorer les caractéristiques conversationnelles auxquelles on peut s'attendre lors d'une conversation avec un humain. Pour cette raison, nous suggérons que l'implication d'une véritable IA peut devenir superflue dans ce contexte, puisque les participants doivent seulement croire qu'ils vont converser avec une IA.

La possibilité de parler à une IA aujourd'hui n'est en effet plus nulle en raison de l'ampleur qu'elle a prise dans de nombreux aspects de notre vie, comme les assistants personnels tels que Siri ¹⁰, Cortana ¹¹, Watson ¹² et bien d'autres.

Turing (1950) décrit l'idée d'un test capable de répondre à la question « Les machines peuvent-elles penser ? ». Dans son article, il a proposé un test analogue à un jeu de devinette sur le genre, où deux personnes (A et B) de sexes différents tentent de convaincre le participant C qu'elles sont une femme alors que le participant sait qu'une seule femme est présente. C s'engage dans une conversation textuelle avec chaque personne, avant de deviner lequel des deux personnages auxquels il a parlé est effectivement la femme entre A et B. Turing a suggéré de remplacer A par une machine. Il convient de noter que le TT était initialement supposé tester l'intelligence des machines et qu'il a fait l'objet de nombreuses critiques à ce sujet. Ces critiques ne s'appliquent pas au contexte de notre étude puisque l'intelligence de l'interlocuteur n'est pas ce qui nous intéresse ici. Ce qui est essentiel, c'est l'utilité du test pour évaluer le caractère humain du comportement de ce partenaire de conversation. Certains aspects peu clairs du TT initial doivent cependant encore être débattus. En particulier, la question de savoir si le participant doit ou non être conscient de la présence d'une IA dans le test, car l'interprétation d'un comportement non humain pourrait en dépendre.

Saygin et Cicekli (2002) ont en effet montré que les participants à qui l'on demande d'élaborer sur leurs sentiments subjectifs à l'égard des productions d'une IA (sans leur donner l'information qu'une IA était en fait présente) ont tendance à l'identifier comme un comportement étrange, mais toujours humain : « Certaines de ces personnes sont-elles des malades mentaux ? » (Saygin & Cicekli, 2002, p. 250). Dans leur expérience, un groupe de participants doit, dans une première phase, répondre à un questionnaire sur la violation de maximes de Grice (1975), puis reçoivent un autre questionnaire concernant le comportement plus ou moins artificiel de l'un des interlocuteurs au cours d'une

10. Assistant personnel développé par Apple

11. Assistant personnel développé par Microsoft

12. Intelligence artificielle développée par IBM

conversation. L'autre groupe reçoit la même tâche dans l'ordre inverse. Dans le premier groupe, les réponses au second questionnaire étaient beaucoup plus radicales que dans le second groupe, ce qui indique que le fait de comprendre les *maximes* aide à déterminer le caractère humain du comportement d'un agent conversationnel.

Pourtant, la violation de certaines *maximes* semble avoir une influence positive sur l'humanité d'un interlocuteur pour les participants :

Parfois la violation de *maximes* peut créer un sentiment d'humanité. En fait, des violations fortes de [la *maxime* de manière] ont invariablement créées des impressions favorables. Il peut être inféré que, si les programmes qui utilisent comme stratégie le fait d'être désagréable ou obscure avaient été conçus pour mieux tenir compte des composantes syntactiques du langage naturel, ils auraient semblé très proche d'humains, même si des humains étranges. Si, en addition, le traitement sémantique avait inclus des méthodes pour partiellement tenir compte de la pertinence et de la quantité, certains auraient peut être même pu passer le Test Loebner¹³ (Saygin & Cicekli, 2002, p. 254).

Si aujourd'hui aucun agent artificiel n'a réussi à passer le *TT*, certains présentent des caractéristiques intéressantes. Ces *chatbots* n'ont pas nécessairement besoin d'être capables d'apprendre par eux-mêmes, surtout les plus anciens d'entre eux. Ils utilisent en effet souvent des mots-clés dans les énoncés de leur interlocuteur pour générer une réponse. Les exemples les plus connus de tels programmes sont ELIZA (Weizenbaum, 1966), A.L.I.C.E. (Wallace, 2009) et plus récemment Zo.

Les conversations avec Zo sortent du lot parmi les agents conversationnels. Pourtant, ce *chatbot* est également loin d'être parfaitement capable de suivre les conversations. En effet, en parlant de quelque chose comme la plage, Zo peut dire qu'elle n'aime pas la

13. "Sometimes maxim violations can create a human-like effect. In fact, strong violations of [Manner] have invariably created favorable impressions. It can be inferred that, had the programs that used being rude or obscure as a strategy been more successfully designed to handle the syntactic components of natural language, they would have appeared quite close to human beings, albeit strange ones. If in addition to this, the semantic processing had included ways to partially handle relevance and quantity, some of these might even have passed the Loebner Test"

plage parce qu'elle n'aime pas le sable. Lorsque l'utilisateur répond « Oui, le sable est ennuyeux », la réponse peut être « Vous êtes ennuyeux »¹⁴. Sur ce point précis, A.L.I.C.E. fait légèrement mieux dans certains cas, comme par exemple, si elle parle d'enfants, de frères et sœurs ou d'animaux domestiques, et si l'utilisateur répond « J'en ai deux », le **chatbot** répondra « Comment s'appellent-ils ? », ce qui est une question attendue, mais dira la même chose si le sujet porte sur les ordinateurs au lieu des enfants (ce qui indique clairement qu'il s'agit d'une réponse générique et préprogrammée). L'absence de traitement pragmatique est en effet extrêmement courante chez les agents conversationnels artificiels en général (Jacquet, Masson et al., 2019) malgré des données expérimentales montrant son importance, y compris avec les robots sociaux (Masson, Baratgin & Jamet, 2017b).

Comme les **chatbots** disponibles ne sont pas spécifiques quant aux **maximes** qu'ils violent, nous éviterons d'utiliser une **IA** dans notre expérience afin de dissocier clairement les différents types de violations, et aussi en raison de l'impact potentiel de différents éléments comme le vocabulaire et la grammaire. Nous présenterons tout de même un interlocuteur comme étant une **IA**, car nous souhaitons faire réfléchir les participants sur ce qu'ils attendent d'un humain par rapport à ce qu'ils attendent d'une **IA**.

Temps de réponse

Il est courant en psychologie expérimentale d'enregistrer le délai entre un stimulus et une réponse des participants. Cependant, à notre connaissance, les études portant sur les conversations en ligne qui en font usage restent très rares (Jacquet et al., 2018).

L'idée d'utiliser ces mesures pour développer des interprétations du fonctionnement interne de l'esprit n'est pas nouvelle. Cela repose sur l'hypothèse générale selon laquelle la vitesse de traitement du cerveau humain n'est pas illimitée, notamment car la

14. Cette conversation est tirée d'une conversation publique sur Twitter, disponible à l'adresse suivante : <https://twitter.com/zochats/status/1009141014827761664>. Il est possible de discuter avec le **chatbot** directement sur la même page, par le biais de messages privés.

communication entre les neurones n'est pas immédiate. En effet, la quasi-totalité des synapses du système nerveux central utilise une libération chimique de neurotransmetteurs qui peut prendre individuellement une demi-milliseconde par connexion, mais qui varie en fonction du type de synapses et d'autres facteurs (Katz & Miledi, 1965).

Le délai entre le stimulus et sa réponse est généralement appelé temps de réaction, et se mesure en millisecondes : (Deary & Der, 2005). Dans notre cas, puisque nous n'enregistrons pas le temps entre un stimulus et une réaction, mais entre un stimulus et une réponse écrite, nous avons choisi d'utiliser le terme **temps de réponse (TR)** pour éviter toute ambiguïté. Nous nous attendons à des délais de l'ordre de la seconde.

Nous ne prétendons pas que la mesure des temps de réaction ou des temps de réponse soit la représentation idéale de ce qui se passe réellement dans le cerveau lors du traitement des phrases, ni qu'il s'agisse d'une mesure idéale en soi, et il faut donc éviter de sur-interpréter les valeurs absolues de ces mesures. Elles n'en restent pas moins un outil très écologique du fait qu'elles ne nécessitent pas d'appareil d'enregistrement dédié et peuvent être utilisées pour des données à forte variabilité, contrairement à des mesures de temps plus précises comme l'électroencéphalogramme. Elles sont souvent suffisantes pour démontrer l'impact de divers facteurs sur le traitement de l'information (Bowyer et al., 2009 ; Fitts, 1966 ; Lachman et al., 1974 ; Thorpe et al., 1996, par exemple).

Un autre facteur critique à considérer est que les temps de réaction peuvent varier avec l'âge des participants dans de nombreuses tâches. Par conséquent, l'âge ne doit pas être ignoré en tant que facteur potentiel lors de la mesure des temps de réaction, en particulier avec des participants de moins de 15 ans (Deary & Der, 2005 ; S. Hale, 1990). Nous devrions adopter des considérations similaires pour les mesures des **TR**.

Dans cette expérience, nous considérons les **TR** entre les énoncés d'un interlocuteur et ceux du participant comme un indicateur du coût cognitif du traitement de l'énoncé de l'interlocuteur. Les données ne seront recueillies qu'auprès de participants âgés de plus de 18 ans afin d'éviter les biais potentiels liés à l'âge comme décrit dans le paragraphe

précédent.

5.3 Expérience

Matériel et Méthode

Cette expérience suit un protocole strictement similaire à celui introduit dans une expérience préliminaire, avec l'ajout de la condition des violations de la quatrième *maxime* de Manière, et avec des participants supplémentaires dans toutes les conditions (Jacquet et al., 2018).

Participants

86 personnes de langue maternelle anglaise, habituées aux conversations textuelles par le biais de logiciels de messagerie (Skype, Telegram, Messenger ou autres) ont accepté de participer.

La plupart d'entre elles vivaient en Amérique du Nord (48), et en Europe (31). Trois vivaient en Australie, deux en Afrique et deux en Amérique centrale. Ces personnes ont été recrutés grâce à l'aide de contacts sur les différents continents. Ces contacts devaient trouver une ou plusieurs personnes volontaires, de sexe différent dans la mesure du possible.

Tous les participants étaient des adultes âgés de 18 à 45 ans ($M = 25, SD = 5, 7$). 46 d'entre eux étaient des hommes, tandis que 40 étaient des femmes.

Les participants avaient des formations différentes afin d'éviter de pouvoir généraliser autant que possible. Les résultats de notre questionnaire sur cette question ont révélé des origines très variées dans les domaines généraux des Arts, des Sciences et des Services. 24 n'ont pas répondu à cette question.

Variables

Facteur - Violations des maximes Le principal facteur discriminant entre nos conditions était le type de **maxime** gricéenne qui était violée (la première **maxime** de quantité, la **maxime** de relation et la quatrième **maxime** de manière) par l'expérimentateur (appelé acteur) au cours des conversations avec chaque participant. L'ordre de conversation (acteur IA en premier ou acteur humain en premier) et le sexe des participants ont également été considérés comme des facteurs potentiels et ont été contrôlés.

Variable Principale - Temps de réponse Notre principale variable dépendante était le délai en secondes (le **TR**) entre le moment où l'acteur envoyait un message et le moment où les participants envoyaient leur réponse.

Comme la longueur des messages pouvait potentiellement influencer le **TR**, nous avons conçu une correction mathématique qui a été appliquée à tous les temps de réponse enregistrés afin d'éliminer le retard probablement causé par le nombre de caractères des phrases. Nous avons utilisé un modèle de régression linéaire multiple (avec interaction) entre la longueur du message de l'acteur et la longueur de la réponse du participant sur les délais observés. Le modèle a été déduit des discussions entre le participant et l'acteur humain (qui n'avaient pas de violations intentionnelles des **maximes** de Grice) pour créer des prédictions du délai tel qu'il devrait être sans violations.

Ce modèle nous a permis de calculer un délai théorique (D) pour chacune des réponses des participants.

$$D = (w \times C_e) + (x \times C_p) + (y \times C_c C_e) + z \quad (5.1)$$

Dans lequel C_e est la longueur du message de l'expérimentateur, et C_p la longueur du message du participant (tous deux en nombre de caractères). w , x , y et z sont les coefficients du modèle.

Ce délai théorique est ensuite retranché au délai observé. La différence qui en résulte est ensuite utilisée pour tester nos hypothèses et représente l'écart par rapport aux temps de réponse attendus (de type humain).

$$\Delta d = d - D \quad (5.2)$$

où d est le retard observé.

Variable secondaire - Pourcentage d'identification Le pourcentage d'identification correcte de l'acteur IA a également été enregistré pour être comparé à 50% (chance aléatoire). Une valeur supérieure au hasard signifiant que l'acteur était perçu comme une machine. Une valeur autour de 50% indiquant que l'acteur est reconnu comme étant de type humain (les participants étant alors incapables de distinguer correctement les deux acteurs).

Variables contrôles D'autres variables sont enregistrées afin de contrôler tout biais potentiel. Ces variables de contrôle comprennent le genre des participants, leur âge, la durée de chaque conversation (en secondes), les connaissances auto-évaluées sur l'IA (échelle de Likert de 1 à 7) et les connaissances auto-évaluées sur l'informatique (échelle de Likert de 1 à 7), et enfin la confiance dans leur réponse au TT (échelle de Likert de 1 à 7).

Procédure

L'expérience ayant lieu entièrement en ligne, les participants n'ont besoin que d'un ordinateur avec une connexion internet pour pouvoir participer et peuvent rester à leur domicile. Pour éviter de fausser l'enregistrement des TR, les participants ne pouvaient pas utiliser d'appareils mobiles pour participer à cette étude, car la vitesse de frappe peut bien sûr être influencée par le type d'appareil utilisé. Le chat où se déroule l'expérience est hébergé sur un serveur privé français et est conçu spécifiquement pour cette étude.

Les participants doivent communiquer sur des échelles de Likert leurs connaissances en IA et en informatique, leur domaine d'étude ou leur carrière, leur genre et leur âge. Ces informations n'étaient envoyées au serveur que si le consentement était explicitement donné pour participer à l'expérience.

Une fois dans l'interface de conversation, l'expérimentateur (affiché comme modérateur) explique en détail les règles que les participants doivent respecter pendant les conversations ainsi que leur tâche.

Toutes les conversations ne peuvent durer plus de 15 minutes, pendant lesquelles les participants peuvent décider d'arrêter la conversation s'ils ont deviné avec quel acteur (IA ou humain) ils parlent.

Les deux acteurs (IA et humain) portent le même nom (Andrew) et essayent tous deux d'incarner le même personnage fictif. L'objectif étant d'inviter les participants à poser des questions au personnage fictif au lieu d'interroger les acteurs eux-mêmes. Ainsi, les deux acteurs fournissent la même information sémantique au participant (puisque le personnage qu'ils incarnent est le même), mais de manière différente.

Comme nous l'avons déjà mentionné, il n'y a en fait aucune IA dans notre étude. Les deux acteurs sont, en fait, la même personne (un étudiant en psychologie expérimentale). Les deux rôles ne diffèrent que par leur comportement concernant les violations des *maximes*. L'expérimentateur dans le rôle humain doit se comporter comme un humain « normal », sans ajouter volontairement des violations. Dans le rôle de l'IA, l'expérimentateur est contraint dans son comportement, et ne peut pas répondre normalement, mais doit suivre des directives conçues pour produire autant de violations que possible du type requis pour la condition, et par conséquent pour changer le sentiment d'humanité donné au participant, puisque nous nous attendions à ce que ce comportement contraint soit plus perçu comme étant plus proche de celui d'une IA. Pour chaque phrase, l'acteur indique (à l'insu du participant) si une violation volontaire a été introduite en raison de ces contraintes.

Le choix de ne pas utiliser une IA réelle pendant ce test a été motivé par le désir de limiter les différences entre nos conditions à celles que nous pouvions contrôler. Des différences de vocabulaire, des problèmes de syntaxe ou une compréhension défectueuse des messages du participant auraient pu ajouter d'autres violations au cours des conversations qu'il aurait été difficile de prévoir dans notre protocole. C'est également pour cette raison que nous avons gardé le même expérimentateur pour les deux rôles.

Le TT portant sur des conversations libres, il était impossible de les garder strictement identiques entre les conditions, et même entre les participants d'une même condition. Pour éviter des biais importants, toute nouvelle information donnée sur le caractère d'Andrew n'ayant pas été anticipée est notée afin que la même information puisse être réutilisée dans des conversations différentes. Les participants ont d'ailleurs parfois déclaré être un peu surpris de voir les mêmes informations revenir entre les deux conversations.

Les participants peuvent discuter avec les deux interlocuteurs (dans un ordre aléatoire), et chaque participant est affecté à l'une des trois conditions de manière aléatoire (voir les conditions ci-dessous). L'utilisation de smileys, de liens hypertextes et de double messages n'est pas autorisée. Les messages (des participants et des acteurs) ne peuvent contenir plus de 255 caractères.

Conditions

Le Tableau 5.1 montre le résumé des conditions et de leurs effets attendus.

	Humain	Machine	Hypothèse
Quantité	Pas de Violation	Information Insuffisante	Effet faible
Relation	Pas de Violation	Phrases génériques	Effet fort
Manière	Pas de Violation	Phrases désordonnées	Effet moyen

TABLE 5.1 – **Résumé de l'expérience.** Résumé du protocole expérimental associé aux hypothèses correspondantes aux différentes conditions concernant les temps de réponse et les scores dans le Test de Turing.

Quantité Le comportement contraint de l'acteur IA dans cette condition était l'obligation de répondre avec trop peu d'informations par rapport à ce qui serait attendu d'une réponse « normale ».

Participant : Quelque chose de prévu pour ce soir ?

Andrew : Aucune idée. Et toi? (la réponse alternative eut été : « Pas vraiment. Et toi? »)

ou

Participant : Avez-vous des animaux de compagnie ?

Andrew : Peut-être. Et toi? (la réponse alternative eut été : « J'en ai deux, oui. Et toi? »)

Nous nous attendions à ce que cette condition produise des violations de la première **maxime** de quantité¹⁵. Cette contrainte devait ainsi produire une faible augmentation (si tant est qu'il y en ait une) du **TR** des participants, et un faible ratio de bonne identification au **TT** (proche du hasard).

Relation Le comportement contraint de l'acteur IA dans cette condition était d'éviter d'utiliser les informations contextuelles des messages passés. Dans les cas où trop peu d'informations sont fournies par le participant pour générer une réponse pertinente, l'acteur IA ne pouvait répondre que par des messages génériques :

Participant : Aimes-tu lire ?

Andrew : Oui, j'aime ça. Et toi ?

Participant : Oui, j'aime ça.

Andrew : Ok. Qu'est-ce que tu fais ce soir ? (La réponse alternative aurait pu être « Quel genre de livres aimes-tu »)

15. « Fais que ta contribution soit aussi informative que nécessaire (pour les objectifs actuels de l'échange) »

ou

- (a) Andrew : Avez-vous des enfants ?
- (b) Participant : Oui, j'en ai deux.
- (c) Andrew : Cool. Est-ce que tu aimes les jeux ? (La réponse alternative aurait pu être « Cool. Comment s'appellent-ils ? »)

Nous nous attendions à ce que cette condition produise des violations de la **maxime** de Relation¹⁶ et que cette contrainte produise une augmentation plus importante du **TR** des participants ainsi qu'une augmentation du ratio de bonnes identifications de l'**IA** dans le **TT** (en raison de la faible humanité).

Manière Le comportement contraint de l'acteur **IA** dans cette condition consistait à intervertir aléatoirement l'ordre des clauses dans des énoncés complexes en contenant au moins deux. Que ce soit autour de connecteurs comme *et*, *ou*, de virgules, de points ou de tout autre délimiteur de subordonnées individuelles.

Participant : Qu'est-ce que tu fais ?

Andrew : Je vais probablement dormir un peu et ensuite regarder la télévision. Pas grand-chose. (Une autre réponse aurait pu être : « Pas grand-chose. Probablement regarder la télé et ensuite dormir un peu. »)

Nous nous attendions à ce que cette condition produise des violations de la troisième **maxime** de Manière¹⁷. Ainsi, cette contrainte devait produire une augmentation modérée des **TR** des participants, ainsi qu'un ratio intermédiaire de bonne identification au **TT** (entre les deux **maximes** précédentes).

16. « Sois pertinent ».

17. « Sois ordonné ».

Résultats

Au total, 2095 énoncés écrits par les participants ont été enregistrés. 1018 provenaient des discussions avec l'acteur humain, et 1077 provenaient des discussions avec l'acteur IA. Deux participants hommes (Condition : Relation) ont été retirés de l'analyse car leur conversation ne permettait pas à l'acteur IA de produire des violations intentionnelles correspondant à leur condition. Une participante femme (Condition : Relation) a été retirée pour la même raison, et une autre participante (Condition : Manière) a demandé à être retirée de l'étude plus tard. Dans les conversations avec l'acteur IA, 275 énoncés ont été précédés de messages avec des violations de la *maxime* de quantité, 114 ont été précédés de messages avec des violations de la *maxime* de relation, 164 ont été précédés de messages avec des violations de la *maxime* de manières et les 525 restants ne suivaient aucune violation.

Seuls les énoncés provenant des discussions avec l'acteur IA ont été utilisés dans l'ANOVA (Type III).

L'égalité des variances entre les groupes a été testé pour les analyses post-hoc et l'ANOVA en utilisant le test de Fisher d'égalité des variances. Si les groupes avaient des variances significativement différentes, la comparaison des moyennes des TR était analysée avec un test t de Welch, sinon avec un test t à deux échantillons.

Leurs *p-values* ont été ajustées à l'aide de la correction de Holm-Bonferroni pour éviter les erreurs de type I (notées dorénavant p_{Holm}).

Nous avons utilisé des tests du Chi-carré pour comparer les résultats du TT à une distribution hypothétique de 50% de chances d'identifications correctes.

Les échelles de Likert d'auto-évaluation sur les connaissances en IA et en informatique ont été analysées avec le test de Kruskal-Wallis par rangs. Les scores de confiance ont été analysés de la même manière.

L'âge des participants et la durée des conversations ont été analysés en fonction de la condition et du genre avec un test de Kurskal-Wallis également.

Modèle Linéaire

Le modèle généré à partir des conversations avec l'acteur humain a donné lieu à l'équation suivante ($R^2 = 0,4$) :

$$D = (0,15 \times C_e) + (0,36 \times C_p) - (0,0004 \times C_e C_p) + 9,2 \quad (5.3)$$

en remplaçant dans (5.1) w par 0,15, x par 0,36, y par 0,0004 et z par 9,2.

Variables Contrôles

Il n'y avait aucune influence des conditions sur le score de confiance rapporté par les participants ($\chi^2(2, N = 86) = 0,04; p = 0,98$), mais il y avait une légère tendance à ce que les femmes rapportent des scores de confiance plus faibles que les hommes ($\chi^2(1, N = 86) = 3,19; p = 0,07$).

Dans toutes les conditions, il n'y avait pas de différence significative dans la distribution des connaissances en IA ($\chi^2(1, N = 86) = 1,4; p = 0,5$) et en informatique ($\chi^2(1, N = 86) = 0,3; p = 0,86$), mais les participantes ont rapporté des scores de connaissances en informatique significativement plus faibles que les participants masculins ($\chi^2(1, N = 88) = 8,3; p < 0,01$).

La distribution des âges n'était pas significativement différente entre les conditions ($\chi^2(2, N = 86) = 2,2; p = 0,32$) ou entre les genres ($\chi^2(1, N = 86) = 2,9; p = 0,08$).

Enfin, la durée des conversations lorsque les participants s'adressaient à l'acteur IA n'était pas significativement différente entre les conditions ($\chi^2(2, N = 86) = 2,8; p = 0,25$) ni entre les genres ($\chi^2(1, N = 86) = 0,35; p = 0,55$).

Temps de réponse

Nous n'avons pas détecté de double interaction entre le type de violations, l'ordre de la conversation et le genre sur Δd ($F_{3,1062} = 1,3; p = 0,26$), mais une interaction significative a été trouvée entre le type de violations et le genre ($F_{3,1070} = 3,8; p < 0,01$).

En ce qui concerne les effets principaux, seul le type de violation a influencé de manière significative Δd ($F_{3,1070} = 7,26; p < 0,001$), mais le sexe avait également tendance à avoir une influence ($F_{1,1070} = 3,84; p = 0,07$).

En raison de l'interaction entre le sexe du participant et le type de violation, et compte tenu de la différence significative des connaissances en informatique auto-évaluées entre les genres (que nous avons remarquée plus haut), nous avons également effectué une

	<i>Moyenne(s)</i>	$\sigma(s)$	<i>ES(s)</i>	<i>Nombre</i>	<i>t</i>	<i>pHolm</i>	
Pas de Violation							
Général	-1.20	17	0.75	525			
Hommes	-1.36	17	1.0	276			
Femmes	-1.02	18	1.1	249			
Quantité							
Général	0.937	19	1.2	275	-1.54	.12	
Hommes	3.11	21	1.8	143	-2.19	.05	*
Femmes	-1.41	17	1.5	132	0.21	-	
Relation							
Général	7.57	24	2.2	114	-3.75	.001	***
Hommes	3.88	16	2.1	62	-2.23	.05	*
Femmes	12.0	30	4.1	52	-3.05	.01	**
Manière							
Général	2.68	21	1.6	164	-2.16	.06	
Hommes	5.48	23	2.6	78	-2.48	.05	*
Femmes	0.15	19	2.0	86	-0.52	-	

TABLE 5.2 – **Effet des violations et du genre sur Δd (en secondes)**. Pas de Violation indiquent les cas où les messages des participants dans les conversations avec l'acteur IA suivaient des messages sans violations (contrôle). Le *t* et le *pHolm* comparent les moyennes de chaque groupe avec le contrôle associé (par exemple, Quantité Hommes avec Pas de Violation Hommes.)

autre ANOVA pour tester l'effet potentiel de l'interaction entre le type de violation et les connaissances en informatique auto-évaluées (Faible : ≤ 3 , correspondant à moins du premier quartile, et Élevé : ≥ 5 , correspondant à plus du troisième quartile). L'influence de cette interaction sur Δd n'était pas significative ($F_{3,634} = 0.266; p = 0,85$), et nous n'avons pas pu trouver d'effet des connaissances en informatique ($F_{1,640} = 1,14; p = 0,29$).

Le résumé des analyses sur l'interaction entre le type de violation et le genre du participant sur les TR est disponible dans le Tableau 5.2.

Général Le Δd était significativement plus long après des énoncés avec une violation de la **maxime** de relation qu'après des énoncés sans violation ($t(140) = -3,75; p_{Holm} < 0,001$), tandis que nous avons remarqué une forte tendance à un Δd plus long pour les violations de la **maxime** de manière ($t(237) = -2,16; p_{Holm} = 0,06$), et aucun effet significatif des violations de la **maxime** de quantité ($t(504) = -1,54; p_{Holm} = 0,12$).

Interaction avec le genre Le diagramme en boîtes de cette interaction est présenté dans la Figure 5.1.

Aucune différence significative n'a été trouvée selon le genre après application de la correction de Holm-Bonferroni, ni pour les violations de la **maxime** de quantité ($t(268) = -1,96; p_{Holm} = 0,20$), ni pour les violations de la **maxime** de relation ($t(76) = 1,76; p_{Holm} = 0,25$), ni pour les violations de la **maxime** de manière ($t(162) = -1,64; p_{Holm} = 0,25$).

Messages des participants femme Le Δd était significativement plus long après des messages avec une violation de la **maxime** de relation qu'après des messages sans violation ($t(59) = -3,05; p_{Holm} = 0,01$).

Les autres différences n'étaient pas statistiquement significatives.

Messages des participants homme Chez les hommes, toutes les conditions ont augmenté les TR par rapport aux messages sans violations : pour les violations de la *maxime* de quantité ($t(236) = -2,19; p_{Holm} = 0,05$), pour les violations de la *maxime* de relation ($t(336) = -2,23; p_{Holm} = 0,05$), ainsi que pour les violations de la *maxime* de manière ($t(102) = -2,48; p_{Holm} < 0,05$).

Test de Turing

Dans la section suivante, tous les pourcentages d'identification corrects sont comparés au hasard (50 %). Un résumé est présenté dans le Tableau 5.3.

En général, les participants ont réussi à identifier facilement l'acteur IA (63%, $\chi^2 = 6,7; p < 0,01$). Pourtant, cela variait en fonction de la condition et du genre des participants.

