Le langage est-il vraiment un système de communication ambigu ?

Alexandre Kabbach

Département de linguistique
Université de Genève
<alexandre.kabbach@unige.ch>

Résumé

Nous proposons de discuter de la question de l'ambiguïté du langage du point de vue du modèle de la communication de Shannon et Weaver (1949). Nous soutenons que le langage n'est pas un *système* de communication ambigu car (1) tout signal linguistique, même ambigu, peut être décodé par un récepteur une fois contextualisé ; et (2) seuls les signaux linguistiques contextualisés forment un objet d'étude pertinent pour statuer sur la nature ambiguë du langage en tant que système de communication, car eux seuls caractérisent les signaux échangés en situation réelle de communication. Nous soutenons que le problème du langage est davantage qu'il n'est pas un système de communication *efficace*, en ce qu'il ne permet pas de s'assurer, au moment du décodage, que le message décodé par le récepteur est parfaitement identique au message encodé par l'émetteur. Nous proposons d'analyser cette inefficacité du langage à la lumière des différences structurelles existant entre les domaines contextuels utilisés par émetteurs et récepteurs pour encoder et décoder les messages transmis. Nous démontrons qu'en dépit de ces différences, les agents possèdent une compétence particulière à *faire sens* des signaux linguistiques auxquels ils sont exposés, c'est-à-dire à enrichir chaque signal linguistique d'un contexte non-linguistique permettant son décodage. Nous proposons de formaliser cette compétence comme fonction de réponse à un stimulus linguistique d'une intelligence artificielle. Nous démontrons qu'une telle formalisation permet de caractériser l'intelligence comme un processus de composition de concepts, où chaque concept peut être modélisé comme une combinaison linéaire d'un nombre fixe de concepts dits *primitifs*. Nous concluons sur les connexions théoriques possibles entre notre formalisation et les travaux sur l'évolution du langage vue comme un mécanisme d'externalisation de la pensée (Reboul, 2017).

Mots clés : langage et communication, ambiguïté, optimisation des systèmes de communication, intelligence artificielle

1. Introduction

L'ambiguïté est définie de manière générale comme la non-bijectivité d'une correspondance entre forme et sens (Piantadosi et al., 2012) — ou plus techniquement entre un ensemble de traits et de labels (Manning et Schütze, 1999). Elle est, nous dit-on, ce qui rend le traitement automatique des langues naturelles si difficile, car pour être capable de comprendre une langue comme le français ou l'anglais, une machine devra résoudre toutes formes d'ambiguïtés linguistiques, qu'elles soient phonétiques, morphologiques, syntaxiques, sémantiques ou même pragmatiques (Jurafsky et Martin, 2014).

Pour certains, l'omniprésence de l'ambiguïté linguistique fait du langage un système de communication relativement mal conçu (Chomsky, 2002) quand pour d'autres elle permet au contraire d'optimiser la communication en réduisant l'entropie globale du langage et en permettant la réutilisation efficiente d'un petit nombre de traits linguistiques (Piantadosi et al., 2012).

Mais qu'est-ce exactement qu'un système de communication *optimisé* ? Et de quelle manière l'ambiguïté potentielle du langage conditionnerait-elle son optimisation, ou non, pour la communication ?

Dans ce travail nous nous proposons d'étudier ces questions dans le cadre théorique du modèle de la communication de Shannon et Weaver (1949). Nous définissons un système de communication comme un système où deux agents — un émetteur et un récepteur — encodent et décodent des messages via des signaux échangés à travers un canal potentiellement bruité. Nous proposons de définir un système de communication optimisé comme étant un système (1) *non-ambigu par défaut*, de sorte qu'en l'absence de bruit, un signal donné ne correspond qu'à un seul et unique message ; (2) *efficace*, de sorte que le message décodé par le récepteur est toujours identique au message encodé par l'émetteur ; (3) *efficient*, en ce que le système minimise son entropie globale c'est-à-dire le nombre de bits d'information nécessaires pour encoder chaque message ; et (4) *robuste au bruit*, de sorte que, même en présence d'un canal bruité, un message émis peut toujours être efficacement décodé par le récepteur.

Dans un premier temps, nous discuterons de la qualification du langage comme *système* de communication ambigu, argumentant de la nécessité pour ce faire d'identifier une correspondance non-bijective entre signaux et messages tels que définis en situation réelle de communication. Nous soutiendrons que, bien que le langage puisse être considéré comme un

système *efficient* de par son utilisation de traits compositionnels non redondants, l'ambiguïté potentielle de ces traits seule ne peut suffire à qualifier l'ensemble du langage de système ambigu, car ces traits pris individuellement ne caractérisent pas l'intégralité des signaux transmis.