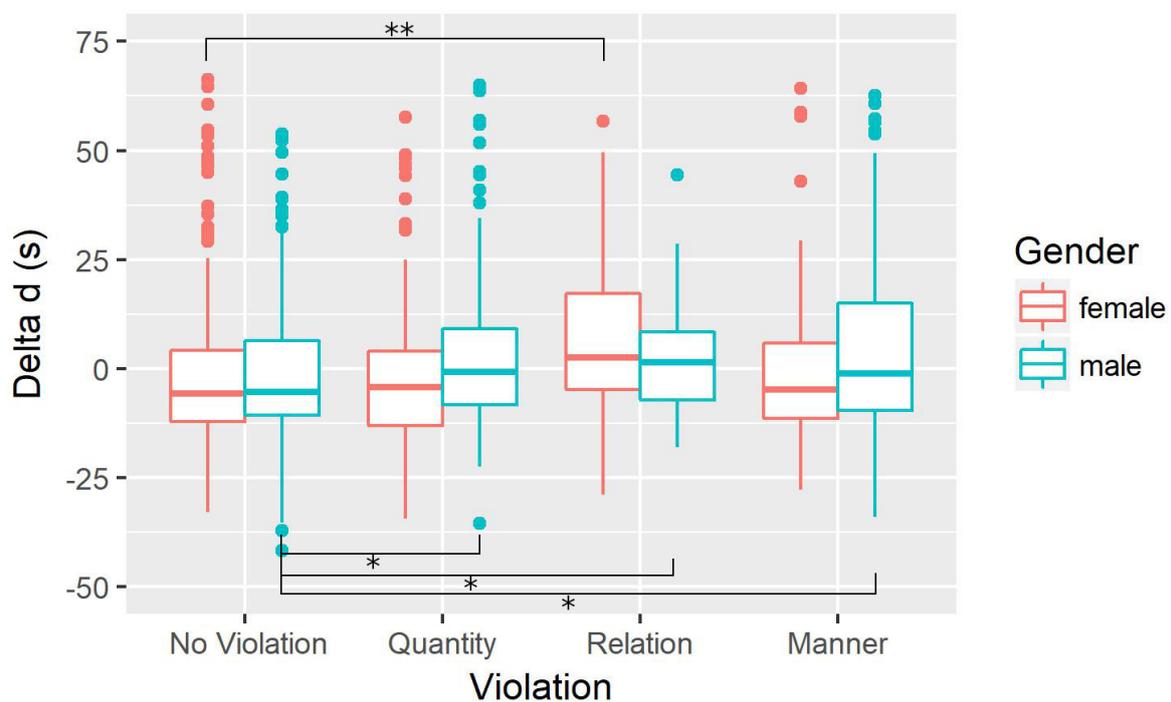


FIGURE 5.1 – **Effet des violations et du genre sur Δd (en secondes)**. « No Violation » indique les cas où les messages des participants dans les conversations avec l'acteur IA suivaient des messages sans violations (contrôle).

** $p < .01$, * $p < .05$

Pour la **maxime** de quantité, il n'y a pas de différence significative par rapport au hasard chez les hommes (44%, $\chi^2 = 0,25$; $p = 0,62$), mais une légère tendance chez les femmes (71%, $\chi^2 = 2,57$; $p = 0,10$).

Pour la **maxime** de relation, il n'y a pas de différence significative par rapport au hasard pour les hommes (56%, $\chi^2 = 1,6$; $p = 0,20$), mais il y a une différence significative pour les femmes (86%, $\chi^2 = 6,2$; $p < 0,01$). En général, la différence par rapport au hasard est également significative (74%, $\chi^2 = 6,3$; $p < 0,01$).

Pour la **maxime** de manière, il y a une différence significative par rapport au hasard pour les hommes (75%, $\chi^2 = 4,0$; $p < 0,05$), mais aucune pour les femmes (46%, $\chi^2 = 0,1$; $p = 0,78$).

	Correct	Erroné	Pourcentage	χ^2	p	
Total	55	31	64%	6.7	.01	**
Quantité						
Général	17	13	57%	0.53	.47	
Hommes	7	9	44%	0.25	.62	
Femmes	10	4	71%	2.6	.10	
Relation						
Général	20	7	74%	6.3	.01	**
Hommes	9	5	64%	1.6	.20	
Femmes	11	2	85%	6.2	.01	**
Manière						
Général	18	11	62%	1.7	.19	
Hommes	12	4	75%	4.0	.05	*
Femmes	6	7	46%	0.077	.78	

TABLE 5.3 – Effet des violations et du genre sur les résultats dans le **TT**. Correct désigne les participants qui ont identifié avec succès l'acteur **IA**, Erroné désigne les autres participants.

5.4 Discussion

De manière générale, il convient de noter que même si l'ANOVA a détecté une interaction significative entre le genre et le type de violation, l'analyse post-hoc n'a pas permis de mettre en évidence précisément dans quelle condition les différences entre les sexes ont contribué à cette interaction. En conséquence, les différences de genre considérées ci-dessous doivent être prises avec précaution, et uniquement en tant qu'hypothèses.

Bien que nous ne nous attendions pas à trouver des différences en fonction du genre, nous avons inclus ce facteur dans notre protocole et dans la discussion des résultats ci-dessous car un effet du genre a été précédemment observé avec des robots sociaux dans des expériences d'interaction homme-machine (Park et al., 2011; Powers et al., 2005; Sussman & Tyson, 2000; Tay et al., 2014) et était également présent dans notre expérience préliminaire (Jacquet et al., 2018). Si l'intensité de l'effet dépend de l'étude, les différences de genre lors de l'interaction avec un partenaire conversationnel artificiel semblent subsister, d'autant plus que le comportement semble différent selon le genre affiché de l'agent artificiel lui-même. Dans notre cas, l'agent artificiel a été présenté comme un partenaire masculin : Andrew.

En tout cas, si de telles différences entre les genres existent dans notre étude, elles ne sont pas dues à une connaissance potentiellement plus grande de l'informatique chez les hommes par rapport aux femmes, car il n'y a pas eu d'interaction ni d'effet principal de ce facteur sur les TR.

Quantité

Les résultats de l'expérience sont pour cette condition partiellement cohérents avec notre hypothèse selon laquelle les violations de la première *maxime* de la quantité ne déclencheront qu'un petit effet. En effet, en général, il n'y a eu qu'une légère augmentation

(qui n'était pas significative) du **TR**, principalement portée par les participants masculins (significative, alors qu'il n'y avait pas de différence significative pour les femmes).

Il est intéressant de noter que le pourcentage d'identification correcte des hommes dans cette condition était en fait pire que le hasard (taux de réussite de 44%, contre 71% pour les femmes dans la même condition), ce qui signifie que les hommes trouvaient que l'acteur machine ressemblait plus à un humain, ne serait-ce que légèrement (bien que la différence par rapport au hasard ne soit pas significative).

Ces conclusions sont cohérentes avec les résultats rapportés par (Saygin & Cicekli, 2002). En effet, dans leur expérience, les violations de la **maxime** ont produit un sentiment d'humanité plutôt qu'un sentiment artificiel. Par conséquent, on peut probablement supposer que les interlocuteurs sont plutôt habitués à des contributions qui ne fournissent pas autant d'informations que prévu, du moins dans les conversations qui ne sont pas axées sur une tâche. De telles violations ont donné aux participants l'impression que leur partenaire de conversation s'ennuyait, était contrarié ou n'était pas très bavard, mais pas qu'il était artificiel.

En ce qui concerne les participants masculins, leurs résultats sont plutôt cohérents avec ceux de Engelhardt et al. (2006) qui indiquaient que les violations de cette **maxime** pouvaient confondre l'interlocuteur. Cependant, leurs résultats pourraient ne pas être directement transposables à ceux de notre protocole expérimental, car notre expérience n'impliquait pas la nécessité pour les participants d'agir sur les informations données, mais plutôt de deviner le sentiment général d'humanité de l'acteur.

Si les **TR** sont effectivement corrélés au coût cognitif requis pour inférer la signification d'un énoncé, cela signifierait qu'en général, ce coût cognitif est en effet très faible ($\Delta d = 0,937s$; $SD = 19s$), comme nous l'avions prédit. Pourtant, puisque les hommes ne semblent pas percevoir cette augmentation comme un signe de manque d'humanité, nous pouvons interpréter qu'une augmentation du coût cognitif ne signifie pas directement une diminution de l'humanité d'un partenaire en conversation qui pourrait également

dépendre du contexte dans lequel cette augmentation se produit.

Relation

Les participants féminins étaient beaucoup plus sensibles aux violations de la *maxime* de relation que les hommes, car elles ont montré la plus grande augmentation de leurs *TR* dans cette condition. Les *TR* enregistrés étaient également plus élevés ($M = 3,88s$; $SD = 16s$) chez les hommes, mais dans une moindre mesure que chez les femmes ($M = 12,0s$; $SD = 30s$).

A notre connaissance, une différence entre les genres dans ce contexte n'a pas été discutée explicitement dans la littérature, ni pour la confirmer ni pour la réfuter. Pourtant, les femmes ont tendance à poser plus de questions que les hommes dans les conversations (Fishman, 1980), et elles ont également tendance à être brèves par rapport aux questions des hommes (Winter, 1993). Cela expliquerait pourquoi les femmes pourraient remarquer les violations de la *maxime* de relation plus que les hommes dans notre expérience, d'autant plus qu'il a été démontré que les filles sont généralement en avance sur les garçons dans les compétences linguistiques, ce qui semble d'ailleurs s'accroître avec l'âge (Eriksson et al., 2012). Alors que l'acteur *IA* pouvait facilement répondre à des questions détaillées sans violer la *maxime* de la relation, il était incapable de répondre à des questions courtes s'appuyant sur le contexte de la conversation. En effet, dans les situations où de telles violations pouvaient se produire, il fallait que les participants fassent référence à des messages ou à des idées antérieurs sans pour autant inclure dans leurs messages une indication explicite de ce dont ils parlaient, comme on peut le voir dans cet exemple¹⁸ (participant masculin) :

A Andrew : Quelque chose de prévu pour ce soir ?

B Participant : Jouer avec mes enfants pendant quelques heures après mon retour du

18. Ce participant a accepté que des parties de sa conversation soient rendues publiques après l'expérience.

travail, avant de me coucher, puis regarder Big Brother Canada avec ma femme.

C Andrew : Cela semble assez agréable. Combien d'enfants avez-vous ?

D Participant : 3

E Andrew : Cool. Qu'est-ce que vous aimez faire en général ?

Dans cet exemple, l'acteur IA n'avait plus aucune idée de ce à quoi le « 3 » de *D* faisait référence. Comme il n'avait aucune idée non plus de ce dont ils avaient parlé auparavant, il a essayé de donner une réponse générique à « 3 » : « Cool », et a continué à faire avancer la conversation avec une autre question générique : « Qu'est-ce que vous aimez faire en général ? ». À cela, l'acteur humain aurait plutôt répondu « Cool, quel âge ont-ils ? » ou « Cool, comment s'appellent-ils ? ».

Alors que dans l'exemple précédent, l'acteur IA ne s'est trouvé qu'une seule fois dans une situation difficile, les conversations avec les participantes pouvaient générer ces situations beaucoup plus souvent ¹⁹ :

A Participant : Je cuisine.

B Andrew : Oh cool, qu'est-ce que tu fais ?

C Participant : Tu peux deviner ?

D Eh bien, ce n'est pas vraiment facile.

E Un participant : Laissons faire le hasard, qu'est-ce qui te vient à l'esprit ?

F Andrew : Je ne suis pas vraiment sûr pour être honnête.

Dans cet exemple, l'acteur IA se trouve dans une situation difficile à la fois en *D* et en *F*. En effet, comme il ne peut pas utiliser le contexte de la conversation, il ne sait plus ce qu'il doit deviner en *D* et en *F*. Dans ce contexte, à titre de comparaison, l'acteur humain aurait au moins essayé de donner des noms d'aliments.

Les TR montrent que ces violations ont, en général, un effet fort, ce qui est cohérent avec l'idée que les TR sont corrélés avec le coût cognitif requis, dans notre cas, pour

19. Cette participante a accepté que des parties de sa conversation soient rendues publiques après l'expérience.

changer de sujet ou continuer la conversation après des réponses décontextualisées. Les résultats au **TT** confirment ce qui a été observé par Saygin et Cicekli (2002) : En effet, la grande majorité des participants ont considéré l'acteur produisant ces violations comme moins humain (74% d'identifications correctes au total, 85% d'identifications réussies pour les femmes, et 64% pour les hommes, bien que ces derniers ne soient pas significativement différents du hasard).

Une autre mesure de la charge cognitive au sein des énoncés existe dans la littérature : la surprise (J. Hale, 2001 ; Levy, 2008). Dans ces études, la surprise est le plus souvent utilisée au niveau du mot. La surprise est, dans ce cas, la probabilité logarithmique négative de ce mot compte tenu de son contexte, en particulier de son contexte syntaxique. Il est tout à fait possible que les participants ne prédisent pas seulement les mots qu'ils s'attendent à voir dans une phrase mais aussi, à un niveau plus élevé, les réponses qu'ils s'attendent à recevoir au sein d'une conversation, compte tenu du contexte de la conversation elle-même, selon le même principe : comme la probabilité logarithmique négative d'un message. Cela pourrait expliquer pourquoi les violations de la **maxime** de relation ont la plus grande augmentation du **TR**, puisque les réponses données par l'acteur dans cette condition sont les plus éloignées de ce que l'on aurait pu attendre d'un partenaire de conversation puisqu'il n'hésitait pas à changer de sujet. Un effet pragmatique similaire a été montré lors de l'étude de la négation (Nordmeyer & Frank, 2015), montrant que les négations attendues n'étaient pas aussi difficiles à traiter que les négations inattendues : dans cette expérience, il y avait une forte corrélation entre le **TR** et la surprise.

Manière

Un effet significatif a été trouvé dans le **TR** dans cette condition, mais seulement pour les hommes qui avaient le **TR** le plus élevé en moyenne ($M = 5,48s$; $SD = 23s$), et les résultats dans le **TT** montrent un effet similaire, puisque les hommes ont réussi à identifier l'acteur **IA** significativement mieux qu'un hasard dans cette condition particu-

lière.

Ceci était au moins partiellement cohérent avec notre hypothèse : cette condition produisant un effet intermédiaire entre les deux conditions précédentes. Pourtant, il semble que cette condition ait eu peu d'effet sur les femmes, malgré le fait que tous les participants qui ont réussi à identifier l'acteur IA dans cette condition ont fait des commentaires du type « C'est juste que le flux faisait un peu bizarre et quelque peu déconnecté ».

Il est possible que cela soit dû à l'absence de violation réelle dans certains cas d'inversion²⁰ :

A Andrew : Vous jouez à ça ?

B Participant : Avant, mais plus maintenant.

C Andrew : Qu'est-ce qui t'a fait arrêter ? Très bien.

En effet, dans *C*, la différence entre « Qu'est-ce qui t'a fait arrêter ? Très bien » et « Très bien. Qu'est-ce qui t'a fait arrêter ? » n'est pas forcément très frappante pour le lecteur. Dans d'autres cas, les violations de la troisième maxime de manière ont fini par amener les participants à mal interpréter ce que l'acteur disait²¹ :

A Participant : Je vais bien, merci, y a-t-il quelque chose dont tu voudrais parler ?

B Andrew : Qu'est-ce que tu fais ? Je ne suis pas vraiment sûr en fait.

C Participant : C'est un peu la même chose ici, donc... Quel est ton passe-temps actuel ?

Dans ce cas, il est évident de voir que le flux de la conversation a été perturbé par la permutation des deux clauses de l'acteur IA autour du point d'interrogation, et donc qu'une violation de la maxime de manière a été produite, créant une ambiguïté. En effet, si nous prenons *B* dans son ordre séquentiel affiché, il a poursuivi la conversation avant de répondre à la question du participant.

20. Cette participante a accepté que des parties de sa conversation soient rendues publiques après l'expérience.

21. Cette participante a accepté que des parties de sa conversation soient rendues publiques après l'expérience.

Nous pouvons avoir l'impression que ce qu'Andrew a dit, c'est que, premièrement, il a ignoré la question posée par le participant « Y a-t-il quelque chose dont tu voudrais parler ? » pour demander directement « Qu'est-ce que tu fais ? ». Le reste de sa réponse pourrait signifier qu'il n'était pas sûr de ce que le participant pouvait être en train de faire, ou qu'Andrew ne savait pas ce qu'il allait faire lui-même. Cette dernière option semble être la plus probable au vu de la réponse du participant, puisqu'il a poursuivi avec *C*, exprimant que c'était « la même chose ici ». Pourtant, Andrew n'a rien dit de ce qu'il faisait, puisqu'il a seulement répondu, dans le mauvais ordre, à la question précédente du participant sur ce dont il aimerait parler en *A*.

Une autre interprétation pourrait être qu'ils ont compris qu'il répondait à leur question en *A*, et ont donc répondu à leur tour qu'ils ne savaient pas non plus de quoi parler. Pourtant, ils ont immédiatement posé des questions sur les loisirs d'Andrew, ignorant sa question sur ce qu'ils faisaient.

Enfin, une autre interprétation encore pourrait être qu'ils ont interprété qu'Andrew voulait dire « Nous pourrions parler de ce que vous faites, mais je ne suis pas vraiment sûr », ce qui expliquerait également leur réponse.

Ainsi, dans les trois interprétations, il y a eu un malentendu, car le participant semble avoir, dans un cas, imaginé plus d'informations que ce qu'Andrew a réellement fourni, et dans les autres interprétations, avoir simplement ignoré la question d'Andrew.

Puisque toutes les informations étaient à la disposition du participant dans de telles inversions, il est possible qu'il ait pu former son interprétation de ce qu'Andrew voulait dire assez facilement, même si cette interprétation ne correspondait pas nécessairement à ce que l'acteur *IA* voulait vraiment communiquer. Cela pourrait expliquer en partie pourquoi l'effet n'a pas été aussi fort chez les femmes. Leurs plus grandes compétences linguistiques (Eriksson et al., 2012) auraient pu leur permettre de générer une interprétation de la phrase plus facilement et avec moins de coûts de traitement que les hommes, qui auraient pu rester un peu bloqués sur de tels messages.

Modèle Linéaire

On pourrait initialement mettre en doute la qualité du modèle linéaire pour représenter le temps de réponse typique des participants, en raison de son R carré ($R^2 = 0,4$). Pourtant, il est important de se rappeler qu'à aucun moment nous n'avons eu l'intention de faire un modèle parfait de ce processus. Le seul objectif de ce modèle était de supprimer l'effet de la longueur des phrases dans notre TR tout en conservant la variabilité des autres facteurs. Nous sommes bien conscients qu'un certain nombre de facteurs différents, non liés à notre protocole expérimental, pourraient jouer un rôle dans l'augmentation des temps de réponse, y compris, mais sans s'y limiter, la récupération en mémoire, ou le temps nécessaire pour traiter le contenu sémantique du message, et d'autres processus qui ne peuvent pas être facilement prédits. Tout bien considéré, notre modèle s'est avéré être un outil important à utiliser, car la longueur des phrases représentait à elle seule environ 40% de la variabilité des données. Il est également intéressant de noter qu'il n'a pas contribué aux différences entre les genres que nous avons observées. En effet, les prédictions données par le modèle pour les messages sans violation dans les conversations avec l'acteur IA n'étaient pas significativement différentes entre les participants hommes et femmes.

Travaux futurs

Dans cet article, nous avons seulement testé notre hypothèse sur les trois maximes qui semblaient avoir des effets d'intensités différentes sur le sentiment d'humanité véhiculé lorsqu'elles étaient violées. D'autres maximes pourraient être étudiées, puisque certaines d'entre elles contribuent également à donner un sentiment artificiel à l'interlocuteur lorsqu'elles sont violées : il s'agit de la deuxième maxime de quantité²². La deuxième maxime de manière²³ pourrait également être une bonne candidate pour des études futures car sa

22. « Ne rends pas ta contribution plus informative que nécessaire »

23. « Évite l'ambiguïté ».

violation aurait probablement un effet important sur les temps de réponse, mais contraire à la deuxième *maxime* de la quantité dont les violations ont un effet positif important sur l'humanité de l'interlocuteur.

Nous pourrions également imaginer de tester d'autres attentes conversationnelles qui ne sont pas directement gricéennes, mais qui partagent des concepts similaires de violations, comme les attentes liées à la politesse conversationnelle (Culpeper & Terkourafi, 2017) dont il a été montré qu'elles avaient un effet dans l'interaction homme-machine (Masson, Baratgin & Jamet, 2017b).

Une autre technique, parallèlement à l'enregistrement des temps de réponse, pourrait consister à utiliser des eye-trackers pour étudier les fixations des participants pour chaque énoncé (Groen & Noyes, 2013, pour un exemple de cette technique). Il est probable que les participants passent plus de temps à fixer des énoncés surprenants et obscurs. Le principal inconvénient de cette technique est le fait qu'elle nécessite la présence physique de l'équipement et du participant, ce qui pourrait rendre le cadre général beaucoup moins écologique par rapport à notre protocole.

En ce qui concerne les différences potentielles observées en fonction du genre des participants, nous pensons qu'une autre expérience impliquant des personnages fictifs de différents sexes pourrait être pertinente afin de générer des conditions avec des conversations de même genre et d'autres avec des conversations mixtes, car Mulac (1989) a montré que les personnes de différents genres ne se comportent pas nécessairement de la même manière selon le genre de leurs partenaires de conversation. Dans notre cas, il s'agissait toujours d'un personnage masculin.

5.5 Conclusion

Notre expérience semble indiquer que l'utilisation du *TT* avec des enregistrements des *TR* des participants est un outil pertinent pour étudier les conversations en ligne.

En effet, l'utilisation des **TR** donne une granularité beaucoup plus fine aux données collectées, les faisant passer du niveau de la conversation (avec le **TT**) au niveau des messages individuels (avec les **TR**). Elle contribue également à l'étude de l'idée selon laquelle les **TR** dans une conversation pourraient effectivement être corrélés dans une certaine mesure au coût cognitif du traitement d'un message et de la génération d'une réponse pertinente.

Nous pouvons également constater que les violations augmentant significativement les **TR** des participants étaient généralement associés à une augmentation de l'effet machine dans l'identification du **TT**, la seule exception étant les violations de la première **maxime** de quantité pour les hommes, ce qui ajoute une indication supplémentaire que les **TR** sont un bon outil pour mesurer les déviations par rapport à l'interaction humaine typique.

Contrairement à d'autres méthodes d'évaluation, l'aspect écologique de notre protocole le distingue d'autres, car les participants sont plutôt habitués à discuter dans des conversations textuelles avec une personne qu'ils ne voient pas réellement. De plus, ils ont la possibilité d'interagir directement avec l'agent qu'ils doivent juger, au lieu de juger des extraits de conversations préalablement enregistrées, offrant ainsi plus de variations de stratégies et contribuant à l'aspect écologique du contexte.

Nous pensons que ce protocole expérimental constitue une nouvelle façon d'évaluer les agents conversationnels dans les conversations en ligne, que ce soit par rapport aux humains ou par rapport à d'autres agents artificiels. Un autre avantage de cette méthode est que les **TR** pourrait potentiellement être analysés pendant que la conversation se déroule, sans avoir à attendre que des échantillons de conversation soient analysés par plusieurs juges humains à une date ultérieure.

Chapitre 6

L'impact des maximes de Grice de qualité, quantité et de manière dans les chatbots

Points remarquables dans ce chapitre

1. Nous répliquons l'expérience du chapitre précédent pour tester l'influence d'autres maximes de Grice.
2. Les maximes évaluées sont les maximes de Qualité (fausse information), de Quantité (trop d'information) et de Manière (ambiguïté).
3. Nous montrons que l'ordre des conversations dans le test de Turing peut avoir un effet sur l'influence de la violation de certaines maximes, et nous proposons quelques explications à ces différences.

6.1 Introduction

Depuis que Facebook et Microsoft ont lancé des outils permettant d'intégrer des [chatbots](#) dans leurs plateformes de messagerie en 2016, les agents conversationnels artificiels se sont répandus comme une traînée de poudre au sein des entreprises et des

services en général (Ask et al., 2016 ; Brandtzaeg & Følstad, 2017 ; Lasek & Jessa, 2013). Les **chatbots** sont des agents conversationnels artificiels capables d'interagir avec leurs utilisateurs sous la forme de conversations naturelles, généralement par le biais de textes, en utilisant des outils tels que la compréhension et la génération du langage naturel.

Malgré les réponses de plus en plus crédibles que les **chatbots** récents peuvent produire, la plupart d'entre eux rencontrent toujours les mêmes écueils que ceux de la première génération : ne pas réussir à déduire l'intention de l'utilisateur à partir du contexte de son énoncé. Pour pouvoir continuer à participer à des conversations avec des humains, certains **chatbots** utilisent des astuces qui ne nécessitent pas de déduire les attentes de l'utilisateur pour générer une réponse.

Certaines de ces astuces sont des violations flagrantes des **maximes** de Grice (1975). Elles correspondent à des attentes pragmatiques, critiques, que les personnes ont sur le contenu et la forme qu'un énoncé devrait avoir dans une conversation avec un autre humain. En voici quelques exemples : 1) les énoncés contenant trop d'informations (une violation de la deuxième **maxime** de quantité de Grice) ; 2) les énoncés contenant trop peu d'informations (une violation de la première **maxime** de quantité de Grice) ; 3) le recours à l'humour au lieu de répondre (ce qui peut violer de nombreuses **maximes**, dont la première **maxime** de qualité) ; 4) le fait de rester ambigu (une violation de la deuxième **maxime** de manière de Grice) ; 5) le changement de sujet pour un sujet plus générique (une violation de la **maxime** de relation de Grice) ; et cette liste n'est pas exhaustive.

L'effet de ces violations sur la perception de l'état d'humanité des **chatbots** a été étudié par Saygin et Cicekli (2002) sur des experts de la conversation avec des **chatbots**. Les temps de réponse des participants à une conversation interactive ont été étudiés pour trois des **maximes** de Grice (la première **maxime** de quantité, la **maxime** de relation, et la quatrième **maxime** de manière) dans le chapitre précédent (également décrits dans Jacquet et al., 2018, 2019), au sein de conversations textuelles interactives. Ces articles ont montré que les violations de la **maxime** de relation avaient un effet négatif particulièrement fort sur l'humanité d'un partenaire de conversation dans un **TT** (Turing, 1950), et

augmentaient significativement les temps de réponse des participants.

Ce chapitre complète l'étude présentée dans le chapitre précédent (Jacquet, Baratgin & Jamet, 2019). Il examine d'une part, l'influence des violations de trois autres *maximes* (la *maxime* de qualité, la *maxime* de quantité, et la *maxime* de manière) sur la perception de l'état d'humanité d'un partenaire conversationnel et d'autre part, sur les temps de réponse des participants à un *TT*. De la même manière que dans cette étude, nous n'utilisons pas d'agent artificiel réel, mais plutôt un protocole type *magicien d'Oz* (prétendre avoir un agent artificiel autonome alors qu'il s'agit en fait d'un humain qui le contrôle). Nous avons émis l'hypothèse que les violations de la *maxime* de qualité auraient : 1) un effet le plus important sur la perception de l'état d'humanité et sur les temps de réponse, 2) suivies par les violations de la *maxime* de quantité, et enfin 3) par les violations de la *maxime* de manière.

6.2 Revue de la littérature

Principe de coopération

Les *maximes* gricéennes (Grice, 1975) ont été largement étudiées dans les domaines de la linguistique et de la pragmatique (voir par exemple (Engelhardt et al., 2006), (Fukumura & van Gompel, 2017) et (Donovan-Kicken et al., 2013)). Nous discuterons certaines de ces études ci-dessous). Les *maximes* représentent des attentes générales que les humains produisent au cours d'une conversation avec un autre humain. Grice suggère que les personnes supposent automatiquement que leur interlocuteur va coopérer avec eux pour transmettre du sens à dans leurs énoncés. Ce principe est ce qu'il a appelé le principe de coopération :

Que votre contribution conversationnelle corresponde à ce qui est exigé de vous, au stade atteint par celle-ci, par le but ou la direction acceptés de

l'échange parlé dans lequel vous êtes engagé¹. (p. 61 Grice, 1975, traduit par Frédéric Berthet et Michel Bozon).

Grice propose ensuite, en supposant que le principe de coopération soit respecté et que les participants à une conversation coopèrent effectivement, que les interlocuteurs auront des attentes réciproques sur les productions au cours de la conversation. Ces attentes, ou *maximes*, se répartissent en quatre catégories : Qualité, Quantité, Relation, et Manière. Les trois premières décrivent les attentes concernant les informations qu'un énoncé est censé contenir, et la quatrième décrit comment ces informations doivent être transmises.

Les deux *maximes* de qualité font référence à la validité des informations contenues dans l'énoncé. La première *maxime* « Ne dites pas ce que vous croyez être faux », est également appelée *maxime* de véracité. Elle suggère que les personnes ne s'attendent pas à ce que leurs interlocuteurs leur donnent des informations qu'ils savent être fausses au sein du principe de coopération. En d'autres termes, si le locuteur dit quelque chose, on suppose qu'il croit en ce qu'il dit ; car sinon il ne l'aurait pas dit. La deuxième *maxime* « Ne dites pas quelque chose dont vous n'avez pas de preuve » suggère que les personnes s'attendent à ce qu'il y ait des raisons à ce que nous disons.

Les deux *maximes* de la quantité font référence à la quantité d'informations contenues dans un énoncé. Il ne doit y avoir ni trop peu d'informations (première *maxime* : « Faites en sorte que votre contribution soit aussi informative que nécessaire ».) ni trop d'informations (deuxième *maxime* : « Faites en sorte que votre contribution ne soit pas plus informative que nécessaire. »). Par exemple, on pourrait dire « Le grand arbre » lorsqu'il n'y a qu'un seul arbre (violation de la deuxième maxime) ou « L'arbre » lorsqu'il y a plus d'un arbre (violation de la première maxime).

La *maxime* de la relation « Soyez pertinent » fait référence à la façon dont l'information contenue dans un énoncé est liée au reste de la conversation. Cela peut être

1. "Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged".

en relation avec les énoncés précédents, mais aussi avec le but de l'échange. Les changements inattendus de sujets de conversation sont un exemple de violation potentielle de cette *maxime*, bien que Grice lui-même ait admis que les détails de cette maxime n'étaient pas clairs pour lui, notamment dans les cas de la pertinence des changements de sujet de conversation (Grice, 1975, p. 61).

Enfin, les quatre *maximes* de la Manière sont des directives pour produire des énoncés efficaces en suivant le principe de coopération. Les énoncés ne doivent pas être obscurs (première *maxime* : « Évitez les propos obscurs ») comme ils ne doivent pas être ambigus (deuxième *maxime* : « Évitez l'ambiguïté »). Ils doivent être également brefs (troisième *maxime* : « Soyez bref (évitez la prolixité inutile) »). Les informations contenues dans les énoncés doivent être disposées dans un ordre logique (quatrième *maxime* : « Soyez ordonné »).

Il est important de souligner que toutes ces *maximes* dans les quatre catégories ne sont pas véritablement des règles que les personnes suivent. Elles constituent des attentes générales implicites. Il existe des exemples connus d'énoncés dans lesquels une ou plusieurs de ces *maximes* sont violées (ignorées à l'insu de l'auditeur) ou au moins bafouées (explicitement ignorées à la connaissance de l'auditeur). Nous pouvons mentionner les ironies verbales (Alba Juez, 1995) comme « La main de son ami était aussi douce qu'une pierre » ou les métaphores (Keysar & Glucksberg, 1992) comme « Il est un requin » qui sont des cas de violation des *maximes* de qualité. Les changements explicites de sujets pour exprimer un refus de discuter plus avant du sujet actuel (Donovan-Kicken et al., 2013) peuvent violer la *maxime* de relation. Par exemple, dans un échange entre A et B : A) « As-tu fait tes devoirs ? » B) « Marc vient ce soir ! ». L'humour peut également violer de nombreuses *maximes* (Attardo, 1990) et en particulier la *maxime* de qualité.

Les situations ne sont pas toujours bien expliquées par le principe de coopération et ses *maximes*. La théorie de la pertinence fournit des prédictions plus précises dans de nombreux cas où les *maximes* semblent échouer² (Wilson & Sperber, 2002).

2. Les autres théories peuvent également expliquer certains cas, comme la théorie de la simplicité

Un aspect intéressant de la théorie de la pertinence est sa définition de ce que signifie une production pertinente. La pertinence d'un énoncé diminue avec le coût (ou l'effort) nécessaire pour traiter sa signification et augmente avec l'effet cognitif (quantité de changement dans les représentations mentales de l'auditeur) produit par le message qu'il véhicule dans un contexte donné. Quelques exemples de *maximes* gricéennes interprétées à la lumière de la théorie de la pertinence sont les deux *maximes* de quantité : fournir des informations inutiles augmente le coût cognitif sans augmenter l'effet cognitif, par conséquent il n'est pas aussi pertinent qu'un énoncé contenant juste la bonne quantité d'informations (similaire à la deuxième *maxime* de quantité). Produire une phrase avec trop peu d'informations est également moins pertinent. Dans ce cas, c'est l'effet cognitif qui est diminué. De même, un énoncé ambigu (une violation de la deuxième *maxime* de manière) pourrait augmenter le coût cognitif tout en diminuant l'effet cognitif. Dans le cas d'une fausse déclaration, l'effet cognitif pourrait également être plus faible, tandis que le coût cognitif pourrait augmenter car l'auditeur pourrait se demander ce qu'implique le fait de mentir dans un contexte donné.

Evaluer les chatbots

On peut se demander s'il est véritablement nécessaire d'aller aussi loin dans le fonctionnement de la communication humaine pour créer des *chatbots* utiles. Pourtant, malgré les progrès réalisés récemment dans l'humanisation des *chatbots*, les utilisateurs restent frustrés par leurs faibles capacités linguistiques et leur incapacité à comprendre l'intention de leur interlocuteur : (Jacquet, Masson et al., 2019 ; Jain et al., 2018). Bien que la capacité à produire des phrases syntaxiquement correctes se soit considérablement améliorée, ils rencontrent toujours des problèmes pour produire des phrases significatives et pertinentes dans leur contexte (Coniam, 2014). Ces problèmes existent depuis les premiers *chatbots* : ELIZA (Weizenbaum, 1966) et survivent encore dans d'autres plus ré-

qui formule également des prédictions intéressantes pour le changement de sujet et la dérive (Dessalles, 2017). La *Asif*-Theorie propose une approche gricéenne corrigée pour expliquer l'ironie tout en intégrant des parties du compte échoïque de la théorie de la pertinence (Garmendia, 2015).

TABLE 6.1 – **Résumé des études.** Synthèse des études comparant l’influence des violation des maximes conversationnelles de Grice sur la perception de l’état humanité des agents conversationnels.

Violations	Saygin et Cicekli (2002)	Jacquet, Baratgin et Jamet (2019)	Étude actuelle (Attendu)
1st Qualité	-	-	Fort effet négatif
1nd Quantité	Effet positif	Aucun effet	-
2st Quantité	Effet négatif	-	Effet négatif intermédiaire
Relation	Fort effet négatif	Fort effet négatif	-
2th Manière	Effet positif?	-	Faible effet
4th Manière	Effet positif?	Effet négatif moyen	-

cents malgré la très large diffusion des agents conversationnels artificiels avec A.L.I.C.E. et le langage AIML (Wallace, 2009). Ces problèmes subsistent avec des [chatbots](#) plus récents comme Zo de Microsoft.

L’évaluation des agents conversationnels artificiels reste un processus compliqué. La raison est la variété des situations auxquelles ils doivent faire face ainsi que le fait qu’il est très facile pour eux de se tromper sur les attentes de l’utilisateur, et donc de ne pas y répondre convenablement. Certains processus d’évaluation d’intérêt existent (voir (Radziwill & Benton, 2017), pour un exemple). Cependant, ils se limitent généralement à fournir des données pour l’ensemble de la conversation plutôt que pour des réponses spécifiques de l’agent.

Les utilisateurs étant habitués à interagir avec d’autres humains plus qu’avec des [chatbots](#), il n’est pas surprenant qu’ils comparent facilement leurs compétences linguistiques à ce qu’ils attendent d’un être humain, à moins qu’on ne leur dise spécifiquement de se préparer à autre chose. Il s’agit d’un élément important à prendre en compte. Plus un agent conversationnel est proche du comportement d’un humain, plus tout écart par rapport à ce comportement humain affectera l’utilisateur (Ciechanowski et al., 2019). Ce principe est également connu sous le nom de « Uncanny Valley », ou vallée de l’étrange,

de l'inconfort.

Il existe déjà un outil permettant d'évaluer dans quelle mesure les productions d'un *chatbot* sont proches de ce que l'on attend d'un humain : le *TT* (Turing, 1950). Bien qu'il ait été initialement conçu par son créateur pour tester l'intelligence d'une machine (en laissant un juge comparer deux conversations, l'une avec un humain, l'autre avec une machine), il serait peut-être plus approprié de le considérer comme un test « d'humanité ». En effet, ce test a été fortement critiqué dans le domaine de l'informatique car le *chatbot*, pour paraître plus humain, devait parfois prétendre avoir plus de difficultés à réaliser certaines tâches, qu'une calculatrice aurait pu réaliser très rapidement (voir French, 2000 ; Saygin et al., 2000, pour des revues sur son histoire).

L'humanité d'un *chatbot* semble être liée dans une certaine mesure à son respect des *maximes* gricéennes (Saygin & Cicekli, 2002). La violation de la *maxime* de relation semble en effet avoir l'effet le plus négatif sur l'humanité de l'agent conversationnel artificiel par rapport aux autres *maximes*. Ceci a été confirmé dans (Jacquet, Baratgin & Jamet, 2019) et (Jacquet et al., 2018). Les violations des *maximes* de manière, en revanche, ont donné au *chatbot* un plus fort sentiment d'humanité (Saygin & Cicekli, 2002), bien qu'il ne soit pas clair à partir de quelle sous-maxime ce sentiment ait pu être produit. Les violations de la quatrième *maxime* de manière (« être ordonné ») n'ont pas produit un effet aussi positif sur l'humanité d'un partenaire conversationnel dans Jacquet, Baratgin et Jamet (2019), et ont plutôt semblé produire un petit effet négatif. Les violations de la *maxime* de quantité produisent des effets opposés selon que le *chatbot* donne trop peu ou trop d'informations. Dans le premier cas, les violations de la *maxime* ont tendance à produire un effet positif sur la perception (Saygin & Cicekli, 2002) de l'humanité, probablement parce que le *chatbot* semble alors s'ennuyer ou ne pas être intéressé. Cet état émotionnel inféré aide à catégoriser le *chatbot* comme étant plus humain que les humains réels. Ceci est cohérent avec les résultats trouvés dans les conversations interactives (comme montré dans le chapitre précédent, Jacquet et al., 2018, 2019). L'inverse semble être vrai pour la sur-description, car cela donne à l'énoncé du *chatbot* un senti-

ment encyclopédique, qui a été reconnu comme artificiel par leur partenaire (Saygin & Cicekli, 2002). On remarque que la sur-description n'est pas nécessairement rare dans les conversations humaines. En particulier dans celles orientées vers un but, comme nous l'avons mentionné ci-dessus. Enfin, les résultats de Saygin et Cicekli (2002) n'étaient pas concluants en ce qui concerne l'influence des violations de la *maxime* de qualité. En effet, ces violations allaient généralement de pair avec la violation de la *maxime* de quantité.

Temps de Réponse

Il est souvent difficile d'étudier la capacité pragmatique des *chatbots* dans les conversations. Il est possible d'utiliser des extraits des conversations précédentes et de les faire vérifier par des juges (Saygin & Cicekli, 2002). Il s'agit d'une méthode *post-hoc* et longue. Elle nécessite à la fois des interlocuteurs et des juges. De plus, elle ne fournit des données quantifiables qu'au niveau de conversations entières. Ce qui peut ne pas être idéal pour des violations très localisées. Une autre méthode consiste à enregistrer le délai entre l'énoncé du *chatbot* et la réponse de l'utilisateur (Jacquet, Baratgin & Jamet, 2019), (Jacquet et al., 2018). Si cette nouvelle méthode s'avère suffisamment fiable, elle pourrait devenir utile dans les *chatbots* utilisant des algorithmes d'apprentissage continu. L'évaluation de la différence entre un temps de réponse prédit et un temps de réponse observé pourrait contribuer à détecter des comportements inattendus.

Un autre aspect intéressant de cette méthode est qu'il n'est pas nécessaire d'avoir un *chatbot* fonctionnel pour qu'elle soit utilisable. En effet, tant que les participants croient qu'ils ont à faire à un *chatbot*, un humain peut jouer ce rôle, dans un paradigme de type *magicien d'Oz*. Cette procédure est utile pour éviter les différences d'utilisation du lexique. Des études antérieures ont montré que le vocabulaire utilisé dans les conversations avec un *chatbot* est, en moyenne, plus pauvre que le lexique utilisé dans les conversations entre humains (Hill et al., 2015). L'utilisation de la technique du *magicien d'Oz* a déjà été utilisée dans le contexte des *chatbots* (Jacquet et al., 2018, 2019; Medhi Thies et al., 2017); et avec les robots sociaux (Jamet et al., 2018; Masson, Baratgin & Jamet, 2017a;

Masson et al., 2015, 2017b ; Masson, Baratgin, Jamet et al., 2017).

Les temps de réponse sont souvent utilisés dans le domaine de la psychologie cognitive (voir (Leth-Steensen et al., 2000) pour un exemple, et (Kosinski, 2008) pour une revue). Si les temps de réponse simples ne sont pas toujours le moyen le plus précis d'étudier le fonctionnement interne de l'esprit, ils ont le grand avantage d'être faciles à mesurer et sont non invasifs pour le participant, contrairement à d'autres méthodes comme l'électroencéphalographie. Le lien entre les temps de réponse et la charge cognitive semble assez probable, y compris dans les recherches sur la pragmatique du langage (voir par exemple : (Nordmeyer & Frank, 2015), (Jacquet et al., 2018), et (Jacquet, Baratgin & Jamet, 2019)). En raison de la facilité d'enregistrement des temps de réponse, cette méthode permet de conserver un environnement assez écologique pour le participant. Elle a également le grand avantage de pouvoir être mesurée sans que le participant soit physiquement présent pour l'expérimentation, ce qui en fait une mesure potentiellement utilisable dans l'analyse de grandes quantités de données enregistrées en ligne.

Des différences entre les sexes ont également été remarquées dans des études antérieures portant sur des partenaires conversationnels artificiels et ce, tant avec des robots (Park et al., 2011 ; Powers et al., 2005 ; Tay et al., 2014) qu'avec des *chatbots* (Jacquet et al., 2018, 2019). Les raisons exactes de ces différences restent floues. Il n'est pas impossible que cela soit lié aux différences observées dans le rythme d'apprentissage des compétences linguistiques par les garçons et les filles. En effet, ces dernières sont généralement en avance sur les garçons (Eriksson et al., 2012). En outre, l'anthropomorphisation automatique des robots (en supposant qu'ils se comportent un peu comme des humains) suffit généralement pour que les personnes les considèrent comme des agents sociaux (Masson, Baratgin & Jamet, 2017a ; Masson, Baratgin & Jamet, 2017b). Dans cet article, nous avons choisi de nous focaliser sur trois *maximes* gricéennes dont l'influence sur les temps de réponse de l'utilisateur n'a pas encore été étudiée en (Jacquet, Baratgin & Jamet, 2019) : la première *maxime* de qualité, la deuxième *maxime* de quantité, et la deuxième *maxime* de manière.

6.3 Expérience

La méthodologie vise à compléter les travaux initiés dans (Jacquet, Baratgin & Jamet, 2019) et dans (Jacquet et al., 2018) sur la première *maxime* de quantité, la *maxime* de relation et sur la quatrième *maxime* de manière. Dans notre étude, nous nous intéresserons à la première *maxime* de qualité, à la deuxième *maxime* de quantité et à la deuxième *maxime* de manière.

Participants

Les participants sont tous des adultes de langue maternelle française (âge : $M = 22,5$, écart-type = 4,5), tous familiers des applications de messagerie en ligne telles que Facebook’s Messenger, Microsoft’s Skype, Discord ou Telegram. 32 étaient des femmes et 31 des hommes.

Variables

Nous avons enregistré le délai entre la réponse du participant et l’énoncé de l’agent conversationnel dans les conversations. Ce délai a été ensuite transformé à l’aide d’un modèle de régression linéaire calculé à partir de toutes les conversations entre les participants et l’agent conversationnel humain dans lesquelles il n’y avait pas de violation planifiée des *maximes* conversationnelles. Ce modèle nous a permis d’extraire la quantité de temps dans ce délai qui ne pouvait pas être expliquée simplement par la longueur des deux énoncés. Nous avons appelé ce résultat « delta d » : la différence entre le délai observé d et le délai prédit par le modèle D :

$$\Delta d = d - D \tag{6.1}$$

Le modèle linéaire lui-même utilisait le nombre de caractères contenu dans l'énoncé de l'expérimentateur C_e et le nombre de caractères contenu dans la réponse du participant C_p .

$$D = (w \times C_e) + (x \times C_p) + (y \times C_e C_p) + z \quad (6.2)$$

Nous avons également demandé aux participants de deviner lequel des deux agents conversationnels auxquels ils ont parlé était un **chatbot** (réponses traditionnelles au TT).

Facteurs

Nous avons considéré trois facteurs potentiels pour notre expérience. Le facteur principal de notre hypothèse était que le type de violation dans la conversation avec le *artificiel*³ agent (la première **maxime** de qualité, la deuxième **maxime** de quantité, la deuxième **maxime** de manière). Le deuxième facteur potentiel que nous avons pris en

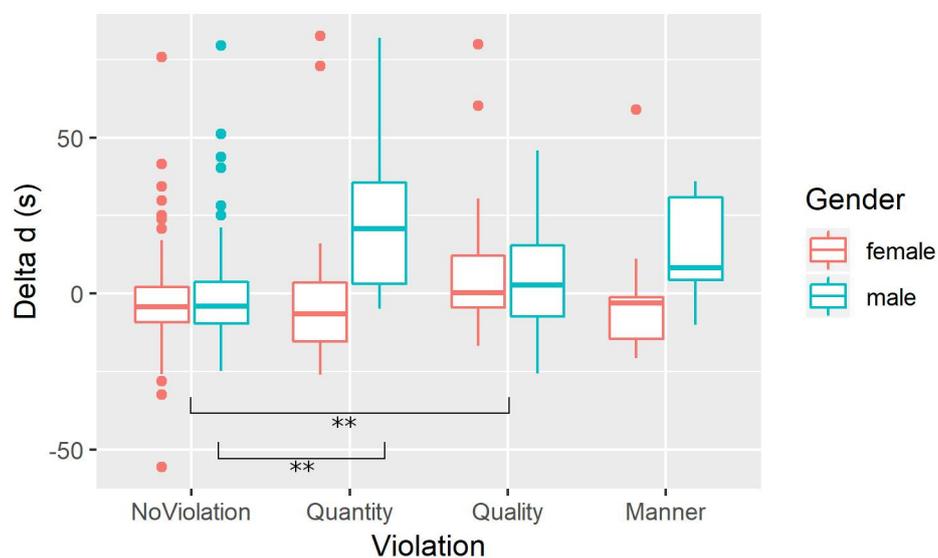


FIGURE 6.1 – Δd dans la première conversation. « No Violation » indique les cas où les messages des participants dans les conversations avec l'acteur IA suivaient des messages sans violations (contrôle).

** $p_{Holm} < 0,01$

3. Aucun **chatbot** n'était en fait présent. Le même expérimentateur s'adressait aux participants dans les deux conversations, mais dans la conversation *artificielle*, il devait introduire des violations volontaires des **maximes** gricéennes de qualité, de quantité et de manière, selon la condition.

compte était le sexe du participant. Enfin, nous avons considéré que l'ordre dans lequel les participants ont conversé avec les différents agents pouvait avoir une influence.

Procédure

Après avoir rempli un questionnaire en précisant leur âge, leur sexe, leurs connaissances auto-évaluées en informatique et en intelligence artificielle, ainsi que leur domaine d'étude/carrière, les participants ont été envoyés sur un site web où un modérateur leur a expliqué les règles qu'ils devraient respecter lors des conversations avec les deux agents. On leur a indiqué qu'ils participaient à un **TT**. Ils allaient parler à une intelligence artificielle et à un humain dans un ordre aléatoire. Leur tâche consistait à deviner, à la fin des deux conversations de 10 minutes, à quel agent ils avaient parlé : l'humain ou l'agent *artificiel*.

Nous avons suivi le protocole de type Wizard-of-Oz. L'agent *artificiel* n'était pas un véritable *chatbot*, mais un humain. En fait, l'agent *artificiel* et l'agent humain étaient joués par le même expérimentateur afin d'éviter les différences potentielles de vocabulaire comme de style d'écriture.

L'expérimentateur a été précisé que les deux agents ne parleraient pas d'eux-mêmes dans les conversations. Ils incarneraient en revanche un même personnage fictif appelé Noam⁴ aussi bien qu'ils le pouvaient. L'objectif poursuivi était d'inviter les participants à prêter attention à la façon dont chaque agent se comportait plutôt que de se concentrer sur le contenu sémantique qui leur a été donné. Par exemple, si un agent disait que Noam étudiait la psychologie et que l'autre disait que Noam travaillait sur des ordinateurs, les participants auraient pu juger de la crédibilité de l'agent sur la base de cette information plutôt que sur les énoncés en général. Par conséquent, toutes les informations relatives au personnage fictif ont été prédéfinies de manière à être aussi identiques que possible entre les deux conversations. La seule exception étant dans la condition de

4. La description du personnage de Noam est disponible en français dans (Jacquet, Hullin et al., 2019b), ainsi que les fichiers, analyses et données du serveur.

violation de la *maxime* de qualité où des informations alternatives (et peu crédibles) ont été choisies.

Les conversations étaient limitées uniquement à du texte. Chaque message ne pouvait contenir que 255 caractères maximum. L'utilisation de smileys n'était pas autorisée. L'expérimentateur et le participant devaient tous deux respecter les principes d'une conversation (le propos de l'un est suivi d'un autre propos). Il postait un message puis attendait la réponse. Bien que la plupart de ces limitations ne soient pas nécessaires, elles ont contribué à donner aux participants l'impression qu'ils allaient effectivement parler à une intelligence artificielle (en particulier l'interdiction d'utiliser des smileys).

Si l'agent de la première conversation faisait une faute de frappe, l'agent de la seconde conversation devait également en introduire une pour éviter que la présence ou l'absence de fautes de frappe ne fasse partie des indices permettant de différencier les deux agents : (Westerman et al., 2018).

À la fin des deux conversations, les participants devaient deviner à quel agent (humain ou intelligence artificielle) avait-il eu à faire lors de la première conversation. Ils indiquaient ensuite dans quelle mesure ils étaient confiants dans leur estimation sur une échelle de Likert de 1 à 7.

Conditions

Nos principales conditions étaient le type de violations que l'agent *artificiel* insérerait volontairement dans ses productions.

Dans la condition où il produirait des violations de la première *maxime* de qualité (comme dire qu'il est un chasseur de baleines au lieu de dire qu'il est un instructeur de plongée), nous nous attendions à ce que le participant soit quelque peu troublé, et que cette réponse soit inattendue. Par conséquent, nous avons prédit que cette condition verrait les temps de réponse des participants augmenter par rapport aux prédictions du

modèle linéaire.

Dans la condition où il produirait des violations de la deuxième *maxime* de quantité (comme dire : le travail de la femme de Noam en plus du sien lorsqu'on lui pose la question : Quel est son travail?), nous nous attendions à ce que le participant ait besoin de plus de temps pour traiter cette phrase en raison de la présence d'informations inattendues. Par conséquent, nous avons prédit que cette condition verrait également les temps de réponse des participants augmenter.

Dans la condition où l'agent *artificiel* produisait des violations de la deuxième *maxime* de manière (comme être ambigu sur le travail de Noam), nous nous attendions à ce que le participant soit moins troublé que dans les autres conditions, et peut-être même qu'il le considère comme plus humain. Néanmoins, nous nous attendions à ce que les temps de réponse augmentent quelque peu en raison du coût cognitif supplémentaire lié au traitement des phrases ambiguës.

En général, nous avons également prévu que les temps de réponse diffèrent entre les participants et les participantes. Les participants auraient des temps inférieurs à ceux des participantes.

Résultats

Un total de 1097 énoncés ont été recueillis au cours de l'expérience : 546 dans les conversations avec l'agent humain et 551 dans les conversations avec l'agent *artificiel*. Parmi les énoncés enregistrés avec l'agent *artificiel*, 272 provenaient de participants lui parlant lors de la première conversation.

Lorsque l'agent *artificiel* était dans la première conversation, 55 violations de la première *maxime* de qualité ont été produites (23 dans les conversations avec des participantes, 32 dans les conversations avec des participants), 37 énoncés contenaient des violations de la deuxième *maxime* de quantité (21 avec des participantes, 16 avec des

TABLE 6.2 – Δd dans la première conversation. Résumé de l'influence des violations en fonction du sexe Δd en seconde

	<i>Moyenne</i>	<i>sd</i>	<i>Somme</i>	<i>se</i>	<i>t</i>	<i>p_{Holm}</i>	
Général							
Aucune violation	-1,0	16	163	1,3			
Qualité	6,7	20	55	2,7	-2,88	0,01	**
Quantité	11	29	37	4,9	-2,5	0,03	*
Manière	6,1	22	17	5,3	-1,65	0,10	.
Femmes							
Aucune violation	-1,9	15	98	1,5			
Qualité	8,7	22	23	4,7	-2,16	0,12	
Quantité	1,2	28	21	6,1	-0,50	-	
Manière	-0,12	23	10	7,3	-0,24	-	
Hommes							
Aucune violation	0,32	18	65	2,2			
Qualité	5,3	18	32	3,1	-1,28	0,20	
Quantité	25	27	16	6,8	-3,50	0,007	**
Manière	15	18	7	6,7	-2,05	0,09	.

. $p_{Holm} < 0,1$; * $p_{Holm} < 0,05$; ** $p_{Holm} < 0,01$

participants), et 17 énoncés contenaient des violations de la deuxième **maxime** de manière (10 avec des participantes, 7 avec des participants).

Les énoncés des conversations avec l'agent *artificial* ont été analysés à l'aide d'une ANOVA de type III (pour les plans non équilibrés). Les comparaisons par paires ont été testées à l'aide d'un test t de Welch ou d'un test t selon que les variances des deux échantillons qui étaient significativement différentes ou non. Nous avons également corrigé toutes les valeurs p de ces tests post-hoc pour tenir compte de la possibilité d'erreurs de type I (rejet trop facile des hypothèses nulles), en utilisant la correction de Holm-Bonferroni. Par conséquent, toutes les p-values corrigées seront notées p_{Holm} .

Les fréquences d'identification correcte de l'agent *artificial* dans le **TT** ont été comparées à 50% avec des tests de Chi-carré pour tester la déviation par rapport aux suppositions aléatoires. Les valeurs supérieures à 50% indiquent que l'agent *artificial* a été facilement détecté (et n'est donc pas de type humain). Les valeurs inférieures à 50% indiquent que l'agent artificiel n'était pas facile à détecter (et donc plus semblable à un

humain). les tests de Krustal-Wallis ont été utilisé pour tester les variables de contrôle entre nos conditions.

Modèle Linéaire

Les 546 énoncés des conversations avec l'agent humain ont été utilisés pour calculer le modèle linéaire. Le résultat est le suivant :

$$D = (0,20 \times C_e) + (0,42 \times C_p) - (0,0009 \times C_e C_p) + 5,3 \quad (6.3)$$

Ce modèle explique environ la moitié de la variance des données ($R^2_{\text{ajusté}} = 0,48$).

Temps de réponse

Nous observons une tendance à une double interaction entre le type de violation, l'ordre des conversations et le genre du participant sur Δd ($F(3,535) = 2,395; p = 0,067$). Cela signifie qu'il existe peut être un effet de l'ordre des conversations. Nous avons poursuivi les analyses en utilisant uniquement les énoncés provenant des participants ayant interagi avec l'agent *artificiel* lors de la première conversation.

Lorsque les participants ont conversé avec l'agent *artificiel* lors de la première conversation, nous observons une interaction significative du type de violation et du genre sur Δd ($F(3,264) = 4,465; p < 0,01$). Cette interaction est illustrée dans la Figure 6.1 et dans le Tableau 6.2.

En comparant les hommes aux femmes dans la première conversation, seules les énoncés suivant les violations de la deuxième *maxime* de quantité avaient un Δd significativement plus élevé que ceux ne suivant aucune violation ($t(35) = -2,62; p_{Holm} = 0,05$). Dans cette condition, le Δd des participants masculins était significativement plus élevé après une violation de la deuxième *maxime* de quantité ($t(18) = -3,50; p_{Holm} < 0,01$)

TABLE 6.3 – **Turing Test**. Les pourcentages représentent la fréquence des identifications correctes de l'*artificiel*.