Nous argumenterons ensuite que le contexte non-linguistique doit être considéré comme faisant partie intégrante des signaux échangés via le langage, puisque celui-ci est utilisé systématiquement par les agents durant l'encodage et le décodage des messages. Nous démontrerons alors que l'ambiguïté linguistique est un artefact du bruit du canal de communication, où la notion de bruit est définie comme perte de tout ou partie du signal émis. Nous poursuivrons en montrant que les exemples illustrant traditionnellement l'ambiguïté du langage sont intrinsèquement bruités car étudiés hors contexte, et qu'ils sont par conséquent non pertinents pour la caractérisation du langage en tant que système de communication puisque, si bruités, ils ne correspondent pas au cas d'utilisation *par défaut* du système.

Nous montrerons ensuite que le langage est bien un système *non-ambigu par défaut*, car une interprétation stricte de l'ambiguïté du point de vue de la communication implique qu'un signal ambigu ne peut être décodé puisque tous les messages auxquels celui-ci peut correspondre sont strictement équiprobables. Nous argumenterons au contraire que la plupart, sinon la totalité, des signaux linguistiques, peuvent être décodés, mais que le problème réside dans l'impossibilité pour le récepteur de garantir que le message décodé corresponde bien au message encodé par l'émetteur, faisant du langage un système de communication particulièrement *inefficace*. Nous montrerons que cette inefficacité découle de différences structurelles entre les domaines contextuels utilisés par les agents pour encoder et décoder les messages, en ce que ces domaines contextuels recouvrent des notions telles que les *expériences et connaissances préalables du sujet*, qui ne peuvent être considérées comme étant parfaitement identiques entre agents au moment du décodage.

Enfin, nous argumenterons que la non-ambiguïté du langage est la conséquence directe d'une compétence particulière des agents à *faire sens* des signaux linguistiques auxquels ils sont exposés, c'est-à-dire à pouvoir trouver, pour chaque signal linguistique, un contexte potentiel leur permettant de le décoder. Nous proposerons de formaliser cette compétence comme la fonction globale de réponse à un stimulus linguistique d'une intelligence artificielle et nous montrerons qu'une telle formalisation permet de caractériser

l'intelligence comme un processus de composition de concepts où chaque concept peut être modélisé comme une combinaison linéaire d'un nombre fixe de concepts dits *primitifs*, via une fonction unique de composition.

2. Communication et optimisation

2.1. Le modèle de la communication de Shannon et Weaver

Dans ce travail nous assimilons le langage à un système de communication suivant le modèle de Shannon et Weaver (voir Figure 1 empruntée à Reboul (2017)), où deux agents—un émetteur et un récepteur—encodent et décodent des messages vers et depuis des signaux transitant par un canal de transmission potentiellement bruité.

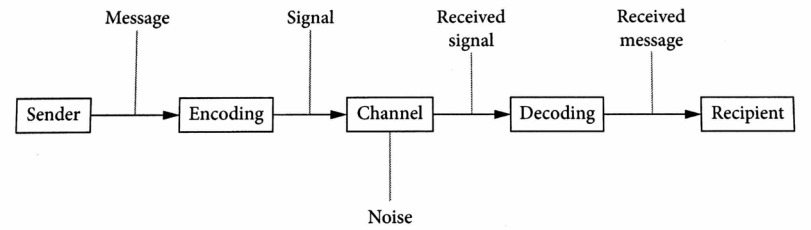

Figure 1 : Le modèle de la communication de Shannon et Weaver (1949)

Nous définissons la notion de *bruit* comme toute forme d'interférence (modification, perte, etc.) appliquée au canal de transmission, conduisant un signal reçu à ne pas être totalement identique au signal émis. Dans la section 4, nous nous focaliserons plus particulièrement sur la notion de bruit définie comme perte de tout ou partie du signal émis, perte conduisant spécifiquement un signal linguistique reçu à être considéré comme ambigu.

2.2. Qu'est-ce qu'un système de communication optimisé ?

Nous proposons de définir un système de communication optimisé comme étant un système :

1. *non-ambigu par défaut*, de sorte qu'en l'absence de bruit, un signal donné ne correspond qu'à un seul et unique message ;
2. *efficace*, de sorte que le message décodé par le récepteur est toujours identique au message encodé par l'émetteur ;
3. *efficient*, en ce que le système minimise son entropie globale c'est-

à-dire le nombre de bits d'informations nécessaires pour encoder chaque message ; et

4. *robuste au bruit*, de sorte que, même en présence d'un canal bruité, un message émis peut toujours être efficacement décodé par le récepteur.

Dans ce travail nous nous appuierons sur la position de Piantadosi et al. (2012) présentée plus en détails dans la section 3.1, en considérant que la nature compositionnelle du langage—caractérisée par une réutilisation de traits minimisant les redondances entre signal et contexte—fait du langage un système de communication *efficient*. Nous questionnerons en revanche la caractérisation du langage comme système ambigu à la section 3.2, en argumentant que les travaux précédents sur l'ambiguïté linguistique et la communication (Wasow et al., 2005 ; Piantadosi et al., 2012) reposent sur des exemples artificiellement ambigus car délibérément bruités. Nous démontrerons à la section 5 que le langage doit être considéré comme un système de communication *non ambigu par défaut* mais néanmoins *inefficace*, étant donné que les agents s'efforcent de décoder les signaux linguistiques auxquels ils sont confrontés, bien qu'en faisant usage de domaines contextuels différents.