	Première Conversation		Seconde Conversation		Les deux
Général	72%	**	48%		60%
Qualité	90%	**	27%		57%
Quantité	58%		60%		59%
Manière	70%		60%		65%
Femmes	71%	.	40%		56%
Qualité	80%		0%	*	40%
Quantité	71%	.	80%		75%
Manière	60%		40%		50%
Hommes	73%	.	56%		65%
Qualité	100%	*	50%		73%
Quantité	40%		40%		40%
Manière	80%		80%		80%

. $p < 0.1$; * $p < 0.05$; ** $p < 0.01$

alors qu'il ne l'était pas pour les participants féminins.

Nous observons un effet général des violations de la première *maxime* de qualité sur Δd , qui était significativement plus élevé ($t(216) = -2,88; p_{Holm} = 0,05$).

Nous observons également une petite tendance à ce que Δd soit plus élevé après des violations de la deuxième *maxime* de manière ($t(178) = -1,65; p_{Holm} = 0,10$).

Test de Turing

Les résultats du **TT** ont été analysés à l'aide d'un test du chi carré par rapport à une probabilité de 50% d'identifier correctement l'agent *artificiel*. Les résultats sont présentés dans le tableau 6.3. Il convient de noter que le nombre de personnes dans chaque condition peut ne pas être suffisant pour donner des généralisations réelles concernant les résultats des tests du chi carré, en particulier pour ceux inférieurs à 10 personnes.

Les participants ont identifié correctement l'agent *artificiel* plus souvent dans la première conversation (72%, $\chi^2(1, N = 30) = 6,12; p = 0,01$) que dans la deuxième conversation, pour laquelle il n'y avait pas de différence par rapport au hasard (48%, $\chi^2(1, N =$

31) = 0,03; $p = 0,86$).

Le pourcentage d'identification correcte était significativement plus élevé que 50% en général dans la première conversation pour la condition produisant des violations de la *maxime* de qualité (90%, $\chi^2(1, N = 10) = 6,4; p = 0,01$), ce qui était principalement le fait des hommes (100%, $\chi^2(1, N = 5) = 5,0; p < 0,05$). Les participantes se sont en fait trompées de manière significative dans cette condition lorsqu'elles ont parlé à l'agent *artificiel* lors de la deuxième conversation (0%, $\chi^2(1, N = 5) = 5,0; p < 0,05$).

Variables Contrôles

La durée des conversations ne variait significativement ni entre les conditions, ni selon le genre des participants, ni selon l'ordre dans lequel ils discutaient avec l'agent humain et l'agent *artificiel*. En effet, les conversations ont presque toujours duré la totalité des dix minutes autorisées.

L'âge n'était pas non plus significativement différent d'une condition à l'autre ($\chi^2(2) = 0,77; p = 0,68$) ni entre les sexes ($\chi^2(1) = 3,1; p = 0,08$).

Les connaissances auto-évaluées en informatique n'étaient pas significativement différentes, ni entre les conditions ($\chi^2(2) = 0,2; p = 0,89$) ni entre les genres ($\chi^2(1) = 2,8; p = 0,09$).

Il n'y avait pas de différence significative entre les conditions sur les connaissances en intelligence artificielle auto-évaluées par les participants ($\chi^2(2) = 3,5; p = 0,17$) mais il y avait une tendance entre les hommes et les femmes ($\chi^2(1) = 4,0; p = 0,06$).

6.4 Discussion

Quantité

Comme l'ordre de la conversation avait tendance à avoir une influence sur les temps de réponse, nous avons décidé de n'analyser que la première conversation. Il est en effet fort probable qu'il y ait eu un effet d'apprentissage dans la condition de violation de la deuxième *maxime* de quantité. En effet, lorsque les participants masculins ont parlé à l'agent *artificiel* lors de la première conversation, nous avons observé une augmentation significative des temps de réponse par rapport à ce que l'on pouvait attendre ($M = 25s, \sigma = 27s, N = 16$). Cette augmentation disparaissait complètement lorsque les participants s'adressaient à l'agent *artificiel* dans la deuxième conversation, après avoir parlé à l'agent humain.

Puisque les deux agents ont donné les mêmes informations de base (avec seulement des informations supplémentaires pour l'agent *artificiel*), le fait de revoir ces informations de base lors de la deuxième conversation pourrait avoir réduit le coût cognitif de leur traitement. Lorsque cette deuxième conversation a eu lieu avec l'agent *artificiel*, certaines des informations données étaient attendues et déjà connues et, par conséquent, il ne restait que les informations supplémentaires à traiter. Ceci est cohérent avec les prédictions données par la théorie de la pertinence (Wilson & Sperber, 2002).

Les résultats indiquent qu'il existe une différence significative entre les participants de genre différent dans cette condition. Ceci est compatible avec le fait que les femmes ont des compétences linguistiques plus élevées que les hommes (Eriksson et al., 2012). Les informations supplémentaires fournies par l'agent *artificiel* pourraient simplement ne pas avoir ajouté autant de charge cognitive pour les femmes que pour les hommes.

Qualité

Les différences de genre mises à part, nous avons observé une augmentation générale des temps de réponse par rapport aux prédictions du modèle linéaire lorsque les énoncés des participants suivaient des violations de la **maxime** de qualité ($M = 6,7s, \sigma = 20s, N = 55$). Ceci est cohérent avec notre hypothèse concernant cette condition (un effet fort). En ce qui concerne le caractère humain de l'agent conversationnel dans le **TT**, cette condition était également la plus polarisée. En effet, c'est dans cette condition que les participants ont jugé l'agent le moins humain (90% d'identifications correctes de l'agent *artificiel*), indiquant que les mensonges évidents de cet agent étaient considérés comme indiquant à quel point il était artificiel. Pourtant, cet effet s'est complètement inversé lorsque les participants se sont adressés à l'agent *artificiel* dans la deuxième conversation (27% d'identifications correctes seulement), ce qui signifie que l'agent *artificiel* donnait l'impression d'être plus humain que l'humain réel.

Il n'est pas tout à fait clair pourquoi nous observons ce résultat, mais une hypothèse qui semble assez convaincante serait l'aspect humoristique des violations de la première **maxime** de qualité. En effet, comme les mensonges étaient assez évidents ou quelque peu ridicules (par exemple : « Je suis un chasseur de baleines. » ou « Les bigorneaux sont en fait ceux qui tuent le plus de baleines. Ils sont féroces avec leurs yeux rouge sang ! »), il est probable que les participants aient considéré ces réponses comme amusantes, surtout en comparaison avec les réponses habituellement plus sérieuses que les participants ont reçues lors de leur première conversation (avec l'agent humain). Puisque dans ce cas, ils disposaient déjà d'informations plus crédibles sur le personnage fictif, il est possible qu'ils aient plus facilement supposé qu'il s'agissait d'un humour authentique plutôt que d'une façon d'éviter de donner une véritable réponse. Lorsque l'agent *artificiel* se trouvait plutôt dans la première conversation, les participants ont pu considérer que l'agent avait été incohérent et quelque peu aléatoire dans ses réponses, par rapport aux réponses plus crédibles de l'agent humain qui semblait répondre sincèrement à leurs questions.

En ce qui concerne l'humanité de l'agent artificiel dans la condition produisant des violations de la première *maxime* de quantité, nous nous attendions à de meilleurs pourcentages d'identification correcte que ce que nous avons obtenu en réalité (83% pour les femmes, et 50% pour les hommes, ni significativement différent de 50%). Nous voyons deux raisons pour des résultats aussi faibles. Soit l'effet sur le caractère humain est en fait plus faible que prévu, et il nous faudrait un plus grand nombre de participants pour pouvoir le détecter (il n'y avait que 6 femmes et 4 hommes dans cette condition, dans la première conversation), soit les violations n'étaient pas assez machinales. En effet, si ces énoncés contenaient des informations supplémentaires qui ne répondaient pas directement à la question posée par les participants, ces ajouts n'étaient pas toujours totalement non pertinents ou non naturels (par exemple, répondre à la fois par le métier du personnage et le métier de sa femme lorsque les participants demandaient quel était leur métier). De telles violations de la première *maxime* de quantité ne sont en fait pas rares (Engelhardt et al., 2006), même si elles peuvent dérouter l'auditeur, et ce qui semble réellement artificiel dans de telles violations est davantage lié au sentiment encyclopédique (Saygin & Cicekli, 2002), qui n'était pas souvent présent dans notre étude.

Manière

Enfin, il n'y a pas eu d'effet réel des violations de la deuxième *maxime* de manière, ni sur les temps de réponse, ni sur le caractère humain de l'agent *artificiel* dans le *TT*. L'ambiguïté a cependant souvent été détectée par les participants. Par exemple, lorsqu'ils ont demandé le métier de Noam, l'agent a répondu « Je suis plongeur », ce qui peut signifier soit « Je suis un plongeur sous marin », soit « Je suis celui qui lave la vaisselle dans un restaurant ». Certains participants ont demandé une désambiguïstation lors de l'énoncé suivant : « Vous voulez dire à la mer ou dans un restaurant ? ». Il est possible que lorsque les participants n'ont pas demandé cette désambiguïstation, qu'ils aient simplement choisi la signification la plus probable avant de continuer la conversation sans même remarquer qu'il y avait une ambiguïté. Notez que nous n'avons pas trouvé d'effet positif des violations

de cette *maxime* sur l'humanité de l'agent artificiel, alors que parmi les *maximes* de manière, l'ambiguïté était l'une des plus susceptibles de produire un tel effet positif. Par conséquent, nous ne savons pas si l'effet positif observé dans (Saygin & Cicekli, 2002) était le résultat de violations de l'une des deux *maximes* de manière restantes, s'il était le résultat d'une interaction entre les *maximes* de manière en général, ou s'il provenait d'autre chose encore. Il est également possible qu'il s'agisse d'un effet trop faible pour être remarqué avec ce nombre de participants, d'autant plus qu'en moyenne, moins de violations ont pu être produites au cours des conversations dans cette condition que dans les deux autres conditions.

6.5 Conclusion

Dans cet article, nous avons élargi les connaissances sur l'interaction entre les violations des *maximes* gricéennes et les temps de réponse des partenaires de conversation dans des conversations textuelles interactives en ligne. Nos résultats étaient généralement cohérents avec nos attentes, bien qu'il y ait eu quelques surprises, en particulier concernant le manque d'effet de la deuxième *maxime* de quantité sur l'humanité de l'agent *artificiel*. Ce résultat spécifique peut potentiellement être expliqué par le fait qu'il n'y avait qu'un petit nombre de participants dans ces subdivisions de nos données. Nous avons également montré que l'ordre des conversations avait une influence sur les résultats, en particulier pour la deuxième *maxime* de quantité (aucun effet dans la deuxième conversation) et pour la *maxime* de qualité (effet négatif sur l'humanité dans la première conversation, effet potentiellement positif sur l'humanité dans la deuxième conversation). Nous avons également constaté ce qui avait déjà été observé dans la littérature, à savoir que les temps de réponse des participants masculins sont plus influencés par les violations des *maximes* que ceux des participantes féminines. Bien que la raison de ce phénomène ne soit pas claire, il est cohérent avec les conclusions selon lesquelles les femmes ont en moyenne de meilleures compétences linguistiques que les hommes (Eriksson et al., 2012).

Pour les travaux futurs sur ces *maximes*, nous recommandons d'étudier l'effet de violations plus encyclopédiques de la deuxième *maxime* de quantité, et concernant la *maxime* de qualité, peut-être d'en atténuer le contenu humoristique, qui pourrait contribuer à transformer des violations potentielles de la *maxime* en le fait de la bafouer. Puisque nous n'avons pas pu observer un effet positif des violations de la deuxième *maxime* de manière sur l'humanité de l'agent conversationnel, la source de l'effet positif des violations de la *maxime* de manière en général observé dans la littérature devrait être étudiée plus en profondeur car, jusqu'à présent, ni la deuxième *maxime* de manière (« Évitez l'ambiguïté ») ni la quatrième *maxime* de manière (« Soyez ordonné ») ne semblent être des causes crédibles d'un tel effet.

À l'issue de cette étude, nous pouvons affirmer que les *chatbots* doivent être soigneusement testés en ce qui concerne la qualité des informations qu'ils fournissent et la quantité d'informations qu'ils donnent dans leurs énoncés. Ils devraient également être capables de répondre aux demandes de désambiguïsation de leurs utilisateurs. En effet, même si l'ambiguïté n'a pas augmenté de manière significative les temps de réponse (trop peu de violations pour pouvoir le dire ?), les participants ont souvent demandé une désambiguïsation suite à de telles violations de la deuxième *maxime* de manière.

Troisième partie

Utiliser les chatbots et le Test de Turing en psychologie

Publications

Le contenu de cette partie a donné lieu à deux publications :

1. A Selfish Chatbot Still Does not Win in the Ultimatum Game, (Beunay et al., 2022)
2. Contextual Information Helps Understand Messages Written with Textisms, (Jacquet, Jaraud et al., 2021)

Chapitre 7

Un chatbot égoïste ne remporte pas le jeu de l'ultimatum

Points remarquables dans ce chapitre

1. Nous proposons un protocole expérimental utilisant des chatbots dans le jeu de l'ultimatum.
2. Nous observons l'effet de différents profils de personnalité de ces chatbots sur le comportement des participants.
3. Nous montrons que l'utilisation de chatbots mène à des comportements plus stéréotypés par rapport à ceux observés avec des expérimentateurs humains (moins de désirabilité sociale?).

7.1 Introduction

Le jeu de l'ultimatum (Güth et al., 1982) illustre la phase finale d'une négociation entre deux entités représentées par deux joueurs A et B. Le joueur A offre au joueur B une part des unités qui lui appartiennent (généralement 10 unités), et le joueur B

peut accepter ou refuser son offre. Si B accepte l'offre, les unités sont échangées entre les deux joueurs. S'il refuse au contraire, les deux joueurs perdent toutes les unités (les 10, quelle que soit la quantité qu'il était proposé de partager). Dans un deuxième tour de jeu, c'est le joueur B qui fait une offre au joueur A. Du point de vue de la rationalité économique, celui qui répond devrait toujours accepter l'offre même si elle est minimale (s'il ne reçoit qu'une unité). Dans les très nombreuses expériences faites sur le jeu de l'ultimatum, il est observé que les offres correspondant à un partage équitable (4, 5 unités) sont largement acceptées alors que les offres minimales (celles inférieures à 3 unités) sont très souvent rejetées (voir Tisserand, 2016, pour une revue). Les individus sont sensibles aux inégalités envers eux-mêmes et envers les autres dans leur prise de décision. Certains paramètres semblent atténuer ces résultats. Par exemple, si le proposant est fortement contraint de faire une offre inéquitable, alors le répondant sera également plus indulgent à son égard (Blount, 1995). Les attributions du répondant quant aux intentions du proposant modulent l'acceptation de l'offre. Il a également été montré que l'offre faite dans la deuxième étape par le joueur A, devenu un proposant, dépend de l'offre faite par le joueur B dans la première étape alors que d'après la rationalité économique, la seconde offre devrait être indépendante de la première. Il a été observé également une réciprocité des comportements. Ainsi, les joueurs sont prêts à sacrifier leurs propres biens matériels afin de récompenser une personne jugée juste (qui a précédemment fait une offre juste) ou de punir une personne jugée injuste (qui a précédemment fait une offre basse). Ce jugement est également modulé par les attributions faites sur les motivations et les contraintes de l'autre joueur, ainsi que par le contexte de la situation (Dufwenberg & Kirchsteiger, 2000 ; Falk et al., 2008 ; Rabin, 1993 ; Sonnemans et al., 1999).

Notre recherche se concentre sur l'attribution de l'intention du proposant avec une offre minimale dans un jeu de l'ultimatum à deux tours de contre-offre. Plus précisément, nous faisons l'hypothèse que cette attribution peut être transmise par le contexte conversationnel créé par des informations linguistiques pertinentes (Wilson & Sperber, 2002) sur le proposant. Plusieurs études sur le jeu de l'ultimatum ont examiné les mécanismes qui sous-tendent la réaction des proposant et des répondants en fonction de

la soumission d'informations descriptives et psychologiques sur le proposant avant le début du jeu (Marchetti et al., 2011) et de l'induction d'émotions positives ou négatives avant le début de la négociation (Forgas, 1998; Petit, 2009). Ces résultats ont montré que les participants peuvent être impactés par leurs expériences passées. Ainsi, chaque acceptation ou refus d'une offre doit être interprété dans l'ensemble du système. Si nous considérons l'interaction comme importante dans l'expérience des négociations, il est alors essentiel de la considérer comme un système similaire à la communication où toutes les parties et éléments de ce système interagissent les uns avec les autres (von, 2012). Notre deuxième question est de savoir si nous pouvons obtenir les mêmes attributions intentionnelles lorsque le contexte conversationnel est limité à un échange textuel de type *chatbot*. L'attribution intentionnelle véhiculée par un contexte conversationnel pertinent est-elle identifiée de la même manière lorsque le participant pense avoir une discussion avec un agent artificiel à travers un échange textuel? Nous supposons que la présence d'une intention véhiculée par un agent artificiel via un contexte conversationnel pertinent restreint au texte influence la prise de décision de refuser ou d'accepter une offre d'une manière similaire au contexte conversationnel plus informatif entre deux personnes en face à face. Un avantage important de ce type d'interaction est qu'il permet de se concentrer uniquement sur l'énoncé et d'isoler complètement les comportements non verbaux. En effet, nous avons démontré dans la partie précédente de cette thèse que les personnes participant à des expériences de type *TT* produisent des inférences et des attentes concernant les conversations textuelles. Les partenaires conversationnels essaient notamment de coopérer pendant une conversation en s'attendant à ce que les participants respectent les 4 *maximes* de Grice (1975); 1) la *maxime* de qualité, axée sur la véracité et la certitude d'une information donnée, 2) la *maxime* de quantité, axée sur la quantité d'informations données (ni trop ni trop peu), 3) la *maxime* de relation, qui suggère que les participants à une conversation s'attendent à ce que leur interlocuteur soit pertinent, et 4) la *maxime* de manière, axée sur la façon dont les informations sont données (brièvement, clairement, de façon ordonnée et sans ambiguïté) (Jacquet et al., 2018, 2019; Jacquet, Hullin et al., 2019a; Jacquet, Jaraud et al., 2021). Ainsi, la connaissance préalable de

la conversation semble faciliter l'interprétation des messages écrits en diminuant le coût cognitif nécessaire pour en déduire le sens.

Ici, nous donnerons à nos participants la tâche de jouer à un jeu d'ultimatum à deux tours avec trois joueurs dont les profils et les intentions peuvent être déduits par les participants avant le début du jeu grâce à des indices explicites et implicites. Les participants commenceront toujours le premier tour en faisant une première offre et au deuxième tour du jeu, les trois profils de joueurs différents proposeront l'offre minimale d'une unité. Contrairement aux expériences classiques où l'offre minimale est perçue comme une offre agressive parce qu'elle n'est pas justifiée, les participants joueront avec des joueurs dont les profils particuliers peuvent suggérer une offre minimale.

1. Le joueur « stochastique » dont l'offre d'une unité apparaîtra au participant qui répondra comme le résultat du hasard (Blount, 1995) et pourra donc être justifiée comme indépendante de la volonté et des intentions du joueur,
2. Le joueur « homo-economicus, rationnel » , dont l'offre d'une unité peut être comprise par le participant répondant comme la conséquence d'une décision rationnelle de maximiser son profit. Cette décision, explicitée au joueur, peut donc être analysée comme non agressive puisqu'elle est soutenue par la théorie de la décision économique et par la théorie du joueur « rationnel » .
3. Le joueur explicitement « égoïste » dont l'offre d'une unité peut être déduite par le participant répondant au début du jeu. Dans ce cas, la décision émane de la personnalité du joueur et doit être perçue comme agressive et comme une violation délibérée de la norme d'équité.

7.2 Expérience

Participants

92 participants (58 hommes et 34 femmes) ont participé à notre expérience. L'âge de nos participants allait de 18 à 39 ans (24,2 ans en moyenne, $\sigma = 2,64$). Les participants ont été recrutés par l'intermédiaire de connaissances. Tous les sujets étaient de langue maternelle française et ont donné leur consentement éclairé. Nous avons effectué 7 pré-tests afin de valider notre méthode expérimentale. Les participants ont été placés au hasard dans l'une des deux conditions de l'expérience :

- La condition **chatbot** était composée de 40 hommes et 22 femmes. Dans cette condition, les participants étaient confrontés à un jeu d'ultimatum via une interface de **chatbot**. Le joueur n'était pas visible et le participant pouvait penser qu'il jouait contre une machine.
- La condition face à face était composée de 18 hommes et 12 femmes. Les participants ont été confrontés à un jeu d'ultimatum en face à face. Le joueur était physiquement présent.

Matériel et Méthode

Les participants ont été confrontés à trois profils de joueurs correspondant aux trois différentes intentions que nous souhaitions transmettre. Dans ce jeu de l'ultimatum « revisité », les participants commencent le jeu en faisant une offre au joueur et ils doivent ensuite répondre à l'offre du joueur (qui sera toujours minimale). Ce choix de placer les participants en tant que premier proposant évite aux participants d'être soumis en même temps à (1) l'intention du joueur et (2) à l'émotion négative d'avoir l'offre « déloyale » (minimale) du joueur (Pillutla & Murnighan, 1996). Le joueur a mené l'ensemble du dialogue jusqu'à sa conclusion. Il se présentait et donnait aux participants des indices leur

permettant de déduire son intention par le biais d'un contexte conversationnel pertinent qui définit son mode d'action (joueur rationnel / joueur stochastique / joueur égoïste). Chaque profil avait un taux d'acceptation différent en fonction de l'offre du participant. Dans un deuxième temps, c'était au tour des participants de répondre à l'offre du joueur. Toutes les offres étaient similaires (offre de 1 unité). Avant d'émettre son offre, le joueur justifiait à nouveau son intention. Lorsque les participants ont fini de discuter avec les trois joueurs expérimentaux, ils ont été invités à répondre à un questionnaire afin d'évaluer leur adaptation stratégique au cours des différents jeux.

Procédure

Notre plan expérimental est un plan semi-complet qui inclut la nature de l'interaction ([chatbot](#) ovs face à face) comme facteur imbriqué et l'intention véhiculée par le contexte conversationnel pertinent des joueurs (stochastique, égoïste et rationnel) comme facteurs croisés. L'expérience prend en compte 2 variables indépendantes. La première variable correspond à l'ordre dans lequel le joueur est utilisé (randomisé sauf pour le joueur stochastique qui est toujours le premier joueur) tandis que la seconde variable représente les différents contextes conversationnels des différents joueurs expérimentaux (stochastique, égoïste et rationnel). Nous avons également 3 variables dépendantes : La valeur de l'offre du participant au premier tour (de 0 à 10 unités), l'acceptation par le participant de l'offre déloyale au second tour (acceptation ou refus), et l'évaluation par les participants de l'adaptation stratégique (échelle de Likert allant de 1 à 5 : 1. Pas du tout d'accord / 2. Pas d'accord / 3. Sans opinion / 4. Plutôt d'accord / 5. Tout à fait d'accord). Nous avons pu discuter et debriefer les participants à la fin de l'expérience afin de recueillir leurs commentaires. Nous voulions également éviter de soulever des soupçons sur la régularité des offres déloyales de nos joueurs, nous avons donc placé le joueur stochastique en premier dans toutes les conditions (justifiant son offre déloyale par un processus aléatoire explicitement présenté comme tel au participant).

Dans le cas du [chatbot](#), l'expérience a été réalisée à l'aide d'une interface de [chat-](#)

bot que nous avons conçue. Tous les scripts des différents acteurs expérimentaux ont été soigneusement préparés à l'avance mais interprétés par un joueur humain. En effet, l'interprétation par un joueur humain permet d'éviter d'importants problèmes techniques ainsi qu'un manque d'adaptation aux réponses des participants. Dans un premier temps, nous avons fait semblant de mener une étude sur l'intelligence artificielle et que l'objectif était de mesurer le comportement de nos joueurs. L'expérimentateur a également pris le rôle de modérateur pour donner les instructions au participant. L'objectif de notre texte de recrutement et de nos instructions était d'éviter tout biais de suspicion sur l'évaluation de leur propre comportement. Nous avons également veillé à ce que les participants comprennent que les joueurs expérimentaux étaient des *chatbots* (Schanke et al., 2021).

Dans la condition de face à face, la procédure était la même chaque profil était joué par des acteurs humains. Les comportements verbaux et non verbaux des acteurs étaient limités et contrôlés afin qu'ils soient cohérents avec chacun des trois profils et identiques pour tous les participants. Les trois profils étaient divisés en un profil non-intentionnel (stochastique) et deux profils intentionnels : l'un était égoïste et prenait des décisions basées sur cette personnalité et l'autre était rationnel et prenait des décisions basées sur la rationalité. Lorsque le participant affirmait avoir compris l'instruction, l'expérience commençait.

Après une brève introduction sur le caractère aléatoire de sa future offre, le joueur stochastique demandait au participant de lui faire une offre. Suite à l'offre du participant, le joueur stochastique pouvait soit refuser, soit accepter l'offre (le seuil d'acceptation était une offre ≥ 3). Dans les deux cas, le joueur stochastique expliquait de manière neutre les conséquences de l'acceptation ou du refus. C'était ensuite au tour du joueur stochastique de faire une offre aléatoire (qui était générée sur l'ordinateur dans la condition face à face) qui était toujours égale à 1. Après la réponse du participant, le joueur stochastique expliquait à nouveau de manière neutre les conséquences de la négociation.

Les deux autres profils, joueur égoïste et joueur rationnel, étaient donnés dans un ordre aléatoire. Tous deux commençaient par se présenter et, ce faisant, donnaient des

indices sur leurs intentions (selon leur profil). Chaque profil avait un seuil d'acceptation différent, lié à sa personnalité. Après l'offre du participant, le joueur donnait son avis en fonction de sa personnalité. Avant de faire son offre, le joueur la justifiait par un contexte conversationnel pertinent (Wilson & Sperber, 2002) et donnait son avis sur la décision du participant après qu'elle ait été prise.

Le joueur au profil égoïste se présentait en affirmant « qu'il était le meilleur à ce jeu, et qu'il ne se souciait que de ses propres gains ». Le joueur égoïste montrait les caractéristiques d'une personne qui se sentait supérieure et qui était condescendante. Si une offre supérieure ou égale à 5 était donnée par le participant, elle était considérée comme « convenable » et « décevante » dans le cas contraire.

Le joueur au profil rationnel se présentait en disant qu'il était « un expert en économie et qu'il était purement rationnel ». Il acceptait toute offre (car refuser signifiait qu'il ne gagnerait rien) et quand il proposait l'offre minimale d'une unité, il se justifiait en précisant que c'était l'offre économiquement juste. Si le participant refusait, il disait alors qu'il ne comprenait pas ce refus qui n'était pas rationnel par rapport à la théorie économique.

A la fin de l'expérience, les participants étaient redirigés vers un questionnaire en ligne pour s'enquérir de leurs stratégies, de leurs préférences et de leurs éventuels retours.

Résultats

Tout d'abord, pour ce qui concerne la valeur de l'offre des participants au premier tour, l'analyse de variance révèle un effet principal significatif du contexte conversationnel sur la valeur de cette offre ($F_{2,180} = 14,20; p < 0,001$). Les participants sont globalement plus généreux dans la condition face à face que dans la condition *chatbot*. Nous observons également un effet principal significatif du profil du joueur sur la valeur de l'offre ($F_{1,90} = 13,68; p < 0,001$). Les différents tests post-hoc indiquent que le joueur stochastique est plus récompensé que les joueurs égoïstes et rationnels ($F_{1,90} = 36,80; p < 0,001$). Ces

derniers sont récompensés de manière similaire (Figure 7.1a).

La Figure 7.1b montre les taux d'acceptation des participants. Nous constatons que le contexte conversationnel influence le taux d'acceptation de l'offre minimale par les participants dans les deux conditions ($Q_{1,3} = 38,26; p < 0,001$ pour la condition `chatbot` et $Q_{1,3} = 12,78; p < 0,001$ pour la condition face à face). Nous avons réalisé différents tests post-hoc afin d'observer les différences entre les contextes pour les deux conditions (`chatbot` et face à face). Pour l'interface `chatbot`, nous observons que le taux d'acceptation est significativement plus élevé pour les joueurs stochastiques et rationnels par rapport au joueur égoïste ($p < 0,001$). Nous observons également que le joueur égoïste est plus rejeté que le joueur rationnel ($p < 0,001$). Concernant le jeu de l'ultimatum en face à face, nous observons une différence significative entre le joueur stochastique et les autres joueurs ($p < 0,01$) mais aucune autre différence n'a pu être trouvée.

Après l'expérience, nous avons demandé à nos participants de remplir un questionnaire concernant leur évaluation stratégique des joueurs (voir Figure 7.1c). Nous avons utilisé une échelle de Likert de 1 à 5 (1. Pas du tout d'accord / 2. Pas d'accord / 3. Pas d'opinion / 4. Plutôt d'accord / 5. Tout à fait d'accord). Nous avons effectué une ANOVA à mesures répétées sur les moyennes des réponses au questionnaire. Nous pouvons observer un effet principal du contexte conversationnel sur l'évaluation stratégique ($F_{2,180} = 170,90; p < 0,001$) mais aussi un effet d'interaction entre le contexte conver-

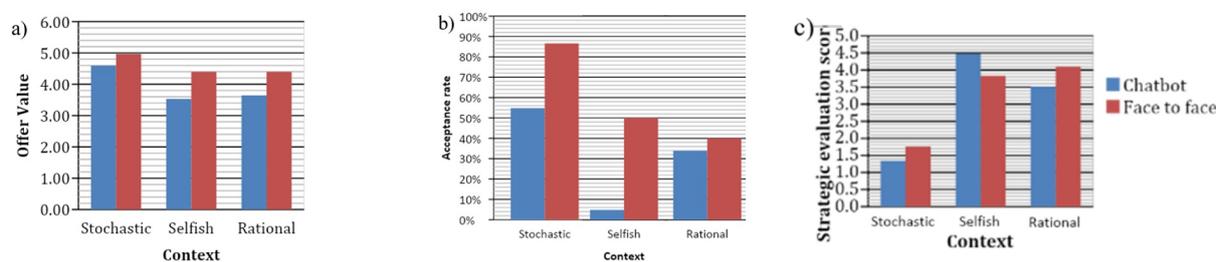


FIGURE 7.1 – a) Offre des participants en fonction du contexte conversationnel (en nombre d'unités); b) Acceptation de l'offre injuste en fonction du contexte conversationnel; c) Evaluation stratégique (Echelle de Likert à 5 points) des participants en fonction du contexte conversationnel. $N_{\text{chatbot}} = 62$; $N_{\text{face à face}} = 30$.

sationnel et la nature de l'interaction ($F_{2,180} = 9,65; p < 0,001$). Nous avons effectué plusieurs tests post-hoc qui montrent que l'adaptation stratégique des participants est plus importante concernant le profil égoïste par rapport aux profils rationnel ($p < 0,01$) et stochastique ($p < 0,001$). De plus, nous nous sommes intéressés à l'effet d'interaction entre la nature de l'interaction et les différents profils de joueurs. Nos tests révèlent que les participants se sont adaptés significativement plus au profil stochastique et égoïste dans la condition **chatbot** que dans l'interaction en face à face ($p < 0,001$). Pour le profil rationnel, nous remarquons que les participants ont eu le sentiment de s'adapter davantage dans l'interaction en face à face ($p < 0,01$).

7.3 Discussion

Dans les deux situations (**chatbots** ou en face à face), nous avons observé que l'attribution d'une intention à l'autre joueur avait un impact sur la décision des participants, de manière similaire à ce qui avait été montré dans (Falk et al., 2008). En effet, les participants acceptent généralement l'offre lorsqu'elle est faite par un joueur stochastique. Cependant, nous observons que les participants de la condition **chatbot** n'ont pas accepté les offres du joueur stochastique aussi facilement que dans la condition face à face. Nos discussions avec les participants pendant le débriefing indiquent que les participants dans la condition **chatbot** ne croyaient pas autant au caractère aléatoire du joueur stochastique que les participants de la condition face à face. Il est donc possible que cela ait eu une influence sur les taux d'acceptation.

Pourtant, lorsqu'il y avait une intention réelle derrière l'offre du joueur (soit en raison de sa personnalité dans le cas du joueur égoïste, soit en raison de sa stratégie dans le cas du joueur rationnel), les taux d'acceptation étaient beaucoup plus faibles. Cela est certainement dû au fait que les participants se réfèrent aux attentes sociales : si l'offre est effectivement aléatoire, le joueur n'y est pour rien et elle est donc acceptable, car refuser ne lui apprendrait pas à offrir davantage lors des tours suivants. Ceci est

particulièrement évident dans le cas du joueur égoïste dans la condition *chatbot*, pour lequel le taux d'acceptation tombe presque à 0. Ce joueur, en offrant 1 unité, évite de manière flagrante de coopérer avec le participant, et donc le participant riposte en refusant l'offre. Ce comportement est probablement rendu encore plus extrême par l'interface du *chatbot*, car les gens n'ont pas à faire face aux conséquences sociales du refus de l'offre, et peuvent donc être plus extrêmes dans leur intention d'apprendre (par leur refus) au joueur de ne plus jouer de la sorte.

Le cas du joueur rationnel est un peu différent. En effet, nous nous attendions à ce que les participants, comprenant les motivations du joueur, offrent un petit nombre d'unités et qu'au tour suivant, ils acceptent l'offre très facilement, sachant que le joueur rationnel devrait de toute façon l'accepter, et pourtant les participants ne semblent pas avoir changé de stratégie malgré le fait que le joueur leur donne la solution du problème, et donc coopère. Une interprétation possible est que les participants ont pu supposer des arrière-pensées, ou un désir personnel caché derrière l'explication rationnelle du comportement qui entre en conflit direct avec l'apparence coopérative du joueur. Ainsi, même si les participants ont probablement compris ce qui aurait été rationnel, la rationalité sociale a tout de même eu une influence plus forte sur leur prise de décision.

7.4 Conclusion

Dans cette expérience, nous avons montré que les gens ont tendance à s'en tenir à leurs habitudes de pensée et à ne s'adapter que marginalement à la rationalité des joueurs. Nous avons également montré que les participants comprennent la différence entre une offre intentionnelle et une offre non intentionnelle et répondent en conséquence lorsqu'ils reçoivent une offre déloyale. Enfin, nous avons également remarqué qu'un comportement égoïste entraîne une réaction beaucoup plus forte des participants lorsque ce trait de personnalité est affiché par un *chatbot* que lorsqu'il s'agit d'un humain en face d'eux. Ces observations soulignent l'importance des traits sociaux et des particularités

des interlocuteurs dans les interactions communicatives dans les [chatbots](#). En fonction du contexte, les participants vont potentiellement afficher des réponses plus extrêmes aux offres déloyales.

Chapitre 8

Les informations contextuelles aident à comprendre les messages écrits en textismes

Points remarquables dans ce chapitre

1. Les textismes sont une forme alternative d'expression écrite qui peut permettre d'étudier la compréhension du langage.
2. Nous utilisons le paradigme du test de Turing pour évaluer le coût cognitif de la lecture de textismes chez des jeunes adultes.
3. Nous observons que la connaissance au préalable des informations communiquées permet de comprendre plus facilement les textismes.

8.1 Introduction

On pourrait penser que les mots ont des significations spécifiques gravées dans la pierre, ou plutôt inscrites dans les dictionnaires. De nombreux éléments semblent indiquer

que cette vision est en fait bien trop simpliste et qu'elle ignore l'influence importante du contexte dans lequel ces mots sont utilisés. Non seulement un même mot peut avoir un sens différent selon le contexte, mais sa signification peut aussi évoluer lentement avec le temps (Michel et al., 2011 ; Reali & Griffiths, 2010 ; Wijaya & Yeniterzi, 2011). Les informations contextuelles peuvent également nous aider à apprendre de nouveaux mots, y compris dans des langues différentes, ou de nouvelles formes de langues (Çetinavcı, 2014 ; Paribakht & Wesche, 1999).

Depuis quelques décennies, de nouvelles formes alternatives de langage écrit destinées à optimiser le contenu pouvant être inséré dans les messages avec un nombre limité de caractères sont apparues avec le développement des téléphones portables, suivi de l'essor des conversations textuelles instantanées. Des mots raccourcis tels que « Kdo » pour « cadeau » ont commencé à être utilisés par les jeunes au grand dam des personnes âgées pour qui déchiffrer les messages envoyés par leurs propres enfants devenait une tâche cognitive à part entière. C'est ainsi que s'est répandue une forme alternative au langage écrit, appelée « textisme » .

Même si cette forme spécifique d'écriture a progressivement diminué, avec la possibilité d'écrire des messages plus longs et des correcteurs orthographiques automatiques¹, le textisme constitue un exemple intéressant de forme alternative de langage écrit qui mérite d'être étudié pour deux raisons principales. Tout d'abord, le textisme prend ses racines dans la sonorité de la langue connue mais a la particularité d'être transcrite sous une forme différente de celle habituellement utilisée. Un mot en textisme est donc difficile à déchiffrer lorsque l'on n'est pas habitué à cette forme d'écriture, d'autant plus s'il n'est pas associé des indices supplémentaires indiquant le sens du message. Il s'agit donc d'un outil précieux pour étudier l'acquisition progressive du sens et l'établissement de liens entre la forme d'un mot et son concept sous-jacent. Ensuite, bien que l'usage intensif du textisme (messages écrits uniquement en textismes) ait diminué, Les symboles et pra-

1. en particulier les formes contractives des textismes, tandis que les messages qui incluent un contenu plus émotionnel comme les smileys ou les signes de ponctuation supplémentaires : « Oh yeah!!!! :) » se sont plutôt répandus (Kemp, 2019 ; Kemp & Grace, 2017)

tiques d'écriture développés sont encore largement utilisés à ce jour (par exemple, les très courants « lol », « mdr » en français, et bien d'autres, mais aussi des contenus qui créent un pont entre la forme écrite et la forme orale de la communication en transcrivant des indices émotionnels qui auraient été présents dans l'intonation ou dans l'expression faciale, comme les smileys) (Kemp, 2019 ; Kemp & Grace, 2017).

Dans cet article, nous examinons si l'accès au contenu sémantique des messages écrits en textisme nécessite un coût cognitif plus élevé pour les personnes qui ne sont pas particulièrement habituées à l'utiliser et, plus important encore, si le contexte peut favoriser par des indices pragmatiques à retrouver le sens des messages.

8.2 Revue de la littérature

L'importance du contexte en conversation

La psychologie et la linguistique, étudient l'importance du contexte dans notre utilisation du langage à l'aide de théories provenant de pragmatique du langage. La grande conclusion de ses études est que le contexte influence grandement la façon dont nous interprétons les énoncés, non seulement dans les conversations (Gallagher, 2020 ; Jacquet & Baratgin, 2020) mais aussi dans les tâches de résolution de problèmes (Baratgin et al., 2020 ; Jamet et al., 2018 ; Pollard & Evans, 1987). Cependant ce résultat de l'importance du contexte dans l'inférence du sens a été en grande partie ignoré dans la conception d'agents numériques conçus pour converser avec les humains (voir Ghosh et al., 2018 ; Jacquet & Baratgin, 2019).

En effet, ce que nous *disons* n'est pas forcément ce que nous *voulons dire*. Cet aspect du langage a été décrit en premier lieu par Grice (1975) qui a essayé de formaliser les règles du discours par des **maximes** conversationnelles. Pour Grice, ce qui est le plus important est que les participants à une conversation suivent un principe de coopération,

ce qui signifie essentiellement qu'ils essaient de rendre ce qu'ils veulent dire aussi clair que possible afin qu'il soit correctement compris par l'autre. En d'autres termes, les participants à une conversation disent ce qu'ils pensent devoir être dit pour transmettre ce qu'ils veulent dire à l'autre. Par exemple, lorsque l'on dit « Il fait froid ici » pour signifier « Pouvez-vous fermer la fenêtre parce que j'ai froid ici et je n'aime pas avoir froid » , ou « Pouvez-vous me passer le sel ? » pour signifier « Donnez-moi le sel parce que vous êtes plus près et je ne peux pas l'atteindre moi-même » .

Cette approche a ensuite été élargi par la théorie de la pertinence (Relevance Theory) (Wilson & Sperber, 2002), qui permet de considérer les *maximes* conversationnelles de Grice en lien avec la cognition humaine. La théorie de la pertinence suggère notamment que la pertinence d'un énoncé dépend de deux attributs principaux : les effets contextuels (c'est-à-dire les effets de l'énoncé sur les représentations mentales de l'auditeur et le coût cognitif nécessaire à leur interprétation (c'est-à-dire combien de ressources dois-je allouer au traitement de la phrase pour la rendre compréhensible).

Dans les exemples cités ci-dessus (fenêtre, sel), on remarque que les phrases sont plus simples, mais surtout qu'elles sont perçues comme étant plus polies grâce à l'utilisation du discours indirect. Être direct pourrait en effet être perçu comme un commandement et serait préjudiciable à la relation de coopération (Leech, 2014).

Aucun de ces exemples ne dit directement à la personne ce qu'elle doit faire, et pourtant, ils seront largement compris comme « Ils veulent que je fasse quelque chose à ce sujet, et ne se contentent pas de me parler de leur état physique (exemple de la fenêtre) ou de m'interroger sur ma capacité à faire quelque chose (exemple du sel) » .

Ainsi, ce que nous disons n'est pas directement ce que nous voulons dire, et cette différence est attendue par nos interlocuteurs et comprise comme telle : les mots sont utilisés comme des indices pour accéder au sens et ce qu'ils représentent peut être plus étroit ou plus large que leur définition habituelle dans un dictionnaire, selon le contexte dans lequel ils sont utilisés ; (Carston, 1997, 2008 ; Sperber & Wilson, 1998 ; Wilson &

Kolaiti, 2017). On peut faire l'hypothèse que la forme produite ne devrait pas empêcher d'accéder au sens, à condition de donner suffisamment d'indices.

Les textismes comme une approche phonologique de l'écriture

Une façon de modifier la forme de la production linguistique sans en changer le sens est d'utiliser des formes d'écriture alternatives. Le textisme est une forme d'écriture utilisée dans les messages textuels lors de conversations téléphoniques. Étant donné qu'au départ, les messages textuels étaient limités par le nombre de caractères pouvant être envoyés sans coûts supplémentaires et par le temps nécessaire pour écrire une seule lettre sur un téléphone à 9 touches, les jeunes adultes du monde entier ont développé diverses formes d'écriture utilisées pour tirer le meilleur parti d'un nombre limité de caractères tout en étant capables de communiquer efficacement.

Pour les variantes françaises, Panckhurst (2009) a décrit et classé les nombreuses libertés prises dans les messages SMS par rapport à l'orthographe traditionnelle pour divers mots et expressions. Fairon et al. (2006) a également rapporté des observations similaires tout en soulignant certaines zones de résistance, comme l'utilisation des digrammes « ph », « rh » ou « th » qui sont contre-intuitivement conservés dans les textismes français, alors que l'orthographe traditionnelle de certaines autres langues les abandonne (comme l'italien). La plupart des changements simplifient radicalement l'orthographe de certains sons. Comme « o » qui remplace « eau » et que l'on retrouve dans d'autres mots comme « Kdo » pour « cadeau », « bato » pour « bateau » et bien d'autres. Certains de ces changements incluent la suppression des lettres muettes : « vou » pour « vous », « douch » pour « douche », et bien sûr des troncatures : « ordi » pour « ordinateur » et des acronymes comme « lol » pour « Laughing out loud » .

Pour les personnes qui ne sont pas habituées à cette forme d'orthographe, cela peut rendre la compréhension des messages assez complexe et devenir une forme d'exclusion. Elle contribue souvent à rendre les conversations inter-générationnelles difficiles, d'autant

plus que cette forme d'écriture tend à être utilisée principalement par les jeunes adultes qui utilisent généralement les textos plus souvent que les autres groupes d'âge : (Ling, 2010).

Il est donc probable que l'utilisation du textisme nécessite un coût cognitif plus élevé afin de comprendre le sens d'un message pour les personnes qui n'y sont pas habituées, mais que ce coût cognitif plus élevé pourrait être diminué en donnant des indices contextuels sur le sens des mots. Par exemple, le sens du mot cadeau (« Kdo » en textisme) est plus facile à déduire dans la phrase « tu ve 1 Kdo pr ton anniv? (Tu veux un cadeau pour ton anniversaire?) que dans la phrase « jéT ds la voiture ac un Kdo » (J'étais dans la voiture avec un cadeau). En effet, dans la première phrase, le contexte de l'anniversaire est rapidement lié au concept de cadeau, et cela diminuera très probablement la quantité d'effort (coût cognitif) nécessaire pour arriver à la conclusion que « Kdo » signifie cadeau, car cette interprétation donne le plus d'effet contextuel. Dans l'autre phrase, les mots « voiture » et « cadeau » ne sont pas liés, et la compréhension de l'un ne donne aucun indice permettant de déduire le sens de l'autre.

Les études explorant le coût cognitif de l'utilisation de textismes sont peu nombreuses. Elles semblent indiquer que l'utilisation de textismes est corrélée à une augmentation du temps de lecture et a également un effet sur certaines mesures liées à l'effort cognitif. Ainsi par exemple dans une comparaison de phrases écrites en textismes ou en anglais standard, les phrases en textismes demandent davantage de fixations. Ces fixations durent significativement plus longtemps que pour les phrases contenant de l'anglais standard (McCausland et al., 2015 ; Perea et al., 2009). Des résultats similaires ont été observés entre le Net-Speak² chinois et le chinois standard : Le Net-Speak augmente le temps de réponse lors d'une tâche de reconnaissance de texte (J. Chen et al., 2020).

2. Forme de langage utilisée dans les conversations textuelles en ligne.

Les temps de réponse pour estimer le coût cognitif

Le coût cognitif (la quantité de ressources mentales nécessaires pour traiter l'information) n'est pas directement accessible à l'observation. Il constitue cependant un élément clés de nombreuses théories en psychologie cognitive. En particulier, c'est l'un des deux facteurs (avec l'effet contextuel) permettant de définir la pertinence d'une information (Wilson & Sperber, 2002). En fonction des expériences, des mesures indirectes peuvent être effectuées pour estimer le coût cognitif nécessaire pour traiter les stimuli et y répondre. Certaines de ces mesures indirectes incluent l'enregistrement de la dilatation pupillaire (qui nécessite des instruments spécifiques et un contrôle important pour atténuer l'influence d'autres facteurs) (Van der Wel & van Steenbergen, 2018), le taux de réussite aux tâches primaires ou secondaires (Paas et al., 2003), et les latences telles que les temps de réponse (ou les temps de réaction pour les processus rapides) utilisés dans la présente recherche et la partie précédente de cette thèse (Jacquet, Baratgin & Jamet, 2019 ; Jacquet, Hullin et al., 2019a).

La mesure de temps de réponse a l'avantage d'être discrète et non-invasive pour les participants. Elle a également l'avantage d'être assez facile à mettre en œuvre dans les expériences en ligne et ne nécessite pas d'équipement spécifique. Néanmoins, l'interprétation de cette mesure doit tenir compte du fait qu'elle correspond seulement à un indicateur *potentiel* du coût cognitif. Elle n'en est pas une mesure directe. Les temps de réponse sont faciles à enregistrer dans les conversations en ligne et les expériences menées dans la deuxième partie de cette thèse montrent des résultats compatibles avec la façon dont le coût cognitif devrait varier lors de l'étude des violations des attentes conversationnelles (Jacquet, Baratgin & Jamet, 2019 ; Jacquet, Hullin et al., 2019a).

Dans la présente expérience, nous avons utilisé un protocole expérimental similaire, adapté pour étudier l'influence de l'utilisation de textismes par l'expérimentateur sur les temps de réponse des participants. L'utilisation de textismes augmente-t-elle significativement le temps de réponse des participants dans une conversation par rapport à

l'utilisation d'un style d'écriture plus traditionnel ? Cette augmentation est-elle atténuée quand la conversation écrite en textismes suit directement une première conversation sur le même sujet dans laquelle le textisme n'est pas utilisé ? Si c'est le cas, cela voudrait dire que le contexte est un facteur favorisant le déchiffrement de cette forme d'écriture.

8.3 Expérience

Cette expérience suit un protocole similaire à celles présentées dans la deuxième partie de cette thèse, en remplaçant la violation des maximes de Grice par l'utilisation de textisme dans la conversation.

Le protocole expérimental utilisé dans cette étude a été évalué par le comité d'éthique de l'association P-A-R-I-S.³

Participants

Tous les participants étaient de jeunes adultes, de langue maternelle française, qui utilisaient régulièrement leur téléphone pour envoyer des SMS (cette expertise dans l'utilisation des SMS était évaluée au début de l'expérience). Tous les participants devaient accepter par écrit de participer à cette expérience et étaient informés qu'ils pouvaient arrêter leur participation à tout moment. L'âge moyen était de 23 ans ($\sigma = 3,7$). Il y avait 23 hommes, 11 femmes et 4 participants ont choisi de ne pas répondre à cette question.⁴

3. disponible à l'adresse https://osf.io/eua6w/?view_only=25c5cfc7c0a48649559e817c31263c8

4. Nous avons initialement inclus deux autres participants qui ont ensuite été retirés de l'étude car ils ont montré une expertise dans l'utilisation du textisme dans les conversations (ils communiquaient eux-mêmes avec des textismes) et n'étaient donc pas homogènes avec les autres participants et ne pouvaient donc pas être inclus dans les analyses. Les autres participants n'ont pas essayé d'écrire en textismes en retour à l'expérimentateur, et certains même ont demandé de « parler normalement ». Malgré l'âge des participants, ce comportement n'était pas inattendu pour les deux raisons suivantes : 1) la diminution des formes contractives des textismes et 2) le fait que les textismes sont surtout utilisés dans des conversations informelles avec des personnes que l'on connaît déjà, ce qui n'était pas le cas ici puisque les participants ne connaissaient ni l'expérimentateur ni le personnage fictif.

Matériel

L'expérience a été mise en œuvre via un chat en ligne conçu spécifiquement pour cette étude. Les participants ont d'abord répondu à une série de questions socio-démographiques ainsi que sur la fréquence de leur utilisation de SMS. L'utilisation d'un téléphone était demandée pour participer à l'expérience. Après avoir accepté un formulaire virtuel de consentement éclairé et confirmé qu'ils étaient bien de langue maternelle française, les participants étaient envoyés sur une autre page où se déroulait une conversation interactive avec l'expérimentateur. En réalité, l'expérimentateur jouait deux rôles au cours des conversations : le rôle de modérateur et le rôle de l'interlocutrice appelée Elsa. Elsa était présentée au participant comme un personnage fictif qui serait joué par deux « acteurs » différents, l'un étant un [chatbot](#) et l'autre un humain (même s'ils étaient en réalité tous les deux joués par le même expérimentateur, selon le principe du [magicien d'Oz](#)).

Afin de standardiser les conversations autant que possible malgré la nature potentiellement chaotique des conversations, Elsa disposait d'une fiche de personnage pré-définie contenant toutes ses informations personnelles. Si l'expérimentatrice recevait une question inattendue, elle notait sa réponse afin de s'assurer qu'elle utiliserait toujours la même réponse avec tous les participants à l'avenir si la même question était posée. Les conversations étaient enregistrées et le moment (à la ms près) de la réception de chaque message sur le serveur était également enregistré. Le délai enregistré entre les messages correspondait à la différence entre le moment où le message était reçu par le participant et le moment où il envoyait sa réponse.

Après l'expérience, les participants devaient répondre à des questions supplémentaires afin de tester leurs compétences en matière de compréhension des textismes.

Méthodes

Les participants sont d'abord introduits dans les conversations par l'expérimentatrice affichée comme « Modérateur ». Le modérateur expliquait aux participants qu'ils allaient participer à un TT (leurs objectifs étaient de découvrir à travers deux conversations la vraie nature de leur interlocutrice Elsa. Était-elle un **chatbot** ou était-elle humaine?).

Ce test a été utilisé pour motiver les participants à converser activement avec les deux acteurs dans un but qui n'était pas explicitement lié à l'utilisation de textismes dans les conversations. De cette façon, les participants ne s'attendaient pas à ce que l'une des deux conversations soit écrite avec des textismes. Chaque participant a conversé deux fois avec Elsa, une fois avec l'expérimentateur qui l'incarne par le biais de messages écrits en orthographe traditionnelle (type de conversation : normal) et une fois par le biais de messages écrits en textismes (type de conversation : textismes).

La décision de savoir si les participants conversaient d'abord avec l'Elsa qui écrivait en orthographe traditionnelle ou en textismes était prise au hasard, avec des ajustements pour obtenir un nombre similaire de participants dans chaque condition. À la fin de chacune des deux conversations, les participants devaient indiquer au modérateur s'ils pensaient que le premier acteur avec lequel ils avaient conversé était un chatbot/IA ou s'il s'agissait d'un humain. Chaque conversation durait 7 minutes. Les participants avaient la possibilité d'arrêter la conversation à tout moment s'ils le souhaitaient.

Analyses

En raison de la grande variabilité des messages dans une conversation, il n'est pas possible de relier directement le délai entre les messages au coût cognitif nécessaire pour les traiter. En effet, la partie du délai qui nous intéresse (liée au coût cognitif de la compréhension du message et de la préparation d'une réponse à celui-ci) peut être occultée par d'autres facteurs qui ont un fort effet sur les temps de réponse des participants. Il est

donc important de réduire ce bruit avant de procéder à une analyse plus approfondie.

Un facteur de bruit clairement identifié est la longueur des phrases. En effet, les messages plus longs prennent plus de temps à lire, et la rédaction de réponses longues prend également plus de temps. Ainsi, on diminue fortement le bruit dans les données collectées en éliminant autant que possible ces deux aspects.

Pour ce faire, nous avons utilisé un modèle de régression linéaire prenant en compte la longueur du message de l'expérimentateur (L_e) et la longueur de la réponse du participant (L_p) pour prédire le délai théorique des conversations normales (D).

$$D = wL_e + xL_p + yL_eL_p + z \quad (8.1)$$

Cela nous a permis de calculer Δd : la différence entre le temps de réponse observé (d) et le délai théorique (D), ce qui nous donne une mesure où le bruit provenant de la longueur des messages est filtré.

$$\Delta d = d - D \quad (8.2)$$

Ce processus est similaire au calcul des temps de lecture résiduels : Gibson2004, Ludtke2006, à la différence près que nous prenons également en compte les temps d'écriture résiduels dans le calcul de Δd .

C'est cette différence qui a ensuite été analysée par une Anova de type III, avec l'ordre et le type de conversation comme deux facteurs. La normalité des données a été testée et les valeurs aberrantes ont été nettoyées (au-delà de 3 écarts-types au-dessus ou au-dessous de la moyenne, ce qui a permis d'éliminer 10 mesures sur 873)⁵.