3. Ambiguïté et efficience du langage

3.1. L'argument de l'efficience de la communication linguistique

Les travaux antérieurs sur les notions d'ambiguïté et de communication linguistique (Wasow et al., 2005 ; Piantadosi et al., 2012) s'appuient sur une caractérisation large du phénomène d'ambiguïté, compris comme correspondance non-bijective entre *forme* et *sens* où le concept de sens recouvre un champ relativement vaste de phénomènes linguistiques, allant de la phonétique à la morphologie en passant par la syntaxe, la sémantique ou même la pragmatique. Dans ces travaux, le mot *run*[1] est par exemple considéré comme lexicalement ambigu car possédant plusieurs sens distincts comme illustrés dans les exemples *run a company* ou *run a marathon*. Le même mot est aussi décrit comme ambigu concernant sa catégorie morphosyntaxique, puisqu'il peut faire référence à un *nom* comme dans l'exemple *we went*

1 Dans la suite de ces travaux, les exemples seront en anglais lorsqu'ils sont issus de la littérature scientifique anglophone discutée et en français sinon.

for a run, ou à un *verbe* comme dans les exemples cités plus haut. Ces mêmes travaux illustrent également le large spectre de l'ambiguïté en mentionnant le cas du phonème anglais [tuː] pouvant être considéré comme ambigu concernant sa transcription en morphème (*to*, *two* ou *too*), ou encore celui du morphème *-s* de *runs* concernant son interprétation comme marqueur du pluriel ou comme conjugaison du présent de l'indicatif.

Pour Piantadosi et al. (2012), le phénomène d'ambiguïté linguistique tel que décrit plus haut optimise la communication en ce qu'il : (a) minimise l'entropie globale du système en permettant d'éviter les redondances avec les informations déjà présentes dans le contexte ; et (b) permet la réutilisation d'éléments linguistiques jugés *simples*, tels que les mots courts et/ou fréquents. Selon les auteurs, l'ambiguïté résulte d'un compromis entre les désirs communicationnels contradictoires de l'émetteur et du récepteur : l'émetteur voudrait un signal unique pour encoder tous les messages quand le destinataire voudrait un signal distinct par message. Inspirée du *principe du moindre effort* de Zipf (1949), leur caractérisation de l'optimisation d'un système de communication est articulée autour des notions de *clarté* et de *facilité* (*clarity and ease*). Un système de communication *clair* permet de récupérer le message destiné avec une haute probabilité, alors qu'un système *facile* permet aux signaux d'être produits et communiqués de manière efficiente. Comparativement aux notions présentées précédemment à la section 2.2, la notion de *clarté* correspond à notre notion *d'efficacité*, tandis que la notion de *facilité* correspond à notre notion *d'efficience*. Piantadosi et al. (2012) ne mentionnent en revanche pas explicitement la notion de *robustesse*, bien que celle-ci soit présente de manière implicite dans leur caractérisation probabiliste de l'ambiguïté. De plus, ils considèrent le langage comme étant un système efficace bien qu'ambigu, alors que notre position est inverse. Comme nous le verrons à la section 5, la notion d'efficacité découle plus naturellement de la notion de non-ambiguïté d'un point de vue mathématique, puisque l'efficacité requiert de manière formelle l'inversibilité des fonctions d'encodage et de décodage, et que cette condition n'est satisfaite que si les deux fonctions sont bijectives.

3.2. L'importance du périmètre de l'ambiguïté

L'efficience du langage comme système de communication repose sur l'utilisation de traits compositionnels en correspondance non-bijective avec un ensemble de labels. Pourtant, l'existence d'une correspondance non-bijec-

tive entre traits et labels ne suffit pas à elle seule à caractériser un système de communication comme étant ambigu, tant que la non-bijectivité des traits ne couvre pas l'entièreté du périmètre des signaux transmis.

Pour illustrer ce point, considérons tout d'abord l'exemple des panneaux de signalisation belges. Considérons, suivant Szabó (2017), qu'un système de signalisation de ce type peut être décrit comme un système compositionnel où chaque panneau de signalisation est défini par une composition particulière de traits significatifs — tels que la *forme*, le *schéma de couleur* ou le *symbole* — significatifs dans la mesure où chaque trait peut être décrit comme ayant une contribution identifiable au sens global du panneau (par exemple, une *forme triangulaire* est généralement associée à la notion de *mise en garde*, un *fond de couleur rouge* à la notion *d'interdiction* et le *symbole d'une flèche* aux *directions*). Considérons maintenant ce système suivant le modèle de la communication de Shannon et Weaver (1949), où chaque panneau est un signal encodant un message spécifique — tel qu'une *obligation*, une *indication* ou une *interdiction*. Nous considérons tout d'abord qu'un tel système de communication est *délibérément conçu pour être non-ambigu*, puisque son objectif est de pouvoir communiquer via chaque panneau une instruction claire et unique. Pourtant, d'aucuns pourraient considérer que le trait *fond de couleur bleu* est ambigu, puisqu'il peut indiquer soit la *nature* de l'instruction encodée par le panneau (une *obligation* ou une *indication* ici), soit le *domaine d'application* du panneau (ayant trait au *stationnement*), comme illustré à la Figure 2. Toutefois, la présence de cette "ambiguïté" ne remet pas en cause la non-ambiguïté du *système* de signalisation dans son ensemble, du moment que le trait *fond de couleur bleu* peut être facilement désambiguïsé lorsque contextualisé, c'est-à-dire lorsque combiné avec d'autres traits au sein d'un panneau de signalisation complet.