5. Le script R complet des analyses est disponible à l'adresse https://osf.io/dejkh/?view_only=25c5cffc7c0a48649559e817c31263c8

8.4 Resultats

L'équation finale du modèle linéaire calculé sur les conversations normales était :

$$D = 0,27 \times L_e + 0,36 \times L_p - 0,0025 \times L_e L_p + 2,5 \quad (8.3)$$

Les Δd pour chaque condition sont présentés dans la figure 8.1. Nous avons observé un effet principal significatif du type de conversations (normales ou textismes)

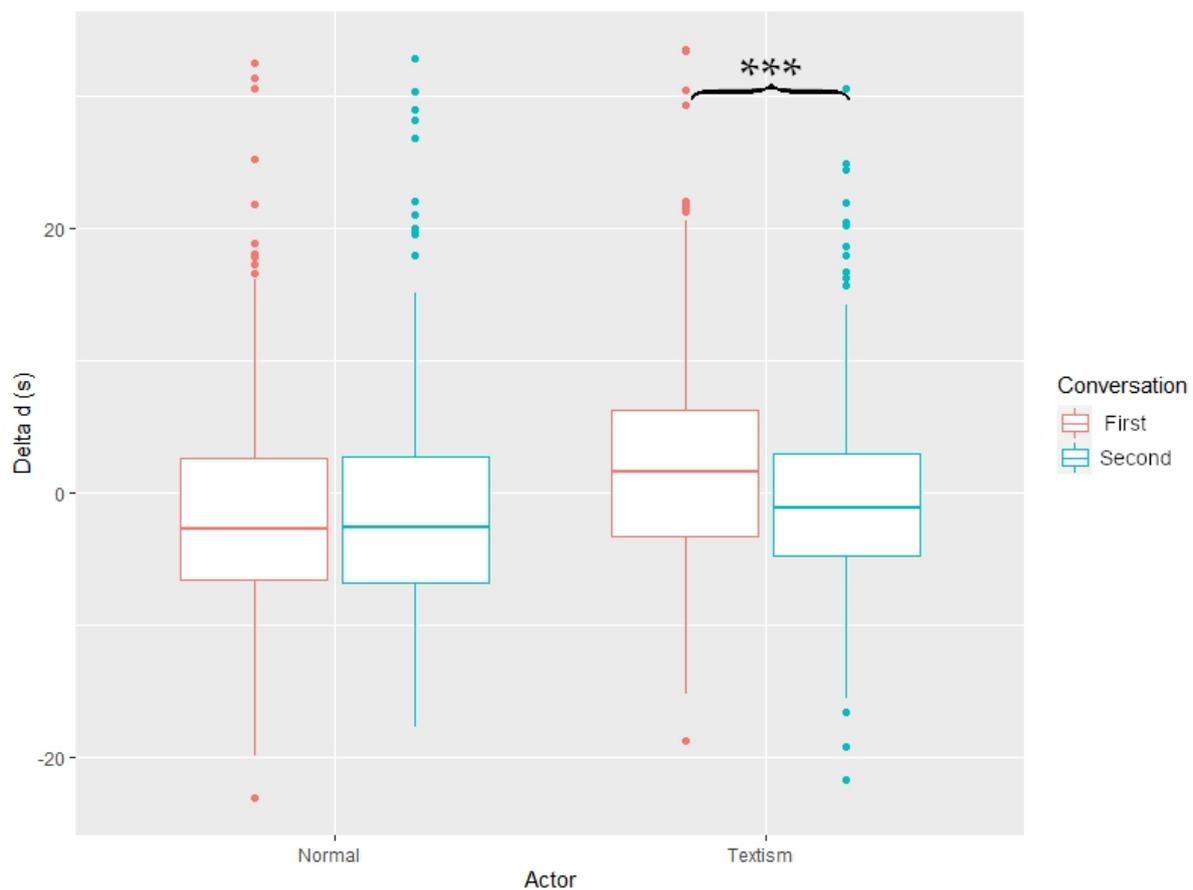


FIGURE 8.1 – **Influence des textismes par rapport au style d'écriture traditionnel sur Δd .** Les boîtes rouges indiquent que les données ont été enregistrées lors de la première conversation, les boîtes bleues indiquent que les données ont été enregistrées lors de la deuxième conversation. Les données enregistrées dans la conversation avec un style d'écriture traditionnel sont présentées à gauche (Normal) et les données enregistrées dans la conversation avec des textismes sont présentées à droite (Textisme).

*** $p < .001$

sur Δd ($F_{1,859} = 23,27; p < 0,001$). Il existe une interaction significative entre l'ordre des conversations (première ou deuxième) et le type de conversations (normales ou textismes) ($F_{1,859} = 8,261; p < 0,01$). Dans les conversations normales, Δd est similaire ($t_{414} = -0,7533; p = 0,45$) quel que soit leur ordre d'arrivées ($-1,35s; \sigma = 8,83s$ quand elles sont en premier et $-0,684s; \sigma = 9,18s$ quand elles sont en second). Dans les conversations avec avec des textismes, nous avons un effet d'ordre : Δd significativement plus élevé pour les premières conversations ($2,63s; \sigma = 8,75s$) que pour les secondes ($-0,109s; \sigma = 8,05s; t_{445} = 3,450; p < 0,001$).⁶

8.5 Discussion

Comme nous l'avions prévu, le textisme a augmenté de manière significative le temps de réponse des participants. Ce résultat est cohérent avec les résultats trouvés dans des études précédentes : (J. Chen et al., 2020; McCausland et al., 2015; Perea et al., 2009). En revanche, l'impact des messages écrits en textisme semble diminuer lorsqu'ils sont utilisés dans une seconde conversation qui suit une conversation avec une orthographe plus traditionnelle. Ce résultat est aussi cohérent avec nos hypothèses initiales. En effet, la diminution du délai lorsque les textismes sont utilisés dans la deuxième conversation indique que les participants ont un accès plus rapide au contenu sémantique des messages. Nous pouvons aisément supposer que cet effet est causé par les connaissances préalables extraites de la première conversation puisque la seule différence entre les deux conversations est la présence ou l'absence d'indices donnés avec la conversation précédente,

Le sujet des conversations étant similaire dans les deux conversations (comment les participants ont vécu les confinements en France durant la pandémie de COVID-19, ce qu'ils aimaient faire et leurs contacts avec leur famille), le participants, pour comprendre

6. Bien qu'il ne s'agisse pas de l'objet direct de cette étude, nous n'avons pas observé d'effet significatif des textismes sur les réponses au TT (à savoir si l'utilisation de textismes rendait Elsa plus humaine ou non). Nous n'avons pas non plus constaté d'effet significatif du genre des participants dans cette étude.

le sens des messages d'Elsa et lui répondre, devaient s'appuyer sur les connaissances préalables acquises lors de la première conversation. Un effet similaire avait été observé dans l'expérience décrite dans le chapitre 7 (Jacquet, Hullin et al., 2019a) : une façon plus compliquée de lire l'information (dans ce cas, en donnant trop d'informations inutiles) n'a pas eu d'effet sur les temps de réponse lorsque cela se produisait dans une deuxième conversation. Nous expliquions que cette différence est probablement due à un effet de surprise lorsque les violations conversationnelles se sont produites dans la première conversation. En effet, dans ce cas, l'information non-pertinente ajoutée a changé les représentations mentales des participants (effet contextuel fort), tandis que la phrase plus longue augmentait l'effort nécessaire pour la traiter. Dans la deuxième conversation, les participants connaissaient déjà l'élément d'information donné dans la réponse. Pour eux, il ne s'agissait que d'une confirmation de ce qu'ils savaient déjà (effet contextuel faible), mais pour cette raison, les éléments d'information non pertinents ajoutés n'ont pas augmenté de manière significative l'effort requis pour traiter la phrase (le coût cognitif est resté assez faible).

De même, dans le cas de cette expérience, les mêmes informations sur Elsa ont été données dans les deux conversations. Lorsque les participants ont rencontré et conversé avec Elsa, ils n'avaient aucune connaissance préalable de ce qui allait être dit. Lorsque ces informations étaient données sous forme de textismes, la seule façon d'en comprendre le sens était de les déchiffrer directement (effet contextuel élevé et coût cognitif élevé). Lorsque les textismes étaient utilisés dans la deuxième conversation, les participants connaissaient déjà la plupart des informations qui allaient être dites. Les mêmes informations étaient simplement données d'une manière différente (effet contextuel faible car il n'y avait pas beaucoup de changement dans les représentations mentales du participant). Ainsi l'effort nécessaire pour obtenir ces informations était plus faible car deux chemins alternatifs pouvaient être utilisés : L'un impliquait la mémoire du participant de la conversation précédente (effet du contexte de la conversation avec l'influence de la conversation précédente sur la seconde conversation), l'autre impliquait le déchiffrement des textismes. Ces deux méthodes pouvaient également être utilisées en parallèle pour

converger vers une seule interprétation du message, diminuant ainsi le coût cognitif (les participants n'avaient pas besoin de passer autant de temps à déchiffrer les textismes car les informations dans leur mémoire fournissaient déjà une partie du sens). Cette interprétation est cohérente avec des effets observés lors de la lecture. Des informations données en amont sur l'apparition d'inférences prédictives (Guéraud et al., 2008).

On peut se poser la question si l'effet que nous avons observé est causé par un amorçage lexical (le fait de voir un mot écrit une fois incite le lecteur à s'y attendre à nouveau à l'avenir) (Bock, 1986 ; Hoey, 2012) ou est un effet topique (la connaissance de l'information sur le point d'être donnée aide à déchiffrer le texte non familier). Dans notre expérience, les sujets de discussion étaient similaires entre les deux conversations, mais aucune restriction n'était imposée sur les mots exacts utilisés dans les réponses de l'expérimentateur tant que l'information donnée restait cohérente (Elsa ne pouvait pas avoir deux frères et sœurs dans une conversation et aucun dans une autre par exemple, mais l'expérimentateur pouvait l'exprimer de différentes manières en fonction de la question posée). Cette question pourrait probablement trouver une réponse dans une expérience future en concevant un protocole qui testerait indépendamment l'influence de l'amorçage lexical et de l'effet topique en proposant par exemple une condition dans laquelle l'information serait donnée avec exactement les mêmes mots et une autre condition dans laquelle l'information serait la même mais donnée avec des mots différents.

Tout ceci indique, comme nous nous y attendions, que le déchiffrement des textismes, sans information préalable sur le contexte de la conversation, est une tâche cognitive qui demande un effort cognitif important. Cependant, le déchiffrement des textismes semble plus facile lorsque les participants s'attendent au contenu sémantique des messages.

Il est également important de noter que dans notre expérience, les participants étaient tous de jeunes adultes. Ainsi, même pour les participants qui se déclaraient ne pas être des experts en textismes, ces derniers possédaient vraisemblablement quelques rudiments de textismes. Nous pouvons certainement faire l'hypothèse que les différences trouvées entre l'écriture traditionnelle et les textismes seraient encore plus marquées avec

des participants plus âgées qui ne seraient pas habituées aux textismes. En revanche, il est difficile de prédire quelle diminution des temps de réponse serait observée lorsque les participants plus âgés rencontreraient des textismes dans la deuxième conversation par rapport à la première. En effet, si nous pouvons envisager une plus grande difficulté des participants plus âgés pour déchiffrer des textismes que pour les participants plus jeunes (et donc trouver un Δd plus élevé), il est difficile de dire si ces participants plus âgés seraient ou non capables d'utiliser les connaissances obtenues lors de la première conversation aussi facilement que les jeunes participants. Deux versions alternatives de résultats pourraient alors être envisagées pour une telle expérience avec des participants plus âgés :

- 1) Un Δd plus élevé que celui observé dans la première conversation avec des textismes (en raison du coût cognitif plus élevé requis), mais dans la deuxième conversation, un Δd proche de la ligne de base (autour de 0s) si les informations préalables étaient suffisantes pour désambiguïser complètement les textismes ; ou 2) Δd pourrait atteindre un niveau intermédiaire entre la valeur observée dans la première conversation et la ligne de base si les informations préalables n'ont aidé que partiellement à déchiffrer les textismes.

8.6 Conclusion

Nos résultats semblent indiquer que les participants ont moins de difficultés à traiter les messages contenant des textismes français lorsqu'ils connaissent déjà le contexte de la conversation. Ceci suggère qu'une façon de s'habituer à des formes alternatives de langage pourrait consister à s'entraîner en utilisant un contexte familier. D'autres études doivent être menées pour examiner si cette accoutumance à cette forme d'écriture (le textisme) peut être considérée comme un réel apprentissage des textismes ou si elle n'est que circonstancielle et ne permettrait donc pas de comprendre les textismes dans un contexte différent.

Les résultats de cette étude fournissent également des preuves supplémentaires sur le fait que la compréhension de la langue ne dépend pas strictement de la forme des mots

utilisés, mais repose plutôt sur les attentes concernant le message lu, celles-ci influençant grandement la façon dont nous lisons ce message et dont nous le comprenons.

Conclusion Générale

Vers des agents conversationnels pragmatiques

Au cours de cette thèse nous avons cherché à répondre à la problématique de l'évaluation de la qualité de la pragmatique conversationnelle d'agents artificiels, et tout particulièrement pour les [chatbots](#). Comme nous l'avons montré au cours des chapitres successifs, la pragmatique conversationnelle est un élément fondamental de l'évaluation de chatbots qui est encore trop peu évalué de manière systématique.

Nous avons en effet vu à quel point la pragmatique conversationnelle est importante car les conversations ne se limitent pas au seul contenu verbal, mais également à tous les indices permettant aux interlocuteurs de mieux inférer ce que chacun *veut réellement dire* en disant ce qu'il *dit*. Ces indices vont des intonations de la voix, au regard, en passant par les gestes et les expressions faciales. Tous ces éléments permettent de construire une relation avec l'entité artificielle comme le montrent Mutlu et al. (2009) avec les regards du robot dirigés vers son interlocuteur humain et permettent de faciliter leur anthropomorphisation (Masson, Baratgin & Jamet, 2017a; Masson et al., 2016). Bien que les [chatbots](#) ne puissent généralement pas utiliser leur corps (puisqu'ils n'en possèdent pas), les codes de la communication non verbale peuvent tout de même être appliqués car certains éléments physiques (intonation de la voix, expressions faciales...) se sont déplacés vers le domaine textuel avec un codage spécifique (l'utilisation de ponctuation supplémentaire ou de majuscules pour représenter l'intonation, les émoticônes pour les expressions faciales...). Ainsi, la pragmatique du langage, y compris dans ces éléments

non-verbaux, s'applique également aux [chatbots](#). La bonne compréhension et utilisation de ces codes est donc primordiale pour permettre la coopération humain-machine.

Or, comme nous l'avons montré dans le deuxième chapitre de la première partie, les [chatbots](#) qui existent à l'heure actuelle ont beaucoup de difficulté à comprendre ne serait-ce que le contenu verbal des messages des utilisateurs, non seulement le contenu implicite bien sûr, mais également parfois le contenu explicite. En effet aucun des [chatbots](#) que nous avons utilisés pour notre expérience n'a été en capacité de maintenir une conversation sur un même sujet pendant plus de deux messages... et uniquement lorsque cette suite de message avait été prévue par les programmeurs du [chatbot](#) (exemple de la couleur préférée de Mitsuku). Cette incapacité à maintenir le sujet de conversation est le principal obstacle à l'émergence de réelles conversations avec ces outils.

Afin de résoudre ce problème, nous avons proposé un modèle de [chatbot](#), inspiré des modèles actuels d'encodeur/décodeur mais en appuyant sur l'aspect modélisation des pensées et attentes de l'utilisateur. Nous pensons qu'avec un tel module le [chatbot](#) pourrait alors biaiser ses réponses en fonction de ce qu'attend de lui son utilisateur. Bien sûr nous n'avons à ce stade pas encore proposé d'implémentation permettant de réaliser ce but. Le but de cette thèse étant avant tout de poser le problème et de le définir le plus clairement possible. Cependant, des modèles d'IA récents basé sur la structure du cortex des mammifères ([HTM](#), voir Hawkins et al., 2019, pour des détails) pourraient permettre de réaliser ce but car ils sont doté d'une capacité d'apprentissage rapide tout en gardant une bonne flexibilité sur le modèle du monde qu'ils infèrent des données. Si beaucoup de travail reste à accomplir, nous pensons qu'un système de [chatbot](#) performant doit pouvoir : 1) inférer l'état mental de l'utilisateur, 2) prédire le message de l'utilisateur en fonction de l'état de la conversation, 3) mettre à jour son modèle de l'état mental de l'utilisateur en fonction de si sa prédiction était correcte ou non, et 4) de proposer une réponse à l'utilisateur tenant compte de ce modèle tout en mettant celui-ci à jour.

Dans la deuxième partie de cette thèse, nous avons montré que le [test de Turing](#) ([TT](#)) (Turing, 1950), bien que très critiqué dans le domaine de l'informatique, reste un

outil qui demeure à ce jour inégalé dans sa capacité à évaluer la qualité générale d'un [chatbot](#) par ressemblance à l'être humain. Si nous convenons volontier que le [TT](#) n'est pas directement un test pour mesure l'intelligence d'un [chatbot](#) (en effet, une réponse artificielle peut donner des informations plus « intelligentes » que celles données par un être humain), il reste néanmoins le seul outil capable de s'approcher d'un test de la qualité pragmatique des [chatbots](#), et en particulier de la capacité (ou l'incapacité) de ceux ci à témoigner d'une théorie de l'esprit (Premack & Woodruff, 1978). En effet, aucune autre technique ne permet de faire ce type d'évaluation avec une précision convenable, même si elles ont l'avantage considérable par rapport au [TT](#) d'être plus facilement automatisable (Comme BLEU et METEOR).

Ce manque d'automatisme du [TT](#) peut potentiellement être palié grâce à l'utilisation des temps de réponse des utilisateurs en conversation comme nous l'avons montré dans les deux derniers chapitres de la deuxième partie de cette thèse. En effet, ceux-ci permettent une évaluation en direct des anomalies dans le temps de réponse des participants, permettant d'attirer l'attention vers ces messages problématiques plutôt que vers d'autres. Le fait que les [TR](#) à l'intérieur du [TT](#) soient cohérent avec les prédictions de la théorie de la pertinence (Wilson & Sperber, 2002) montre qu'ils représentent une piste d'étude intéressante pour permettre une meilleur automatisation de l'évaluation des [chatbots](#) afin qu'ils puissent provoquer, chez le participants, un coût cognitif moindre. Reste cependant à trouver une solution pour permettre la détection des effets contextuels sur les représentations mentales de l'utilisateur. Si ces deux éléments étaient combinés il devrait alors être possible d'automatiser presque entièrement l'évaluation pragmatique d'agents conversationnels artificiels.

Nous avons vu dans la troisième et dernière partie de cette thèse que l'utilisation des [chatbots](#) et du [TT](#) ne se limitent pas à des applications purement informatiques, mais que ces deux outils peuvent très bien être utilisés pour répondre à des questions théoriques sur l'humain et son comportement en interaction, aussi bien avec des machines qu'avec d'autres humains. En effet, nous avons pu voir que l'utilisation de [chatbots](#) dans un

paradigme de l'ultimatum game (Güth et al., 1982) permet de révéler des réactions parfois beaucoup plus fortes des utilisateurs selon la personnalité de leur interlocuteurs lorsqu'ils interagissent avec des *chatbots*. L'utilisation de *chatbots* dans des études concernant l'influence des normes sociales sur le raisonnement est donc particulièrement intéressant et pourrait permettre de comprendre davantage la rationalité humaine. On peut par exemple citer le comportement de participants dans le paradigme de l'effet de dotation (Thaler, 1980). En effet l'effet observé du refus de l'échange d'un objet pour un objet équivalent est neutralisé lorsque la tâche est donnée par un robot NAO, indiquant bien que cet effet est dû aux normes sociales appliquées aux humains (Masson et al., 2015).

Enfin, le *TT* et les *chatbots* peuvent également être utilisés pour explorer d'autres formes de communication. Nous avons pu observer que l'utilisation de textismes (Fairon et al., 2006 ; Panckhurst, 2009) impliquait un coût cognitif plus important pour les interlocuteurs, en particulier si ils ne connaissaient pas le sujets de conversation. Cet effet était en revanche neutralisé si les participants avaient connaissance du sujet de conversation au préalable. Cela indique bien que les connaissances préalables peuvent influencer le coût de certaines inférences, comme le montrent également Cook et Guéraud (2005) et Guéraud et al. (2008). Ainsi peut-être serait il possible d'envisager par exemple l'apprentissage de langues différentes avec ce type de méthode : une première conversation dans la langue de l'utilisateur suivie d'une autre conversation dans une autre langue, mais dans laquelle les informations données seraient sensiblement les mêmes. Peut-être cela permettrait-il de neutraliser le coût cognitif plus élevé d'une conversation avec un interlocuteur communiquant dans une autre langue que la sienne. Reste à savoir si ce type de méthode permettrait un apprentissage sur le long terme de la langue en question.

Ainsi, et pour résumer, cette thèse met en valeur l'importance de l'évaluation de la pragmatique pour les agents conversationnels artificiels et propose d'utiliser le *TT* associé à la mesure des temps de réponse au cours des conversations. En effet, aucune autre mesure n'est à l'heure actuelle capable d'obtenir une évaluation aussi fine des aspects pragmatiques indispensables au bon déroulé d'une conversation. Nous pouvons désormais

confirmer que le paradigme de l'utilisation du TT associé aux temps de réponse permet de détecter des effets de violations y compris lorsque tout autre aspect (compétence orthographique, compréhension des messages par l'agent conversationnel) est neutralisé. Comme nous l'avons vu dans la deuxième partie de cette thèse, le TT permettait au participant de différencier clairement deux interlocuteurs en fonction de leur usage de la pragmatique, toute chose ayant été égale par ailleurs (le style d'écriture était le même, la personnalité était la même, la qualité de l'orthographe était la même, les informations données étaient les mêmes...). Tout autre type de mesure aurait donné des faux positifs, considérant comme valide des réponses qui sont, d'un point de vue syntaxique ou sémantique, parfaitement correctes mais qui ne respectent pour autant pas les attentes conversationnelles des interlocuteurs, ou auraient pu au contraire générer des faux négatifs en considérant comme non-valides des contributions pour l'unique raison que celles-ci ne correspondaient pas, sur le plan sémantique, à la réponse de référence, tout en étant pourtant parfaitement valides sur le plan de la pragmatique conversationnelle.

Comme nous l'avons vu, le plus gros problème du TT reste son manque d'automatisation. Il n'est pas toujours possible de proposer des TT chaque fois que l'on souhaite évaluer un agent conversationnel artificiel. C'est en cela que la mesure des temps de réponse peut avoir une grande valeur. En effet comme nous l'avons vu, une augmentation des temps de réponse est généralement visible lorsqu'il y a violation des attentes des utilisateurs, et cette augmentation va généralement de pair avec le sentiment d'humanité donné par l'agent conversationnel, évalué par la réponse binaire au TT. Si l'objectif est donc l'évaluation automatisée (ou semi-automatisée) d'agents conversationnels, nous suggérons fortement de tenir compte de la mesure de ces temps de réponse (ajustés pour y retrancher la variabilité attendue du fait de la longueur des messages). Un message de l'utilisateur arrivant avec un retard significatif par rapport au délai théorique de base lié au nombre de caractères ou de mots dans la phrase indique que l'utilisateur a dû faire un effort supplémentaire pour répondre au chatbot. La détection automatique de ce type d'anomalies pourrait donc permettre de mettre en évidence des points sensibles dans des conversations sans avoir à demander un score de satisfaction à l'utilisateur suite à

chaque message du *chatbot*. Néanmoins il demeure important d'inclure le *TT* à la boucle d'évaluation des *chatbots*, même si d'autres outils peuvent être utilisés au préalable pour résoudre des problèmes autres que les problèmes liés à la pragmatique. Cette faculté du *TT* associé aux temps de réponse à évaluer le respect des attentes, comme nous l'avons vu, permet de répondre à des questions de natures plus théoriques en permettant de donner un élément de mesure du coût cognitif tel que décrit par (Wilson & Sperber, 2002) afin d'objectiver cette notion cruciale dans l'étude de la pragmatique conversationnelle. Il devient ainsi relativement simple de tester des hypothèses en évaluant si tel ou tel type de réponse produit effectivement un coût cognitif plus important ou non au sein d'une conversation.

Les *chatbots* et autres objets sociaux, comme les robots sociaux, ont un avenir prometteur dans de nombreux domaines plus large que la relation client. Ils pourraient être des outils particulièrement intéressants pour s'adapter au maximum aux situations particulières des utilisateurs, par exemple dans le cas de l'éducation. Il serait possible d'envisager le développement de *chatbots* afin de faciliter la diffusion de connaissances (Jaillet, 2020), notamment dans l'éducation à distance (voir par exemple Heller et al., 2005 ; Palasundram et al., 2019). L'utilisation de robots sociaux peut aussi être pertinente dans le domaine éducatif en présence, notamment grâce à la méthode du *learning-by-teaching* (Jamet et al., 2018), comme par exemple aider à lutter contre la dysgraphie à l'aide de robots (Gargot et al., 2021). Pour pouvoir être pleinement adoptés en évitant les problèmes de frustrations, il est cependant important de choisir entre une approche pragmatique du traitement du langage en conversation, et donc des travaux plus importants dans le domaine de la compréhension et de la production du langage naturel, ou au contraire utiliser des techniques qui ne nécessitent pas de telles technologies et en font part explicitement aux utilisateurs afin de ménager leurs attentes, comme le fait par exemple Woebot⁷. Ne pas faire de choix entre ces deux options résulterait, comme nous l'avons montré, au déclenchement de problèmes de non-respect des attentes de l'utilisateur, provoquant ainsi des frustrations. Le traitement de l'implicite commence néanmoins

7. Woebot Labs Inc. <https://woebot.io>

à faire son chemin dans les agents conversationnels, notamment avec la classification des intentions de l'utilisateur (Li et al., 2017), mais la route vers des agents conversationnels capables de maîtriser l'art de la conversation ne fait que commencer, et nous voulons appuyer le fait que l'aspect théorique des obstacles qui seront rencontrés ne doit pas être négligé : ces obstacles ne seront pas uniquement des obstacles techniques, mais seront aussi des obstacles théoriques sur le fonctionnement des conversations entre humains...

Travaux futurs

De nombreux travaux restent à mener, en particulier sur le développement du modèle d'agent conversationnel artificiel présenté dans la première partie. Il serait pertinent de chercher à l'implémenter concrètement afin de permettre d'améliorer progressivement ce modèle, ainsi que de tester la validité de chacune de ses fonctions en conversant avec le [chatbot](#) qui en résulterait. Cela permettrait certainement de développer plus rapidement un modèle cognitif de l'humain en conversation. Ainsi non seulement la connaissance de l'humain permettrait d'améliorer le [chatbot](#), mais l'utilisation du [chatbot](#) permettrait de tester des hypothèses pour mieux comprendre la cognition humaine du traitement du langage en conversation. Nous pensons que le modèle [HTM](#) est particulièrement prometteur pour une implémentation de ce type (Hawkins et al., 2019).

En effet, les chatbots actuels sont généralement développés de manière procédurale (si-alors), en définissant des règles qui s'apparentent au livre d'instructions décrit par Searle et al. (1980) dans son argument de la chambre Chinoise. Comme nous l'avons montré, il est très peu probable qu'une [IA](#) fonctionnant de la sorte puisse un jour atteindre un niveau de performance pragmatique similaire à celui d'un humain lambda (McDermott, 2014), même s'il faut bien noter qu'il est de plus en plus plaisant de converser avec elles. Un agent conversationnel doit pouvoir s'adapter, doit pouvoir modifier sa représentation du monde et de l'utilisateur en fonction de ses expériences conversationnelles (voir ses expériences physiques dans le cas d'agents incarnés). La tâche qui sera certainement la

plus importante dans le but d'obtenir un agent doté d'un sens pragmatique sera le développement d'une [ToM](#) (Premack & Woodruff, 1978). Bien que cela puisse sembler être une montagne à dépasser, il est important de se rappeler que la [ToM](#) consiste avant tout dans le fait d'inférer les états mentaux d'un autre à partir d'observations concrètes, et de déductions. Sans aller dans les questions philosophiques et chercher à savoir si il est possible d'avoir une [ToM](#) sans avoir d'esprit ou de conscience, il ne semble pas insurmontable de créer un système capable de générer, par inférence, un modèle lui permettant de faire des prédictions sur le comportement d'un utilisateur. Exprimée ainsi, la tâche semble bien moins insurmontable. En effet, inférer un modèle plus ou moins complexe dans le but de générer des prédictions, c'est exactement ce pourquoi des milliers de systèmes d'[IA](#) sont développés chaque jour à travers le monde. Les réseaux de neurones artificiels sont excellents dans ce domaine et, selon la tâche qu'ils doivent effectuer, de nombreux types d'organisation de ces réseaux existent, plus ou moins spécifique à certains types de tâches, comme les [RNN](#) pour l'inférence de modèles permettant d'effectuer des prédictions sur des informations arrivant de manière séquentielle. Alors même que ces circuits de neurones n'ont pas d'esprit, il est possible aujourd'hui de développer une proto [ToM](#) en détectant par exemple l'intention implicite de l'utilisateur (Li et al., 2017). Bien d'embryonnaire à l'heure actuelle, nous pensons possible le développement d'une forme de [ToM](#) dans de tels systèmes.

Le problème majeur de ces systèmes de réseaux de neurones classiques concerne leur entraînement. L'inférence d'un modèle complexe s'appuyant sur plusieurs centaines, voir milliers de neurones connectés entre eux (et donc potentiellement dotés de dizaines de milliers de connections selon l'organisation choisie) prend du temps, et surtout des exemples. Un nombre conséquent, voir massif, d'exemples. Il n'est pas rare aujourd'hui d'avoir des systèmes (notamment de *deep-learning*, ou d'apprentissage profond) qui requièrent des centaines de milliers d'exemples pour être entraînés convenablement. Tout repose alors sur le corpus disponible pour permettre au système d'inférer un modèle convenable. Dans le cas de l'apprentissage de l'art de la conversation, ce problème est particulièrement vrai. Est-il nécessaire de trouver des centaines d'exemples de conver-

sation (au moins) pour entraîner un [chatbot](#) pour chacun des sujets de conversations potentiels ? Le propriétaire du [chatbot](#) pourrait il/elle tout de même influencer de manière simple le [chatbot](#) à répondre de la façon dont il/elle le souhaite dans certains cas particuliers (par exemple pour donner des réponses spécifiques à un client tout en bénéficiant des capacités d'inférence et de la flexibilité de ces systèmes) ?