Figure 2 : Comparaison de la sémantique du trait *fond de couleur bleu* dans le système de panneaux de signalisation belges

Les conclusions précédentes demeurent valables lorsque la notion de *contexte* est étendue à un ensemble de traits distincts des traits composition-

nels formant le cœur des signaux transmis. Prenons pour exemple le système de numération indo-arabe utilisé aujourd'hui dans la plupart des sociétés occidentales. Ce système de numération dit *positionnel* repose sur un ensemble de dix chiffres (de 0 à 9) et encode chaque nombre comme une séquence desdits chiffres. Considérons maintenant que chaque nombre puisse être caractérisé par un ensemble de symboles (les chiffres) et un contexte, défini comme un ensemble de positions de symboles dans une séquence. Le nombre 123 peut dès lors être défini comme l'ensemble de symboles {1, 2, 3} combiné au contexte {*premier*, *deuxième*, *troisième*}, puisqu'il est formé par la séquence comprenant le chiffre 1 en première position, le chiffre 2 en deuxième, etc. Ainsi défini, le système de numération indo-arabe peut être considéré comme étant plus efficient que, p. ex., le système de numération romain, puisqu'il repose sur un plus petit nombre de symboles et possède donc une entropie symbolique plus faible. Pourtant, le fait que l'ensemble des symboles {1, 2, 3} soit ambigu vis-à-vis du nombre auquel il fait référence *hors contexte* (123, 321, 213, etc.) ne suffit pas à qualifier l'intégralité du système de numération comme ambigu, puisque le contexte précédemment défini doit être considéré comme faisant partie intégrante du signal transmis.

Les deux systèmes compositionnels présentés ci-dessus peuvent être considérés comme des systèmes *efficients* étant donné qu'ils font tous deux usage d'un nombre limité de traits compositionnels, réutilisés avec un minimum de redondances au sein des signaux transmis. En tant que tels, ils constituent des exemples pertinents permettant d'illustrer les thèses de Piantadosi et al. (2012), puisque la réutilisation des traits et la minimisation de l'entropie sont rendues possibles par la correspondance non-bijective entre certains traits et certains labels. Pourtant, ces systèmes demeurent *non-ambigus par conception*, et les phénomènes d'ambiguïté observés ne caractérisent jamais l'ensemble du périmètre des signaux complets et peuvent même être considérés comme artificiels puisque générés par une altération délibérée des signaux complets.

4. Ambiguïté du langage et nature des signaux complets

4.1. L'ambiguïté comme artefact du bruit

Afin d'étendre les considérations précédentes à la question de l'ambiguïté linguistique, revenons tout d'abord aux exemples présentés à la section 3.1. Nous avons mentionné précédemment que, selon la définition usuelle de

l'ambiguïté, le mot *run* était considéré comme ambigu du point de vue de son sens et de sa catégorie morphosyntaxique. Concernant l'ambiguïté lexicale, nous avons notamment cité pour exemples *run a company* et *run a marathon*. Ce que ces exemples illustrent en revanche, c'est qu'un contexte lexical minimal — tel que l'ensemble des mots au voisinage direct du verbe comme (*company*, *marathon*) — peut suffire à désambiguïser le verbe *run*. Il en va de même de l'ambiguïté morphosyntaxique, où un contexte combinant à la fois item lexical et catégorie morphosyntaxique des éléments précédant le mot *run* dans une séquence suffit bien souvent à différencier la catégorie *nom* de la catégorie *verbe*, comme dans {*a*|DET, *run*|N} et {*they*|PRO, *run*|V}, où *DET* indique le déterminant, *N* le nom, *PRO* le pronom et *V* le verbe. Ces quelques considérations forment la base des modèles computationnels pour l'étiquetage morphosyntaxique ou la désambiguïsation des mots en contexte en traitement automatique des langues naturelles (Manning et Schütze, 1999 ; Jurafsky et Martin, 2014). Si ces considérations n'éludent pas, bien entendu, les défis posés par l'ambiguïté linguistique pour le traitement automatique des langues, elles questionnent en revanche la pertinence de considérer les éléments linguistiques hors contexte — comme un mot isolé du reste de la phrase — pour statuer sur le caractère ambigu du langage en tant que *système* de communication.