C'est en cela que le principe de [HTM](#) proposé par Hawkins et al. (2019) nous paraît intéressant. Sa capacité à apprendre rapidement de ses expériences avec un risque très faible de détruire d'autres apprentissages précédemment effectués le rend particulièrement attractif dans le cas de l'utilisation pour des [chatbots](#). Ainsi, quand bien même le système aurait appris à interagir d'une façon particulière par inférence, le propriétaire pourrait conserver la main pour lui apprendre des interactions spécifiques à la tâche pour laquelle il serait utilisée, avec un risque faible de détruire l'apprentissage général précédent, car étant dans un contexte différent. Sa faculté de conserver un apprentissage en continu est également particulièrement attractif pour la modélisation de conversations. En effet les systèmes de réseaux de neurones ont besoin d'une phase d'apprentissage longue, et lorsqu'ils sont utilisés avec des utilisateurs ils deviennent alors inertes, figés, pour ne pas risquer de perturber le modèle conversationnel appris.

Il est tout à fait envisageable également que différentes fonctions du modèle conversationnel proposé dans la première partie de cette thèse utilisent différentes technologies. En effet, l'encodeur de phrases et le décodeur de phrases pourraient certainement voir leurs fonctions remplies par des systèmes de [RNN](#) ou des variantes de ce type car beaucoup de travaux existent déjà sur ce type de réseaux, notamment en ce qui concerne le décodage de la grammaire (voir Archambault & Bassano, 1994, par exemple). Le processeur d'information et le modèle du monde et de l'utilisateur pourraient eux voir leurs fonctions remplies par un système [HTM](#), ces deux types de systèmes n'ayant pas d'incompatibilité de principe, rien n'empêche *a priori* de les faire communiquer dans l'attente d'obtenir un système plus unifié.

Nous pensons enfin que de nombreuses tâches de raisonnement pourraient bé-

néficier de l'utilisation plus généralisée de *chatbots* du fait que les participants ne leur attribuent généralement pas les mêmes attentes sociales. En effet les humains sont souvent plus rationnels lorsqu'ils interagissent avec des machines qu'ils ne le sont avec d'autres humains (Baratgin et al., 2020 ; Masson, Baratgin & Jamet, 2017a ; Masson et al., 2015), il est donc tout à fait possible que de nombreux comportements considérés comme des biais soient neutralisés avec des *chatbots* ou des robots. L'utilisation du *TT* associé aux temps de réponse peut aussi devenir un outil généralisable permettant l'évaluation du coût cognitif en conversation. En effet, ce paradigme a l'avantage de donner une tâche à réaliser au participant pendant laquelle ils doivent converser avec un interlocuteur. Si la seule différence entre les deux conversations dans le *TT* est la variable dont on cherche à évaluer le coût cognitif, alors les temps de réponse devraient refléter cette différence.

Nous avons étudié certaines des maximes de Grice (1975) dans la deuxième partie de cette thèse (la première sous-maxime de qualité, les deux sous-maximes de quantité, la maxime de relation et deux des maximes de manière), mais d'autres pourraient encore être explorées (la deuxième sous-maxime de qualité ainsi que les deux autres sous-maximes de manière) afin d'avoir une vision complète du coût cognitif lié à la violation de chacune de ces maximes, ainsi que de leur importance dans la définition de ce qu'est un comportement « humain ». D'autres règles pourraient également être testées selon un principe similaire et dans le même paradigme du *TT*, comme par exemple les maximes de politesse de Leech (2014).

Nous pensons que cette thèse s'insère donc à deux niveaux distincts mais complémentaires dans la recherche en *IA* et en psychologie expérimentale : elle offre un état des lieux et une réponse aux problèmes rencontrés dans un domaine, et ajoute un paradigme expérimental nouveau (bien que le *TT*, lui, ne l'est pas) à l'autre domaine avec des possibilités de répondre à des questions théoriques dans un nouveau cadre qui ne demande qu'à être exploré.

Bibliographie

- Alba Juez, L. (1995). Verbal irony and the Maxims of Grice's cooperative principle. *Revista alicantina de estudios ingleses*, 8, 25-30. <https://doi.org/10.14198/raei.1995.8.02> (cf. p. 86, 118)
- Archambault, D. & Bassano, J.-C. (1994). A neural network for supervised learning of natural language grammar. *Proceedings of the Sixth International Conference on Tools with Artificial Intelligence. TAI 94*, 267-273. <https://doi.org/10.1109/TAI.1994.346481> (cf. p. 57, 176)
- Arora, A., Srivastava, A. & Bansal, S. (2019). Graph and neural network-based intelligent conversation system. *Nature-Inspired Algorithms for Big Data Frameworks* (p. 339-357). IGI Global. <https://doi.org/10.4018/978-1-5225-5852-1.ch014>. (Cf. p. 44)
- Ask, J. A., Facemire, M. & Hogan, A. (2016). The State Of Chatbots : Pilot Chatbots As Part Of Your App+ Mobile Strategy. *Forrester.com report*, 20 (cf. p. 115).
- Attardo, S. (1990). The violation of Grice's maxims in jokes. *Annual Meeting of the Berkeley Linguistics Society*, 16(1), 355-362. <https://doi.org/10.3765/bls.v16i0.1726> (cf. p. 118)
- Bagassi, M. & Macchi, L. (2006). Pragmatic approach to decision making under uncertainty : The case of the disjunction effect. *Thinking & Reasoning*, 12(3), 329-350. <https://doi.org/10.1080/13546780500375663> (cf. p. 14, 79)
- Banerjee, S. & Lavie, A. (2005). METEOR : An automatic metric for MT evaluation with improved correlation with human judgments. *Proceedings of the acl workshop*

on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 65-72 (cf. p. 59).

Baratgin, J. (2002). Is the human mind definitely not Bayesian? A review of the various arguments. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition*, 21, 653-682 (cf. p. 14).

Baratgin, J. (2009). Updating our beliefs about inconsistency : The Monty-Hall case. *Mathematical Social Sciences*, 57(1), 67-95. <https://doi.org/10.1016/j.mathsocsci.2008.08.006> (cf. p. 14)

Baratgin, J., Dubois-Sage, M., Jacquet, B., Stilgenbauer, J.-L. & Jamet, F. (2020). Pragmatics in the false-belief task : let the robot ask the question! *Frontiers in Psychology*, 11, 3234. <https://doi.org/doi.org/10.3389/fpsyg.2020.593807> (cf. p. 67, 153, 177)

Baratgin, J. & Noveck, I. A. (2000). Not only base rates are neglected in the Engineer-Lawyer problem : An investigation of reasoners' underutilization of complementarity. *Memory & cognition*, 28(1), 79-91. <https://doi.org/10.3758/BF03211578> (cf. p. 14, 79)

Baratgin, J. & Politzer, G. (2006). Is the mind Bayesian? The case for agnosticism. *Mind & Society*, 5, 1-38. <https://doi.org/10.1007/s11299-006-0007-1> (cf. p. 14, 79)

Baratgin, J. & Politzer, G. (2007). The psychology of dynamic probability judgment : Order effect, normative theories and experimental methodology. *Mind & Society*, 5, 53-66. <https://doi.org/10.1007/s11299-006-0025-z> (cf. p. 14, 79)

Baratgin, J. & Politzer, G. (2010). Updating : A psychologically basic situation of probability revision. *Thinking & Reasoning*, 16, 253-287. <https://doi.org/80/13546783.2010.504581> (cf. p. 14, 79)

Beunay, B., Jacquet, B. & Baratgin, J. (2022). A Selfish Chatbot Still Does not Win in the Ultimatum Game. In T. Ahram & R. Taiar (Éd.), *Human Interaction, Emerging Technologies and Future Systems V* (p. 255-262). Springer International Publishing. https://doi.org/10.1007/978-3-030-85540-6_33. (Cf. p. 73, 138)

- Benton, M. A. (2016). Gricean quality. *Noûs*, 50(4), 689-703. <https://doi.org/10.1111/nous.12065> (cf. p. 80)
- Berant, J. & Liang, P. (2014). Semantic parsing via paraphrasing. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, 1, 1415-1425. <https://doi.org/10.3115/v1/P14-1133> (cf. p. 75)
- Blackmore, D. & Carston, R. (2005). The pragmatics of sentential coordination with 'and'. *Lingua*, 115, 569-589. <https://doi.org/10.1016/j.lingua.2003.09.016> (cf. p. 85, 86)
- Bless, H., Betsch, T. & Franzen, A. (1998). Framing the framing effect : the impact of context cues on solutions to the 'Asian disease' problem. *European Journal of Social Psychology*, 28(2), 287-291. [https://doi.org/10.1002/\(SICI\)1099-0992\(199803/04\)28:2<287::AID-EJSP861>3.0.CO;2-U](https://doi.org/10.1002/(SICI)1099-0992(199803/04)28:2<287::AID-EJSP861>3.0.CO;2-U) (cf. p. 79)
- Block, N. (1981). Psychologism and behaviorism. *The Philosophical Review*, 90(1), 5-43. <https://doi.org/10.1080/09515080500169470> (cf. p. 64)
- Blount, S. (1995). When Social Outcomes Arent Fair : The Effect of Causal Attributions on Preferences. *Organizational Behavior and Human Decision Processes*, 63(2), 131-144. <https://doi.org/10.1006/obhd.1995.1068> (cf. p. 140, 142)
- Bock, J. K. (1986). Meaning, sound, and syntax : Lexical priming in sentence production. *Journal of Experimental Psychology : Learning, Memory, and Cognition*, 12(4), 575. <https://doi.org/10.1037/0278-7393.12.4.575> (cf. p. 165)
- Bowyer, S. M., Hsieh, L., Moran, J. E., Young, R. A., Manoharan, A., Liao, C.-c. J., Malladi, K., Yu, Y.-J., Chiang, Y.-R. & Tepley, N. (2009). Conversation effects on neural mechanisms underlying reaction time to visual events while viewing a driving scene using MEG. *Brain Research*, 1251, 151-161. <https://doi.org/10.1016/j.brainres.2008.10.001> (cf. p. 90)
- Brandtzaeg, P. B. & Følstad, A. (2017). Why people use chatbots. In K. I. et al. (Éd.), *International Conference on Internet Science* (p. 377-392). Springer, Cham. https://doi.org/10.1007/978-3-319-70284-1_30. (Cf. p. 115)

- Brandtzaeg, P. B. & Følstad, A. (2018). Chatbots : changing user needs and motivations. *Interactions*, 25(5), 38-43. <https://doi.org/10.1145/3236669> (cf. p. 69)
- Bretherton, I., McNew, S. & Beeghly-Smith, M. (1981). Early person knowledge as expressed in gestural and verbal communication : When do infants acquire a “theory of mind”. *Infant social cognition*, 333, 73 (cf. p. 67).
- Broekens, J., Heerink, M., Rosendal, H. et al. (2009). Assistive social robots in elderly care : a review. *Gerontechnology*, 8(2), 94-103. <https://doi.org/10.1016/j.jamda.2010.10.002> (cf. p. 25)
- Carston, R. (1997). Enrichment and loosening : complementary processes in deriving the proposition expressed? *Pragmatik* (p. 103-127). Springer. https://doi.org/10.1007/978-3-663-11116-0_7. (Cf. p. 154)
- Carston, R. (2008). *Thoughts and utterances : The pragmatics of explicit communication*. John Wiley & Sons. (Cf. p. 154).
- Çetinavcı, B. M. (2014). Contextual factors in guessing word meaning from context in a foreign language. *Procedia-Social and Behavioral Sciences*, 116, 2670-2674. <https://doi.org/10.1016/j.sbspro.2014.01.633> (cf. p. 152)
- Chakrabarti, C. & Luger, G. F. (2015). Artificial conversations for customer service chatbots : Architecture, algorithms, and evaluation metrics. *Expert Systems with Applications*, 42(20), 6878-6897. <https://doi.org/10.1016/j.eswa.2015.04.067> (cf. p. 22, 74, 76, 77)
- Chaves, A. P. & Gerosa, M. A. (2020). How Should My Chatbot Interact? A Survey on Social Characteristics in Human–Chatbot Interaction Design. *International Journal of Human–Computer Interaction*, 1-30. <https://doi.org/10.1080/10447318.2020.1841438> (cf. p. 15, 38, 69)
- Chen, D. & Manning, C. (2014). A fast and accurate dependency parser using neural networks. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 740-750. <https://doi.org/10.3115/v1/D14-1082> (cf. p. 75)

- Chen, J., Huang, S. & Luo, R. (2020). Does Net-Speak Experience Interfere With the Processing of Standard Words? Evidence From Net-Speak Word Recognition and Semantic Decisions. *Frontiers in Psychology*, 11, 1932. <https://doi.org/10.3389/fpsyg.2020.01932> (cf. p. 156, 163)
- Ciechanowski, L., Przegalinska, A., Magnuski, M. & Gloor, P. (2019). In the shades of the uncanny valley : An experimental study of human–chatbot interaction. *Future Generation Computer Systems*, 92, 539-548. <https://doi.org/10.1016/j.future.2018.01.055> (cf. p. 120)
- Cohen, L. J. (1971). Some remarks on Grice's views about the logical particles of natural language. In B.-H. Y. (Éd.), *Pragmatics of natural languages* (p. 50-68). Springer. https://doi.org/10.1007/978-94-010-1713-8_3. (Cf. p. 86)
- Coniam, D. (2014). The linguistic accuracy of chatbots : usability from an ESL perspective. *Text & Talk / An Interdisciplinary Journal of Language Discourse Communication Studies*, 34, 545-567. <https://doi.org/10.1515/text-2014-0018> (cf. p. 119)
- Cook, A. E. & Guéraud, S. (2005). What have we been missing? The role of general world knowledge in discourse processing. *Discourse Processes*, 39(2-3), 265-278. <https://doi.org/10.1080/0163853X.2005.9651683> (cf. p. 171)
- Cui, L., Huang, S., Wei, F., Tan, C., Duan, C. & Zhou, M. (2017). Superagent : a customer service chatbot for e-commerce websites. *Proceedings of ACL 2017, System Demonstrations*, 97-102. <https://doi.org/10.18653/v1/P17-4017> (cf. p. 43, 74)
- Culpeper, J. & Terkourafi, M. (2017). Pragmatic approaches (im) politeness. *The Palgrave Handbook of Linguistic (Im) politeness* (p. 11-39). Springer. https://doi.org/10.1057/978-1-137-37508-7_2. (Cf. p. 112)
- Dahiya, M. (2017). A tool of conversation : Chatbot. *International Journal of Computer Sciences and Engineering*, 5(5), 158-161 (cf. p. 38).
- Danziger, E. (2010). On trying and lying : Cultural configurations of Grice's Maxim of Quality. *Intercultural Pragmatics*, 7, 199-219. <https://doi.org/10.1515/iprg.2010.010> (cf. p. 81)

- Deary, I. J. & Der, G. (2005). Reaction time, age, and cognitive ability : Longitudinal findings from age 16 to 63 years in representative population samples. *Aging, Neuropsychology, and cognition*, 12(2), 187-215. <https://doi.org/10.1080/13825580590969235> (cf. p. 90)
- de Gennaro, M., Krumhuber, E. G. & Lucas, G. (2020). Effectiveness of an Empathic Chatbot in Combating Adverse Effects of Social Exclusion on Mood. *Frontiers in Psychology*, 10, 3061. <https://doi.org/10.3389/fpsyg.2019.03061> (cf. p. 61)
- Dessalles, J.-L. (2017). Conversational topic connectedness predicted by Simplicity Theory. *CogSci* (cf. p. 119).
- Devaram, S. (2020). Empathic Chatbot : Emotional Intelligence for Empathic Chatbot : Emotional Intelligence for Mental Health Well-being. *arXiv preprint arXiv :2012.09130*. <https://arxiv.org/abs/2012.09130> (cf. p. 61)
- Dhyani, M. & Kumar, R. (2021). An intelligent Chatbot using deep learning with Bidirectional RNN and attention model. *Materials Today : Proceedings*, 34, 817-824. <https://doi.org/10.1016/j.matpr.2020.05.450> (cf. p. 57)
- DiSalvo, C. F., Gemperle, F., Forlizzi, J. & Kiesler, S. (2002). All robots are not created equal : the design and perception of humanoid robot heads. *Proceedings of the 4th conference on Designing interactive systems : processes, practices, methods, and techniques*, 321-326. <https://doi.org/10.1145/778712.778756> (cf. p. 70)
- Donovan-Kicken, E., Guinn, T. D., Romo, L. K. & Ciceraro, L. D. L. (2013). Thanks for Asking, but Let's Talk About Something Else : Reactions to Topic-Avoidance Messages That Feature Different Interaction Goals. *Communication Research*, 40. <https://doi.org/10.1177/0093650211422537> (cf. p. 116, 118)
- Ducrot, O. (1972). *Dire et ne pas dire : principes de sémantique linguistique*. Hermann. (Cf. p. 14).
- Dufwenberg, M. & Kirchsteiger, G. (2000). Reciprocity and wage undercutting. *European Economic Review*, 44(4), 1069-1078. [https://doi.org/10.1016/S0014-2921\(99\)00047-1](https://doi.org/10.1016/S0014-2921(99)00047-1) (cf. p. 140)

- Dulany, D. E. & Hilton, D. J. (1991). Conversational implicature, conscious representation, and the conjunction fallacy. *Social Cognition*, 9(1), 85-110. <https://doi.org/10.1521/soco.1991.9.1.85> (cf. p. 14)
- Engelhardt, P. E., Bailey, K. G. & Ferreira, F. (2006). Do speakers and listeners observe the Gricean Maxim of Quantity? *Journal of Memory and Language*, 54. <https://doi.org/10.1016/j.jml.2005.12.009> (cf. p. 82, 105, 116, 135)
- Eriksson, M., Marschik, P. B., Tulviste, T., Almgren, M., Pérez Pereira, M., Wehberg, S., Marjanovič-Umek, L., Gayraud, F., Kovacevic, M. & Gallego, C. (2012). Differences between girls and boys in emerging language skills : Evidence from 10 language communities. *British journal of developmental psychology*, 30(2), 326-343. <https://doi.org/10.1111/j.2044-835X.2011.02042.x> (cf. p. 106, 110, 123, 133, 136)
- Fairon, C., Klein, J. R. & Paumier, S. (2006). Le langage SMS : révélateur d'1compétence. *Le français m'a tuer. Actes du colloque "L'orthographe française à l'épreuve du supérieur"*. Louvain-la-Neuve, Presse universitaire de Louvain, 33-42 (cf. p. 17, 155, 171).
- Falk, A., Fehr, E. & Fischbacher, U. (2008). Testing theories of fairness—Intentions matter. *Games and Economic Behavior*, 62(1), 287-303. <https://doi.org/10.1016/j.geb.2007.06.001> (cf. p. 140, 148)
- Faulds, D. J., Mangold, W. G., Raju, P. & Valsalan, S. (2018). The mobile shopping revolution : Redefining the consumer decision process. *Business Horizons*, 61(2), 323-338. <https://doi.org/10.1016/j.bushor.2017.11.012> (cf. p. 38)
- Fishman, P. (1980). Conversational Insecurity. In H. GILES, W. P. ROBINSON & P. M. SMITH (Éd.), *Language* (p. 127-132). Pergamon. <https://doi.org/10.1016/B978-0-08-024696-3.50024-7>. (Cf. p. 106)
- Fitts, P. M. (1966). Cognitive aspects of information processing : III. Set for speed versus accuracy. *Journal of experimental psychology*, 71(6), 849-857. <https://doi.org/10.1037/h0023232> (cf. p. 90)

- Fitzpatrick, K. K., Darcy, A. & Vierhile, M. (2017). Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot) : A Randomized Controlled Trial. *JMIR Ment Health*, 4(2), e19. <https://doi.org/10.2196/mental.7785> (cf. p. 69)
- Forgas, J. P. (1998). On feeling good and getting your way : Mood effects on negotiator cognition and bargaining strategies. *Journal of personality and social psychology*, 74(3), 565. <https://doi.org/10.1037/0022-3514.74.3.565> (cf. p. 141)
- French, R. M. (2000). The Turing Test : the first 50 years. *Trends in Cognitive Sciences*, 4, 0–122. [https://doi.org/10.1016/s1364-6613\(00\)01453-4](https://doi.org/10.1016/s1364-6613(00)01453-4) (cf. p. 121)
- Fukumura, K. & van Gompel, R. P. (2017). How do violations of Gricean maxims affect reading? *Journal of Memory and Language*, 95, 1-18. <https://doi.org/10.1016/j.jml.2017.01.008> (cf. p. 116)
- Gallagher, S. (2020). What in the world : Conversation and things in context. *Minimal Cooperation and Shared Agency* (p. 59-70). Springer. https://doi.org/10.1007/978-3-030-29783-1_4. (Cf. p. 153)
- Gargot, T., Asselborn, T., Zammouri, I., Brunelle, J., Johal, W., Dillenbourg, P., Archambault, D., Chetouani, M., Cohen, D. & Anzalone, S. M. (2021). “It Is Not the Robot Who Learns, It Is Me.” Treating Severe Dysgraphia Using Child–Robot Interaction. *Frontiers in Psychiatry*, 12. <https://doi.org/10.3389/fpsy.2021.596055> (cf. p. 173)
- Garmendia, J. (2015). A (Neo)Gricean Account of Irony : An Answer to Relevance Theory. *International Review of Pragmatics*, 7. <https://doi.org/10.1163/18773109-00701003> (cf. p. 119)
- Ghosh, D., Fabbri, A. R. & Muresan, S. (2018). Sarcasm Analysis Using Conversation Context. *Computational Linguistics*, 44(4), 755-792. https://doi.org/10.1162/coli_a_00336 (cf. p. 153)
- Gockley, R., Bruce, A., Forlizzi, J., Michalowski, M., Mundell, A., Rosenthal, S., Sellner, B., Simmons, R., Snipes, K., Schultz, A. & Wang, J. (2005). Designing robots for long-term social interaction. *2005 IEEE/RSJ International Conference on Intelli-*

- gent Robots and Systems*, 1338-1343. <https://doi.org/10.1109/IROS.2005.1545303>
(cf. p. 22, 26)
- Goodman, K. S. (1967). Reading : A psycholinguistic guessing game. *Journal of the Reading Specialist*, 6(4), 126-135. <https://doi.org/10.1080/19388076709556976>
(cf. p. 46)
- Grice, H. P. (1975). Logic and conversation. *Speech acts* (p. 41-58). Brill. (Cf. p. 2, 4, 14, 16, 17, 30, 40-43, 65, 70, 75, 79, 80, 83-85, 87, 115-118, 141, 153, 177).
- Groen, M. & Noyes, J. (2013). Establishing Goals and Maintaining Coherence in Multiparty Computer-Mediated Communication. *Discourse Processes*, 50(2), 85-106. <https://doi.org/10.1080/0163853X.2012.731968> (cf. p. 112)
- Guéraud, S., Tapiero, I. & O'Brien, E. J. (2008). Context and the activation of predictive inferences. *Psychonomic bulletin & review*, 15(2), 351-356. <https://doi.org/10.3758/PBR.15.2.351> (cf. p. 165, 171)
- Güth, W., Schmittberger, R. & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, 3(4), 367-388. [https://doi.org/10.1016/0167-2681\(82\)90011-7](https://doi.org/10.1016/0167-2681(82)90011-7) (cf. p. 3, 5, 17, 139, 171)
- Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, 1-8. <https://doi.org/10.3115/1073336.1073357> (cf. p. 108)
- Hale, S. (1990). A global developmental trend in cognitive processing speed. *Child development*, 61(3), 653-663. <https://doi.org/10.2307/1130951> (cf. p. 90)
- Ham, J., Bokhorst, R., Cuijpers, R., van der Pol, D. & Cabibihan, J.-J. (2011). Making Robots Persuasive : The Influence of Combining Persuasive Strategies (Gazing and Gestures) by a Storytelling Robot on Its Persuasive Power. In B. Mutlu, C. Bartneck, J. Ham, V. Evers & T. Kanda (Éd.), *Social Robotics* (p. 71-83). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-25504-5_8. (Cf. p. 22, 27)
- Harnad, S. (2001). What's Wrong and Right About Searle's Chinese Room Argument? (M. Bishop & J. Preston, Éd.). <http://cogprints.org/4023/>. (Cf. p. 62)

- Hawkins, J., Lewis, M., Klukas, M., Purdy, S. & Ahmad, S. (2019). A Framework for Intelligence and Cortical Function Based on Grid Cells in the Neocortex. *Frontiers in Neural Circuits*, 12, 121. <https://doi.org/10.3389/fncir.2018.00121> (cf. p. 63, 169, 174, 176)
- Heller, B., Proctor, M., Mah, D., Jewell, L. & Cheung, B. (2005). Freudbot : An investigation of chatbot technology in distance education. *EdMedia+ Innovate Learning*, 3913-3918. <https://www.learntechlib.org/primary/p/20691/> (cf. p. 173)
- Heo, M. & Lee, K. J. (2018). Chatbot as a New Business Communication Tool : The Case of Naver TalkTalk. *Business Communication Research and Practice*, 1(1), 41-45. <https://doi.org/10.22682/bcrp.2018.1.1.41> (cf. p. 38)
- Herring, S. (2013). Relevance in computer-mediated conversation. In. T. V. S. C. Herring D. Stein (Éd.), *Handbook of Pragmatics of Computer-Mediated Communication* (p. 245-268). Mouton de Gruyter. <https://doi.org/10.1515/9783110214468.245>. (Cf. p. 86)
- Hill, J., Ford, W. R. & Farreras, I. G. (2015). Real conversations with artificial intelligence : A comparison between human–human online conversations and human–chatbot conversations. *Computers in Human Behavior*, 49, 245-250. <https://doi.org/10.1016/j.chb.2015.02.026> (cf. p. 69, 122)
- Hilton, D. J. (1995). The social context of reasoning : Conversational inference and rational judgment. *Psychological Bulletin*, 118(2), 248. <https://doi.org/10.1037/0033-2909.118.2.248> (cf. p. 79)
- Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780. <https://doi.org/10.1162/neco.1997.9.8.1735> (cf. p. 57)
- Hoey, M. (2012). *Lexical priming : A new theory of words and language*. Routledge. <https://doi.org/10.1093/ijl/ecl017>. (Cf. p. 165)
- Horn, L. R. (1984). Toward a new Taxonomy for Pragmatic Inference : Q-Based and R-Based Implicature. *Meaning, form, and use in context : Linguistic applications*, 11-42 (cf. p. 83).