Quand bien même la segmentation des signaux linguistiques reposerait sur un périmètre prédéfini comme la *phrase* ou *l'énoncé*, ce périmètre n'en demeurerait pas moins relativement arbitraire au regard d'une situation de communication réelle. Les cas d'ambiguïté résultant d'une isolation des signaux après segmentation pourraient donc être considérés comme artificiels, comme dans l'exemple suivant :

(1) Il écrit assez petit […] je n'arrive pas bien à voir sa figure.

Si l'on considère en effet chaque séquence autour de la pause […] comme un énoncé complet, la segmentation des deux énoncés conduira à générer de l'ambiguïté de manière artificielle sur le sens du mot figure (qui doit être comprise ici comme *figure géométrique*), puisque le contexte utilisé dans une conversation naturelle pour désambiguïser le sens (ex. *écrit*) aura été retiré délibérément du signal.

Outre la segmentation, la méthode de transcription du langage peut aussi introduire de l'ambiguïté de manière artificielle, en altérant des signaux linguistiques autrement complets. Une transcription textuelle sans

marqueurs prosodiques pourra par exemple être considérée comme bruitée, puisque ces marqueurs sont naturellement présents dans le discours et qu'ils permettent de résoudre certains types d'ambiguïtés syntaxiques (Carlson, 2009).

Outre les cas d'ambiguïté artificielle résultant d'une altération délibérée du signal linguistique, il est intéressant de reconsidérer les cas d'ambiguïté sémantique présentés dans la littérature scientifique, comme l'exemple (2), car ceux-ci illustrent en effet l'importance du contexte non-linguistique dans toute communication humaine.

(2) I had a book stolen

Selon Chomsky (1965) cet énoncé offre trois interprétations possibles : (i) *Someone stole a book from my car* (ii) *I had someone steal a book* et (iii) *I had almost succeeded in stealing a book*. Pourtant, aucune de ces interprétations n'impliquerait exactement le même contexte non-linguistique, que l'on caractérise ce contexte en terme de cognition située (Barsalou, 2008), de connaissances préalables potentiellement partagées entre les agents (Morris, 1995), ou d'intentions de communication (Grice, 1969). Savoir par exemple si l'émetteur de (2) est connu pour être un voleur de livres, si la situation d'énonciation se déroule dans une bibliothèque, si l'émetteur pointe du doigt une pile de livres sur l'étagère d'une librairie, ou s'il est en train de déménager, sont autant d'éléments d'information utiles qui, si portés à la connaissance du récepteur, pourront être utilisés pour décoder le message et pourront donc influer sur la nature de ce dernier.

L'intervention du contexte non-linguistique dans le processus de décodage semble indéniable. En revanche, *dans quelle mesure* celui-ci intervient dans ce processus demeure une question ouverte. Dans la suite de ce travail, nous considérerons que le contexte non-linguistique n'est pas négligeable, de sorte qu'un signal échangé via le langage ne peut être approximé par le seul signal linguistique mais doit être considéré comme la combinaison d'un signal linguistique et d'un contexte non-linguistique. La question que nous nous posons dès lors est de savoir si un tel contexte est identique pour tous les agents, de sorte à ce que l'on puisse garantir que les messages encodés et décodés soient parfaitement égaux en tout temps.

4.2. Le contexte non-linguistique est-il spécifique à l'agent ?

Supposons qu'un individu donné soit engagé dans une mission écologique en mer et reçoive de la part d'un membre de l'équipe l'instruction suivante : *Comptez tous les mammifères que vous observez*. Supposons que cet individu ne soit pas très au point concernant la biologie marine et ne sache pas que les dauphins sont des mammifères, contrairement à l'émetteur du message. Lorsque les membres de l'équipe se réuniront pour comparer leurs décomptes, il deviendra probablement évident que le récepteur aura mal décodé le message, et que cette erreur de décodage provient d'une différence de connaissances préalables entre émetteur et récepteur concernant la notion de mammifère marin. Cet exemple nous semble pertinent pour deux raisons : (i) il démontre que si l'on considère que la notion de contexte non-linguistique englobe des concepts tels que les croyances, les expériences ou les connaissances préalables du sujet, il n'est pas possible de considérer le contexte comme étant indépendant des agents, puisqu'il n'y a aucune raison de présupposer que deux agents distincts puissent partager des croyances, des expériences ou des connaissances parfaitement identiques à tout moment ; et (ii) il montre également qu'il n'existe pas de mécanisme d'évaluation, au moment du décodage, permettant de garantir l'alignement entre message décodé et message encodé. En d'autres termes, sans retour de l'émetteur, le récepteur d'un signal ne peut s'assurer que le message qu'il décode est *correct* en ce qu'il est identique au message de l'émetteur. Cela suggère donc que le fonctionnement *par défaut* de la communication linguistique est de décoder les messages sans considérations pour leur exactitude.