- Hung, V., Elvir, M., Gonzalez, A. & DeMara, R. (2009). Towards a method for evaluating naturalness in conversational dialog systems. *Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on*, 1236-1241. <https://doi.org/10.1109/ICSMC.2009.5345904> (cf. p. 77)
- Jacquet, B. & Baratgin, J. (2019). Towards a pragmatic model of an artificial conversational partner : opening the blackbox. *International Conference on Information Systems Architecture and Technology*, 169-178. https://doi.org/10.1007/978-3-030-30604-5_15 (cf. p. 19, 72, 153)
- Jacquet, B. & Baratgin, J. (2020). Mind-Reading Chatbots : We Are Not There Yet. *International Conference on Human Interaction and Emerging Technologies*, 266-271. https://doi.org/10.1007/978-3-030-55307-4_40 (cf. p. 19, 36, 56, 72, 153)
- Jacquet, B., Baratgin, J. & Jamet, F. (2018). The gricean maxims of quantity and of relation in the turing test. *Proceedings of the 11th International Conference on Human System Interaction (HSI)*, 332-338. <https://doi.org/10.1109/HSI.2018.8431328> (cf. p. 21, 42, 43, 52, 71, 89, 91, 104, 115, 121-124, 141)
- Jacquet, B., Baratgin, J. & Jamet, F. (2019). Cooperation in online conversations : the response times as a window into the cognition of language processing (F. M. Bosco, Éd.). *Frontiers in Psychology*, 10, 1-15. <https://doi.org/10.3389/fpsyg.2019.00727> (cf. p. 21, 25, 31, 42, 43, 52, 71, 115, 116, 120-124, 141, 157)
- Jacquet, B., Hullin, A., Baratgin, J. & Jamet, F. (2019a). The impact of the gricean maxims of quality, quantity and manner in chatbots. *2019 international conference on information and digital technologies (idt)*, 180-189. <https://doi.org/10.1109/DT.2019.8813473> (cf. p. 25, 31, 42, 43, 52, 71, 141, 157, 164)
- Jacquet, B., Hullin, A., Baratgin, J. & Jamet, F. (2019b). The Impact of the Gricean Maxims of Quality, Quantity and Manner in Chatbots - Data Repository. osf.io/jkxr4. (Cf. p. 126)

- Jacquet, B., Jamet, F. & Baratgin, J. (2021). On the Pragmatics of the Turing Test. *2021 International Conference on Information and Digital Technologies (IDT)*, 123-130. <https://doi.org/10.1109/IDT52577.2021.9497570> (cf. p. 52)
- Jacquet, B., Jaraud, C., Jamet, F., Guéraud, S. & Baratgin, J. (2021). Contextual Information Helps Understand Messages Written with Textisms. *Applied Sciences*, 11(11). <https://doi.org/10.3390/app11114853> (cf. p. 138, 141)
- Jacquet, B., Masson, O., Jamet, F. & Baratgin, J. (2019). On the Lack of Pragmatic Processing in Artificial Conversational Agents. *IHSED 2018. Advances in Intelligent Systems and Computing*, 876, 394-399. https://doi.org/10.1007/978-3-030-02053-8_60 (cf. p. 19, 38, 75, 89, 119)
- Jaillet, A. (2020). Les ressources éducatives libres pour sauver le monde (de l'éducation) ou repenser l'éducation avec le numérique pour améliorer la qualité? *Distances et médiations des savoirs. Distance and Mediation of Knowledge*. <https://doi.org/10.4000/dms.5523> (cf. p. 173)
- Jain, M., Kumar, P., Kota, R. & Patel, S. N. (2018). Evaluating and informing the design of chatbots. *Proceedings of the 2018 Designing Interactive Systems Conference*, 895-906. <https://doi.org/10.1145/3196709.3196735> (cf. p. 58, 69, 119)
- Jamet, F., Masson, O., Jacquet, B., Stilgenbauer, J.-L. & Baratgin, J. (2018). Learning by Teaching with Humanoid Robot : A New Powerful Experimental Tool to Improve Children's Learning Ability. *Journal of Robotics*, 2018, 1-11. <https://doi.org/10.1155/2018/4578762> (cf. p. 122, 153, 173)
- Jokinen, K. & De Smedt, K. (2012). Computational pragmatics. In J.-O. O. J. Verschueren (Éd.), *Handbook of Pragmatics* (p. 1-39). John Benjamins Publishing Company. (Cf. p. 75).
- Katz, B. & Miledi, R. (1965). The measurement of synaptic delay, and the time course of acetylcholine release at the neuromuscular junction. *Proceeding of the Royal Society of London series B*, 161(985), 483-495. <https://doi.org/10.1098/rspb.1965.0016> (cf. p. 90)

- Kemp, N. (2019). Textese : Language in the online world. *The Oxford Handbook of Cyberpsychology*, 151. <https://doi.org/10.1093/oxfordhb/9780198812746.013.11> (cf. p. 152, 153)
- Kemp, N. & Grace, A. (2017). Txting across time : undergraduates' use of 'textese' in seven consecutive first-year psychology cohorts. *Writing Systems Research*, 9(1), 82-98. <https://doi.org/10.1080/17586801.2017.1285220> (cf. p. 152, 153)
- Keysar, B. & Glucksberg, S. (1992). Metaphor and Communication. *Poetics Today*, 13(4), 633-658. <https://doi.org/10.2307/1773292> (cf. p. 118)
- Kosinski, R. J. (2008). A literature review on reaction time. *Clemson University*, 10. <http://www.cognaction.org/cogs105/readings/clemson.rt.pdf> (cf. p. 123)
- Krupenye, C. & Call, J. (2019). Theory of mind in animals : Current and future directions. *WIREs Cognitive Science*, 10(6), e1503. <https://doi.org/10.1002/wcs.1503> (cf. p. 67)
- Lachman, R., Shaffer, J. P. & Henrikus, D. (1974). Language and cognition : Effects of stimulus codability, name-word frequency, and age of acquisition on lexical reaction time. *Journal of Memory and Language*, 13(6), 613-625. [https://doi.org/10.1016/S0022-5371\(74\)80049-6](https://doi.org/10.1016/S0022-5371(74)80049-6) (cf. p. 90)
- Lasek, M. & Jessa, S. (2013). Chatbots for Customer Service on Hotels' Websites. *Information Systems in Management*, 2(2), 146-158 (cf. p. 115).
- Lassègue, J. (1996). What kind of Turing test did Turing have in mind? *Tekhnema : Journal of Philosophy and Technology*, 3, 37-58 (cf. p. 62).
- Leech, G. N. (2014). *The pragmatics of politeness*. Oxford University Press, USA. (Cf. p. 154, 177).
- Leth-Steensen, C., Elbaz, Z. K. & Douglas, V. I. (2000). Mean response times, variability, and skew in the responding of ADHD children : a response time distributional approach. *Acta Psychologica*, 104, 167-190. [https://doi.org/10.1016/s0001-6918\(00\)00019-6](https://doi.org/10.1016/s0001-6918(00)00019-6) (cf. p. 123)
- Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3), 1126-1177. <https://doi.org/10.1016/j.cognition.2007.05.006> (cf. p. 108)

- Li, C., Du, Y. & Wang, S. (2017). Mining Implicit Intention Using Attention-Based RNN Encoder-Decoder Model. In D.-S. Huang, A. Hussain, K. Han & M. M. Gromiha (Éd.), *Intelligent Computing Methodologies* (p. 413-424). Springer International Publishing. https://doi.org/10.1007/978-3-319-63315-2_36. (Cf. p. 174, 175)
- Ling, R. (2010). Texting as a life phase medium. *Journal of Computer-mediated communication*, 15(2), 277-292. <https://doi.org/10.1111/j.1083-6101.2010.01520.x> (cf. p. 156)
- Liu, C.-W., Lowe, R., Serban, I. V., Noseworthy, M., Charlin, L. & Pineau, J. (2016). How not to evaluate your dialogue system : An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv :1603.08023* (cf. p. 59).
- Lowe, R., Noseworthy, M., Serban, I. V., Angelard-Gontier, N., Bengio, Y. & Pineau, J. (2017). Towards an automatic turing test : Learning to evaluate dialogue responses. *arXiv preprint arXiv :1708.07149* (cf. p. 60).
- Luger, E. & Sellen, A. (2016). “Like Having a Really Bad PA” The Gulf between User Expectation and Experience of Conversational Agents. *Proceedings of the 2016 CHI conference on human factors in computing systems*, 5286-5297. <https://doi.org/10.1145/2858036.2858288> (cf. p. 69)
- Macchi, L. (2000). Partitive formulation of information in probabilistic problems : Beyond heuristics and frequency format explanations. *Organizational behavior and human decision processes*, 82(2), 217-236. <https://doi.org/10.1006/obhd.2000.2895> (cf. p. 14)
- Macchi, L. & Bagassi, M. (2012). Intuitive and analytical processes in insight problem solving : a psycho-rhetorical approach to the study of reasoning. *Mind & Society*, 11(1), 53-67. <https://doi.org/10.1007/s11299-012-0103-3> (cf. p. 14)
- Macchi, L., Caravona, L., Poli, F., Bagassi, M. & Franchella, M. A. (2020). Speak your mind and I will make it right : the case of “selection task”. *Journal of Cognitive Psychology*, 32(1), 93-107. <https://doi.org/10.1080/20445911.2019.1707207> (cf. p. 14)

- Macchi, L., Poli, F., Caravona, L., Vezzoli, M., Franchella, M. A. & Bagassi, M. (2019). How to Get Rid of the Belief Bias : Boosting Analytical Thinking via Pragmatics. *Europe's Journal of Psychology*, 15(3). <https://doi.org/10.5964/ejop.v15i3.1794> (cf. p. 14)
- Marchetti, A., Castelli, I., Harlé, K. M. & Sanfey, A. G. (2011). Expectations and outcome : The role of Proposer features in the Ultimatum Game. *Journal of Economic Psychology*, 32(3), 446-449. <https://doi.org/10.1016/j.joep.2011.03.009> (cf. p. 141)
- Masson, O., Baratgin, J. & Jamet, F. (2017a). NAO robot as experimenter : Social cues emitter and neutralizer to bring new results in experimental psychology. *Proceedings of the International Conference on Information and Digital Technologies, IDT 2017*, 256-264. <https://doi.org/10.1109/DT.2017.8024306> (cf. p. 22, 26, 42, 79, 122, 123, 168, 177)
- Masson, O., Baratgin, J. & Jamet, F. (2015). NAO robot and the “endowment effect”, 1-6. <https://doi.org/10.1109/ARSO.2015.7428203> (cf. p. 42, 122, 171, 177)
- Masson, O., Baratgin, J. & Jamet, F. (2016). NAO robot and the “endowment effect”. *2015 IEEE International Workshop on Advanced Robotics and its Social Impacts (ARSO), 2016-March*. <https://doi.org/10.1109/ARSO.2015.7428203> (cf. p. 22, 26, 79, 168)
- Masson, O., Baratgin, J. & Jamet, F. (2017b). NAO Robot, Transmitter of Social Cues : What Impacts ? In S. Benferhat, K. Tabia & M. Ali (Éd.), *Advances in Artificial Intelligence : From Theory to Practice. IEA/AIE 2017. Lecture Notes in Computer Science* (p. 559-568). Springer, Cham. https://doi.org/10.1007/978-3-319-60042-0_62. (Cf. p. 22, 26, 42, 89, 112, 123)
- Masson, O., Baratgin, J., Jamet, F., Fabien, R. & Filatova, D. (2017). Use a robot to serve experimental psychology : Some examples of methods with children and adults. *2016 International Conference on Information and Digital Technologies (IDT)*, 190-197. <https://doi.org/10.1109/DT.2016.7557172> (cf. p. 79, 123)

- McCausland, S., Kingston, J. & Lyddy, F. (2015). Processing costs when reading short message service shortcuts : an eye-tracking study. *Writing Systems Research*, 7(1), 97-107. <https://doi.org/10.1080/17586801.2014.943150> (cf. p. 156, 163)
- McCornack, S. A. (1992). Information manipulation theory. *Communications Monographs*, 59(1), 1-16. <https://doi.org/10.1080/03637759209376245> (cf. p. 81)
- McDermott, D. (2014). On the claim that a table-lookup program could pass the Turing test. *Minds and Machines*, 24(2), 143-188. <https://doi.org/10.1007/s11023-013-9333-3> (cf. p. 64, 174)
- Medhi Thies, I., Menon, N., Magapu, S., Subramony, M. & O'Neill, J. (2017). How Do You Want Your Chatbot ? An Exploratory Wizard-of-Oz Study with Young, Urban Indians. In R. Bernhaupt, G. Dalvi, A. Joshi, D. K. Balkrishan, J. O'Neill & M. Winckler (Éd.), *Human-Computer Interaction - INTERACT 2017* (p. 441-459). Springer International Publishing. https://doi.org/10.1007/978-3-319-67744-6_28. (Cf. p. 122)
- Mei, H., Bansal, M. & Walter, M. R. (2017). Coherent dialogue with attention-based language models. *Thirty-first AAAI conference on artificial intelligence*. <https://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14164/14210> (cf. p. 44)
- Meira, M. & Canuto, A. (2015). Evaluation of Emotional Agents' Architectures : an Approach Based on Quality Metrics and the Influence of Emotions on Users. *Proceedings of the World Congress on Engineering*, 1 (cf. p. 77).
- Menenti, L., Gierhan, S. M. E., Segaert, K. & Hagoort, P. (2011). Shared Language : Overlap and Segregation of the Neuronal Infrastructure for Speaking and Listening Revealed by Functional MRI [PMID : 21841148]. *Psychological Science*, 22(9), 1173-1182. <https://doi.org/10.1177/0956797611418347> (cf. p. 49)
- Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J. et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014), 176-182. <https://doi.org/10.1126/science.1199644> (cf. p. 152)

- Motley, M. T. (1993). Facial Affect and Verbal Context in Conversation. *Human Communication Research*, 20(1), 3-40. <https://doi.org/10.1111/j.1468-2958.1993.tb00314.x> (cf. p. 26)
- Mulac, A. (1989). Men's and women's talk in same-gender and mixed-gender dyads : Power or polemic?' *Journal of Language and Social Psychology*, 8(3-4), 249-270. <https://doi.org/10.1177/0261927X8983006> (cf. p. 112)
- Mutlu, B., Shiwa, T., Kanda, T., Ishiguro, H. & Hagita, N. (2009). Footing in Human-Robot Conversations : How Robots Might Shape Participant Roles Using Gaze Cues. *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction*, 61-68. <https://doi.org/10.1145/1514095.1514109> (cf. p. 22, 26, 168)
- Nayak, N., Hakkani-Tür, D., Walker, M. A. & Heck, L. P. (2017). To Plan or not to Plan? Discourse Planning in Slot-Value Informed Sequence to Sequence Models for Language Generation. *INTERSPEECH*, 3339-3343. <https://doi.org/10.21437/Interspeech.2017-1525> (cf. p. 43)
- Nordmeyer, A. E. & Frank, M. C. (2015). Negation is only hard to process when it is pragmatically infelicitous. *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, 23-25. http://langcog.stanford.edu/papers_new/nordmeyer-2015-underrev.pdf (cf. p. 108, 123)
- Noveck, I. A. (2001). When children are more logical than adults : experimental investigations of scalar implicature. *Cognition*, 78(2), 165-188. https://doi.org/10.1007/978-3-319-60042-0_62 (cf. p. 79)
- Paas, F., Tuovinen, J. E., Tabbers, H. & Van Gerven, P. W. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational psychologist*, 38(1), 63-71. https://doi.org/10.1207/S15326985EP3801_8 (cf. p. 157)
- Palasundram, K., Sharef, N. M., Nasharuddin, N., Kasmiran, K. & Azman, A. (2019). Sequence to sequence model performance for education chatbot. *International Journal of Emerging Technologies in Learning (iJET)*, 14(24), 56-68 (cf. p. 173).

- Panckhurst, R. (2009). Short Message Service (SMS) : typologie et problématiques futures. (Cf. p. 17, 155, 171).
- Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311-318. <https://doi.org/10.3115/1073083.1073135> (cf. p. 59)
- Paribakht, T. S. & Wesche, M. (1999). Reading and “incidental” L2 vocabulary acquisition : An introspective study of lexical inferencing. *Studies in second language acquisition*, 21(2), 195-224. <https://doi.org/10.1017/S027226319900203X> (cf. p. 152)
- Park, E., Kim, K. J. & Del Pobil, A. P. (2011). The effects of robot’s body gesture and gender in human-robot interaction. *Human-Computer Interaction*, 6, 91-96. <https://doi.org/10.2316/P.2011.747-023> (cf. p. 104, 123)
- Paroubek, P., Chaudiron, S. & Hirschman, L. (2007). Principles of evaluation in natural language processing. *Traitement Automatique des Langues*, 48(1), 7-31. <https://hal.archives-ouvertes.fr/hal-00502700> (cf. p. 77)
- Pasupat, P. & Liang, P. (2015). Compositional semantic parsing on semi-structured tables. *arXiv preprint arXiv :1508.00305* (cf. p. 75).
- Penn, D. C. & Povinelli, D. J. (2007). On the lack of evidence that non-human animals possess anything remotely resembling a ‘theory of mind’. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 362(1480), 731-744. <https://doi.org/10.1098/rstb.2006.2023> (cf. p. 67)
- Perea, M., Acha, J. & Carreiras, M. (2009). Eye movements when reading text messaging (txt msgng). *Quarterly Journal of Experimental Psychology*, 62(8), 1560-1567. <https://doi.org/10.1080/17470210902783653> (cf. p. 156, 163)
- Petit, E. (2009). Émotions et prise de décision dans le jeu de l’ultimatum. *Les Cahiers internationaux de psychologie sociale*, (3), 71-90. <https://doi.org/10.3917/cips.083.0071> (cf. p. 141)

- Pillutla, M. M. & Murnighan, J. (1996). Unfairness, Anger, and Spite : Emotional Rejections of Ultimatum Offers. *Organizational Behavior and Human Decision Processes*, 68(3), 208-224. <https://doi.org/10.1006/obhd.1996.0100> (cf. p. 143)
- Politzer, G. (1986). Laws of language use and formal logic. *Journal of Psycholinguistic Research*, 15(1), 47-92. <https://doi.org/10.1007/BF01067391> (cf. p. 79)
- Politzer, G. (2016). The class inclusion question : a case study in applying pragmatics to the experimental study of cognition. *SpringerPlus*, 5(1), 1133. <https://doi.org/10.1186/s40064-016-2467-z> (cf. p. 79)
- Politzer, G. & Macchi, L. (2000). Reasoning and pragmatics. *Mind & Society*, 1(1), 73-93. <https://doi.org/10.1007/BF02512230> (cf. p. 14)
- Pollard, P. & Evans, J. S. B. T. (1987). Content and Context Effects in Reasoning. *The American Journal of Psychology*, 100(1), 41-60. <https://doi.org/10.2307/1422641> (cf. p. 153)
- Powers, A., Kramer, A. D., Lim, S., Kuo, J., Lee, S.-l. & Kiesler, S. (2005). Eliciting information from people with a gendered humanoid robot. *International Workshop on Robot and Human Interactive Communication (ROMAN)*, 158-163. <https://doi.org/10.1109/ROMAN.2005.1513773> (cf. p. 104, 123)
- Premack, D. & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4), 515-526. <https://doi.org/10.1017/S0140525X00076512> (cf. p. 2, 4, 14, 16, 29, 31, 66, 170, 175)
- Prochaska, J. J., Vogel, E. A., Chieng, A., Kendra, M., Baiocchi, M., Pajarito, S. & Robinson, A. (2021). A Therapeutic Relational Agent for Reducing Problematic Substance Use (Woebot) : Development and Usability Study. *J Med Internet Res*, 23(3), e24850. <https://doi.org/10.2196/24850> (cf. p. 69)
- Ptaszynski, M., Dybala, P., Rzepka, R. & Araki, K. (2010). An Automatic Evaluation Method for Conversational Agents Based on Affect-as-Information Theory. *Journal of Japan Society for Fuzzy Theory and Intelligent Informatics*, 22(1), 73-89. <https://doi.org/10.3156/jsoft.22.73> (cf. p. 77)

- Qiu, M., Li, F.-L., Wang, S., Gao, X., Chen, Y., Zhao, W., Chen, H., Huang, J. & Chu, W. (2017). Alime chat : A sequence to sequence and rerank based chatbot engine. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, 498-503. <https://doi.org/10.18653/v1/P17-2079> (cf. p. 38, 57)
- Qiu, X. & Huang, X. (2015). Convolutional neural tensor network architecture for community-based question answering. *Twenty-Fourth international joint conference on artificial intelligence* (cf. p. 57).
- Rabin, M. (1993). Incorporating Fairness into Game Theory and Economics. *The American Economic Review*, 83(5), 1281-1302. <http://www.jstor.org/stable/2117561> (cf. p. 140)
- Radziwill, N. M. & Benton, M. C. (2017). Evaluating Quality of Chatbots and Intelligent Conversational Agents. (Cf. p. 120).
- Reali, F. & Griffiths, T. L. (2010). Words as alleles : connecting language evolution with Bayesian learners to models of genetic drift. *Proceedings of the Royal Society B : Biological Sciences*, 277(1680), 429-436. <https://doi.org/10.1098/rspb.2009.1513> (cf. p. 152)
- Saygin, A. P. & Cicekli, I. (2002). Pragmatics in human-computer conversations. *Journal of Pragmatics*, 34, 227-258. [https://doi.org/10.1016/S0378-2166\(02\)80001-7](https://doi.org/10.1016/S0378-2166(02)80001-7) (cf. p. 21, 25, 31, 42, 69-71, 77, 81, 83, 84, 87, 88, 105, 108, 115, 120-122, 135, 136)
- Saygin, A. P., Cicekli, I. & Akman, V. (2000). Turing Test : 50 Years Later. *Minds and Machines*, 10, 463-518. <https://doi.org/10.1023/a:1011288000451> (cf. p. 64, 121)
- Schanke, S., Burtch, G. & Ray, G. (2021). Estimating the Impact of “Humanizing” Customer Service Chatbots. *Information Systems Research*, 32(3), 736-751. <https://doi.org/10.1287/isre.2021.1015> (cf. p. 145)
- Searle, J. R. et al. (1980). Minds, brains, and programs. *The Turing Test : Verbal Behaviour as the Hallmark of Intelligence*, 201-224. <https://doi.org/10.1017/S0140525X00005756> (cf. p. 62, 64, 67, 174)

- Shannon, C. E., McCarthy, J. et al. (1956). *Automata studies* (T. 11). Princeton University Press Princeton, NJ. <https://doi.org/10.1090/S0002-9904-1957-10118-9>. (Cf. p. 64)
- Socher, R., Manning, C. D. & Ng, A. Y. (2010). Learning continuous phrase representations and syntactic parsing with recursive neural networks. *Proceedings of the NIPS-2010 Deep Learning and Unsupervised Feature Learning Workshop, 2010*, 1-9 (cf. p. 75).
- Sonnemans, J., Schram, A. & Offerman, T. (1999). Strategic behavior in public good games : when partners drift apart. *Economics Letters*, 62(1), 35-41. [https://doi.org/10.1016/S0165-1765\(98\)00203-1](https://doi.org/10.1016/S0165-1765(98)00203-1) (cf. p. 140)
- Sperber, D., Cara, F. & Girotto, V. (1995). Relevance theory explains the selection task. *Cognition*, 57, 31-95. [https://doi.org/10.1016/0010-0277\(95\)00666-M](https://doi.org/10.1016/0010-0277(95)00666-M) (cf. p. 14, 79)
- Sperber, D. & Wilson, D. (1998). The mapping between the mental and the public lexicon. *Language and thought : Interdisciplinary themes*, 184, 200 (cf. p. 154).
- Sperber, D. & Wilson, D. (2015). Beyond speaker's meaning. *Croatian Journal of Philosophy*, 15(2 (44)), 117-149 (cf. p. 78).
- Spivey, M. J., Tanenhaus, M. K., Eberhard, K. M. & Sedivy, J. C. (2002). Eye movements and spoken language comprehension : Effects of visual context on syntactic ambiguity resolution. *Cognitive Psychology*, 45. [https://doi.org/10.1016/s0010-0285\(02\)00503-0](https://doi.org/10.1016/s0010-0285(02)00503-0) (cf. p. 83)
- Sussman, N. M. & Tyson, D. H. (2000). Sex and power : Gender differences in computer-mediated interactions. *Computers in Human Behavior*, 16(4), 381-394. [https://doi.org/10.1016/S0747-5632\(00\)00020-0](https://doi.org/10.1016/S0747-5632(00)00020-0) (cf. p. 104)
- Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M. & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268, 1632-1634. <https://doi.org/10.1126/science.7777863> (cf. p. 82)
- Tay, B., Jung, Y. & Park, T. (2014). When stereotypes meet robots : the double-edge sword of robot gender and personality in human-robot interaction. *Computers in*

- Human Behavior*, 38, 75-84. <https://doi.org/10.1016/j.chb.2014.05.014> (cf. p. 104, 123)
- Thaler, R. (1980). Toward a positive theory of consumer choice. *Journal of Economic Behavior & Organization*, 1(1), 39-60. [https://doi.org/10.1016/0167-2681\(80\)90051-7](https://doi.org/10.1016/0167-2681(80)90051-7) (cf. p. 171)
- Thomas, N. T. (2016). An e-business chatbot using AIML and LSA. *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2740-2742. <https://doi.org/10.1109/ICACCI.2016.7732476> (cf. p. 38)
- Thorpe, S., Fize, D. & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, 381(6582), 520-522. <https://doi.org/10.1038/381520a0> (cf. p. 90)
- Tisserand, J.-C. (2016). Le jeu de l'ultimatum, une méta-analyse de 30 années de recherches expérimentales. *L'Actualité économique*, 92(1-2), 289-314. <https://doi.org/10.7202/1039879ar> (cf. p. 140)
- Turello, D. (2015). Brain, Mind, and Consciousness : A Conversation with Philosopher John Searle. <https://blogs.loc.gov/kluge/2015/03/conversation-with-john-searle/>. (Cf. p. 64)
- Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59, 433-460 (cf. p. 2, 4, 16, 17, 42, 53, 54, 61, 77, 86, 87, 115, 121, 169).
- Vallverdú, J., Shah, H. & Casacuberta, D. (2012). Chatterbox challenge as a test-bed for synthetic emotions. *Creating Synthetic Emotions through Technological and Robotic Advancements* (p. 118-144). IGI Global. (Cf. p. 70).
- Van der Wel, P. & van Steenbergen, H. (2018). Pupil dilation as an index of effort in cognitive control tasks : A review. *Psychonomic bulletin & review*, 25(6), 2005-2015. <https://doi.org/10.3758/s13423-018-1432-y> (cf. p. 157)
- Venkatesh, A., Khatri, C., Ram, A., Guo, F., Gabriel, R., Nagar, A., Prasad, R., Cheng, M., Hedayatnia, B., Metallinou, A. et al. (2018). On evaluating and comparing open domain dialog systems. *arXiv preprint arXiv :1801.03625* (cf. p. 60).

- von, B. L. (2012). *Théorie générale des systèmes / Ludwig von Bertalanffy ; [préface de Ervin Laszlo ; traduit par Jean-Benoit (i.e. Benoît) Chabrol ; mise à jour bibliographique (1993) de Bernard Paulré]*. Dunod. (Cf. p. 141).
- Wallace, R. S. (2009). The anatomy of ALICE. *Parsing the Turing Test*, 181-210. https://doi.org/10.1007/978-1-4020-6710-5_13 (cf. p. 13, 22, 24, 31, 55, 88, 120)
- Wei, B., Lu, S., Mou, L., Zhou, H., Poupart, P., Li, G. & Jin, Z. (2019). Why do neural dialog systems generate short and meaningless replies? a comparison between dialog and translation. *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7290-7294. <https://doi.org/10.1109/ICASSP.2019.8682634> (cf. p. 58)
- Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), 36-45. <https://doi.org/10.1145/365153.365168> (cf. p. 13, 22, 54, 88, 119)
- Westerman, D., Cross, A. C. & Lindmark, P. G. (2018). I Believe in a Thing Called Bot : Perceptions of the Humanness of “Chatbots”. *Communication Studies*, 1-18. <https://doi.org/10.1080/10510974.2018.1557233> (cf. p. 127)
- Wijaya, D. T. & Yeniterzi, R. (2011). Understanding semantic change of words over centuries. *Proceedings of the 2011 international workshop on DETecting and Exploiting Cultural diversiTy on the social web*, 35-40. <https://doi.org/10.1145/2064448.2064475> (cf. p. 152)
- Wilson, D. (1995). Is there a maxim of truthfulness? *UCL Working Papers in Linguistics*, 7, 197-212 (cf. p. 81).
- Wilson, D. & Kolaiti, P. (2017). Lexical pragmatics and implicit communication. In P. Cap & M. Dynel (Éd.), *Implicitness. From Lexis to Discourse* (p. 147-175). (Cf. p. 154).
- Wilson, D. & Sperber, D. (2002). Relevance theory. (Cf. p. 2, 4, 14, 16, 21, 30, 41-43, 66, 75, 78, 83, 84, 118, 133, 140, 146, 154, 157, 170, 173).
- Wilson, D. & Sperber, D. (2012). *Meaning and Relevance*. Cambridge University Press. (Cf. p. 86).

- Winter, J. (1993). Gender and the political interview in an Australian context. *Journal of Pragmatics*, 20(2), 117-139. [https://doi.org/10.1016/0378-2166\(93\)90079-5](https://doi.org/10.1016/0378-2166(93)90079-5) (cf. p. 106)
- Xu, A., Liu, Z., Guo, Y., Sinha, V. & Akkiraju, R. (2017). A new chatbot for customer service on social media. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 3506-3510. <https://doi.org/10.1145/3025453.3025496> (cf. p. 74)
- Zamora, J. (2017). I'm sorry, dave, i'm afraid i can't do that : Chatbot perception and expectations. *Proceedings of the 5th International Conference on Human Agent Interaction*, 253-260. <https://doi.org/10.1145/3125739.3125766> (cf. p. 69)
- Zhou, A., Jia, M. & Yao, M. (2017). *Business of Bots : How To Grow Your Company Through Conversation*. Topbots Inc. (Cf. p. 24, 38).