Bien entendu, ces différences entre messages encodés et décodés se révèlent plus facilement lorsqu'elles conduisent à un désaccord ou un malentendu explicité entre les agents. Mais il se pourrait également que la nature du langage soit telle qu'elle tolère des différences entre messages encodés et décodés sans pour autant entraver la communication (Sperber et Wilson, 1995 ; Mercier et Sperber, 2017). Un émetteur peut dire *il y a un chat derrière la porte*, et être compris des récepteurs sans pour autant que tous les agents partagent la même représentation *a priori* de la *taille* du chat ou de sa *couleur*. L'accord entre les agents n'implique pas non plus l'identité des messages, puisque deux agents peuvent s'accorder sur la vérité de la proposition *les chatons sont mignons* sans avoir à partager les *raisons* conditionnant leurs jugements respectifs, ou même *jusqu'à quel âge* chacun d'entre eux considère qu'un chat peut être considéré comme un chaton. Un message peut donc

demeurer *vague* au regard de son contenu propositionnel, ou de ce qu'il dénote, et des exemples parfois considérés comme ambigus tels que *nothing is better than your cooking* (Wasow et al., 2005) ou sous-déterminés comme *pas de ça ici* demeurent décodables et décodés par les récepteurs, puisque le décodage d'un message n'implique pas que celui-ci ait un contenu propositionnel unique ou même un périmètre de dénotation clair.

4.3. Une interprétation stricte de l'ambiguïté du point de vue du décodage

Clarifions maintenant le lien entre ambiguïté et décodage au sein du modèle de la communication de Shannon et Weaver. Nous avons vu précédemment que l'ambiguïté est définie de manière usuelle comme une correspondance non bijective entre traits et labels. Pourtant, une interprétation stricte de cette non-bijectivité implique qu'aucun label unique ne peut être choisi parmi toutes les possibilités, puisque même dans un modèle probabiliste, et sans biais d'initialisation, toutes les correspondances trait-label doivent être considérées comme étant équiprobables. Pour le langage, cela implique que t*out signal ambigu ne peut être décodé*, puisqu'aucun message unique ne peut être assigné au signal correspondant — une question totalement différente de celle de savoir si le message décodé correspond bien au message encodé.

Un point important est souligné par les recherches sur l'ambiguïté et le langage : l'ambiguïté sémantique telle qu'illustrée à la section 4.1 est bien souvent transparente pour les agents qui ne la découvre que lorsque celle-ci leur est explicitée (Chomsky, 1965 ; Wasow et al., 2005). Cela suggère en réalité qu'une telle ambiguïté n'entrave pas le décodage *d'au moins un message* puisque les agents peuvent dans leur grande majorité trouver *une* interprétation d'un énoncé ambigu, aussi vague soit-elle.

Dans la suite de ce travail nous nous appuierons sur ces constatations en argumentant que les cas d'ambiguïté *stricte* sont en réalité excessivement rares en ce que les agents s'efforcent de *faire sens* des signaux linguistiques auxquels ils sont confrontés. Nous ferons l'hypothèse que les agents possèdent une compétence particulière pour déterminer, pour chaque signal linguistique, un contexte non-linguistique correspondant, rendant le signal décodable et donc significatif. Cette intuition est appuyée par des travaux tels que ceux de Chao (1997) tentant par exemple de fournir un

contexte significatif à l'exemple des *colorless green ideas* de Chomsky (1957)[2] ou encore certains résultats expérimentaux en linguistique computationnelle (Bell et Schäfer, 2013 ; Vecchi et al., 2017).

5. Ambiguïté et efficacité du langage

5.1. Caractérisation formelle des notions d'ambiguïté et d'efficacité linguistique

Soit M l'ensemble des messages encodables et S l'ensemble des signaux complets et donc décodables. Pour simplifier, nous considérerons dans ce qui suit que (i) tout message peut être encodé par au moins un signal complet ; et (ii) tout signal complet peut être décodé par au moins un message. Nous dirons d'un système de communication qu'il est *non-ambigu par défaut* si et seulement si, en l'absence de bruit, les protocoles d'encodage et de décodage des signaux peuvent être modélisés par deux fonctions $\varepsilon: \begin{cases} M \longrightarrow S \\ x \longmapsto \varepsilon(x) \end{cases}$ et $\delta: \begin{cases} S \longrightarrow M \\ x \longmapsto \delta(x) \end{cases}$ telles que :

1. La fonction d'encodage de M vers S est bijective :

$$\forall s \in S \exists ! m \in M : s = \varepsilon(m)$$

2. La fonction de décodage de S vers M est bijective :

$$\forall m \in M \exists ! s \in S : m = \delta(s)$$

En outre, nous dirons d'un système de communication qu'il est efficace si et seulement si : $\forall m \in M : \delta(\varepsilon(m)) = m$

Formellement, la non-ambiguïté par défaut est une condition nécessaire mais non suffisante pour garantir l'efficacité, puisqu'un système non-ambigu peut ne pas être efficace comme illustré à la Figure 3, mais qu'un système efficace est nécessairement non-ambigu par défaut.

2 Voir aussi la compétition de Stanford : [https://www.linguistlist.org/issues/2/2-457.html#2]

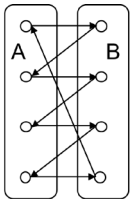

Figure 3 : Correspondances non-ambiguës mais inefficaces entre deux ensembles A et B : les correspondances de A vers B et de B vers A sont toutes deux bijectives, mais ne sont pas l'inverse l'une de l'autre.

5.2. Le langage comme système de communication non-ambigu et inefficace

À la section 4.3 nous avons défendu une interprétation stricte de la notion d'ambiguïté du point de vue du décodage, de sorte qu'un signal ne puisse être considéré comme ambigu que s'il ne peut être décodé par le récepteur. Nous considérons donc dès lors qu'encodage et décodage opèrent sur l'ensemble des signaux non-ambigus, de sorte que les applications et peuvent être modélisées par des fonctions bijectives de l'ensemble des signaux S vers l'ensemble des messages M. Nous concluons donc que le langage est non-ambigu par défaut.

En ce qui concerne l'efficacité, nous avons vu à la section 4.1 qu'un signal échangé via le langage devait être considéré comme la combinaison d'un signal linguistique et d'un contexte non-linguistique. Étant donné L l'ensemble des signaux linguistiques et C l'ensemble des contextes, les fonctions d'encodage et de décodage deviennent :

$$\varepsilon: \begin{cases} M \to S = L \times C \\ x \mapsto \varepsilon(x) \end{cases} \text{ et } \delta: \begin{cases} S = L \times C \to M \\ x \mapsto \delta(x) \end{cases}$$

Considérant maintenant, comme nous l'avons défendu à la section 4.2, que chaque contexte est spécifique à l'agent, nous affinons nos définitions des fonctions d'encodage et de décodage pour un agent i : ε_i et δ_i, en intégrant la notion d'ensemble contextuel spécifique à l'agent C_i. Étant donné deux agents i et j, les fonctions d'encodage et de décodage respectives des agents sont donc définies sur deux domaines distincts $S_i = L \times C_i$ et $S_j = L \times C_j$ avec $S_i \neq S_j$. Les domaines d'application des fonctions étant

distincts, on ne peut pas démontrer l'efficacité du système tel que défini à la section 5.1. Le langage est donc *inefficace*.

6. Conséquences pour les domaines de l'intelligence artificielle et du traitement automatique des langues naturelles

Comme nous l'avons vu à la section 4.3, le langage peut être considéré comme un système de communication non-ambigu par défaut étant donné que les agents possèdent une compétence particulière à *faire sens* des signaux linguistiques auxquels ils sont exposés. Nous proposons dans cette section de discuter des bénéfices de la formalisation d'une telle compétence comme fonction de réponse à un stimulus linguistique d'une intelligence artificielle.

Considérons une intelligence artificielle comme une fonction définie sur l'ensemble des signaux linguistiques L, des contextes non-linguistiques C et des messages M, tous trois modélisés par des espaces vectoriels, suivant les pratiques du domaine (Lazaridou et al., 2016). Le décodage d'un signal linguistique $l \in L$ est alors défini pour la machine comme le fait de trouver un élément $c \in C$ tel qu'il existe un élément $m \in M$ satisfaisant la condition $(l, c) = m$, où (l, c) désigne la concaténation des vecteurs l et c. Si l'on considère maintenant l'ensemble des messages M comme étant assimilable à la notion *d'espace conceptuel* formé de concepts artificiels pré-existants conditionnant potentiellement la génération d'éléments de C, s'ensuit: (i) *l'intelligence* et plus particulièrement *l'acte de penser* peut être formalisé pour une machine comme la combinaison linéaire de vecteurs préexistants de M conduisant à la formation d'un nouveau vecteur dans M ; (ii) l'espace conceptuel M peut être caractérisé par une loi de composition interne unique notée + et une *base* formée d'un nombre fini de vecteurs b_k tels que chaque élément de M puisse être défini comme une combinaison linéaire finie des vecteurs b_k avec +. Conceptuellement, une telle formalisation revient à considérer que l'espace conceptuel est productif, formé d'un ensemble fini de *primitifs* (Fodor, 1975 ; Fodor, 2008) et d'une unique opération de composition (Chomsky, 1995) ce qui serait, du moins dans l'esprit, compatible avec la formalisation de l'espace conceptuel tel que pensée par Reboul (2017) ; et (iii) *la raison précède la communication*, de sorte qu'une machine ne peut correctement traiter le langage humain que si elle possède au préalable un système conceptuel bien défini, productif et compositionnel.

7. Conclusion

Dans ce travail, nous avons démontré que le langage n'était pas un *système* de communication ambigu, mais qu'il demeurait néanmoins un système de communication hautement inefficace et donc non-optimisé. Nous avons en outre montré comment les considérations sur l'inefficacité du langage pouvaient fournir une contribution utile au domaine de l'intelligence artificielle, en théorisant une fonction de réponse à un stimulus linguistique basée sur la compétence particulière des agents à contextualiser les signaux linguistiques auxquels ils sont confrontés. En revanche, nous avons dit peu de choses sur les conséquences de ces considérations sur la question de l'évolution du langage pour la communication. Bien que dépassant très largement le cadre des présents travaux, nous voyons notre approche comme totalement compatible avec celle de (Reboul, 2017) caractérisant le langage comme un système de communication au sens faible ayant évolué non pas pour la communication mais pour l'externalisation de la pensée. Nous voyons en effet la précédence d'un espace conceptuel productif et compositionnel sur sa transmission, ainsi que l'absence de mécanisme de validation des messages décodés, comme des contraintes structurelles immuables découlant directement de la nature même du langage et de son évolution et comme des prérequis à toute formalisation d'une intelligence artificielle.

Bibliographie

Barsalou, W. Lawrence. 2008. Grounded Cognition. *Annual Review of Psychology* 59(1) : 617–645.
DOI : [https://doi.org/10.1146/annurev.psych.59.103006.093639]

Bell, J. Melanie & Martin Schäfer. 2013. Semantic transparency: challenges for distributional semantics. In *Proceedings of the IWCS 2013 Workshop Towards a Formal Distributional Semantics*, 1–10. Association for Computational Linguistics.

Carlson, Katy. 2009. How Prosody Influences Sentence Comprehension. *Language and Linguistics Compass* 3(5) : 1188–1200.
DOI : [https://doi.org/10.1111/j.1749-818X.2009.00150.x]

Chao, Yuen Ren. 1997. Making Sense Out of Nonsense. *The Sesquipedalian* 7(32) : 1996–97.

Chomsky, Noam. 1957. *Syntactic Structures*. The Hague : Mouton & Co.

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge, USA : MIT Press.
DOI : [https://doi.org/10.21236/AD0616323]

Chomsky, Noam. 1995. *The Minimalist Program*. Cambridge, USA : MIT Press.

Chomsky, Noam. 2002. An interview on minimalism. In Adriana Belletti & Luigi Rizzi (eds.), *On Nature and Language*, 92–161. Cambridge : Cambridge University Press. DOI : [https://doi.org/10.1017/CBO9780511613876.005]

Fodor, A.Jerry. 1975. *The Language of Thought*. New York : Thomas Y. Crowell.

Fodor, A. Jerry. 2008. *LOT 2: The Language of Thought Revisited*. Oxford : Clarendon Press.

Grice, H. Paul. 1969. Utterer's Meaning and Intention. *The Philosophical Review* 78(2) : 147–177. DOI : [https://doi.org/10.2307/2184179]

Jurafsky, Dan & Martin, James H. 2014. *Speech and Language Processing*, volume 3. London : Pearson.

Lazaridou, Angeliki, Alexander Peysakhovich & Marco Baroni. 2016. Multi-Agent Cooperation and the Emergence of (Natural) Language. *CoRR* abs/1612.07182.

Manning, D. Christopher & , Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, USA : MIT Press.

Mercier, Hugo & Dan Sperber. 2017. *The Enigma of Reason*. Harvard University Press. DOI : [https://doi.org/10.4159/9780674977860]

Morris, Stephen. 1995. The Common Prior Assumption in Economic Theory. *Economics and Philosophy* 11(2) : 227–253. DOI : [https://doi.org/10.1017/S0266267100003382]

Piantadosi, T. Steven, Harry Tily & Edward Gibson. 2012. The Communicative Function of Ambiguity in Language. *Cognition* 122(3) : 280 – 291. DOI : [https://doi.org/10.1016/j.cognition.2011.10.004]

Reboul, Anne. 2017. *Cognition and Communication in the Evolution of Language*. Oxford : Oxford University Press. DOI : [https://doi.org/10.1093/acprof:oso/9780198747314.001.0001]

Shannon, E. Claude & Warren Weaver. 1949. *The Mathematical Theory of Communication*. Urbana : University of Illinois Press.

Sperber, Dan & Deirdre Wilson. 1995. Postface to the second edition of Relevance. *Communication and cognition* : 255–279.

Szabó, Zoltán Gendler. 2017. Compositionality. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2017 edition.

Vecchi, Eva M., Marco Marelli, Roberto Zamparelli & Marco Baroni. 2017. Spicy Adjectives and Nominal Donkeys: Capturing Semantic Deviance Using Compositionality in Distributional Spaces. *Cognitive Science* 41(1) : 102–136. DOI : [https://doi.org/10.1111/cogs.12330]

Wasow, Thomas, Amy Perfors, & David Beaver. 2005. The puzzle of ambiguity. Morphology and the Web of Grammar. *Essays in Memory of Steven G. Lapointe* : 265–282.

Zipf, K. George. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley.