# Améliorer les modèles de langue pour l'analyse des émotions : perspectives venant des sciences cognitives

Constant Bonard, Gustave Cortal

# Améliorer les modèles de langue pour l'analyse des émotions : perspectives venant des sciences cognitives

Constant Bonard[1] Gustave Cortal[2, 3]
(1) Université de Berne, Département de Philosophie, Hochschulstrasse 4, 3012 Berne, Suisse
(2) Université Paris-Saclay, ENS Paris-Saclay, CNRS, LMF, 91190, Gif-sur-Yvette, France
(3) Université Paris-Saclay, CNRS, LISN, 91400, Orsay, France
constant.bonard@gmail.com, gcortal@ens-paris-saclay.fr

RÉSUMÉ

Nous proposons d'exploiter les recherches en sciences cognitives sur les émotions et la communication pour améliorer les modèles de langue pour l'analyse des émotions. Tout d'abord, nous présentons les principales théories des émotions en psychologie et en sciences cognitives. Puis, nous présentons les principales méthodes d'annotation des émotions en traitement automatique des langues et leurs liens avec les théories psychologiques. Nous présentons aussi les deux principaux types d'analyses de la communication des émotions en pragmatique cognitive. Enfin, en s'appuyant sur les recherches en sciences cognitives présentées, nous proposons des pistes pour améliorer les modèles de langue pour l'analyse des émotions. Nous suggérons que ces recherches ouvrent la voie à la construction de nouveaux schémas d'annotation et d'un possible *benchmark* pour la compréhension émotionnelle, prenant en compte différentes facettes de l'émotion et de la communication chez l'humain.

ABSTRACT

**Improving Language Models for Emotion Analysis : Insights from Cognitive Science.**

We propose leveraging cognitive science research on emotions and communication to improve language models for emotion analysis. First, we present the main emotion theories in psychology and cognitive science. Then, we introduce the main methods of emotion annotation in natural language processing and their connections to psychological theories. We also present the two main types of analyses of emotional communication in cognitive pragmatics. Finally, based on the cognitive science research presented, we propose directions for improving language models for emotion analysis. We suggest that these research efforts pave the way for constructing new annotation schemes and a possible benchmark for emotional understanding, considering different facets of human emotion and communication.

MOTS-CLÉS : Analyse des émotions, modèle de langue, sciences cognitives, sciences affectives, pragmatique.

KEYWORDS: Emotion analysis, language model, cognitive science, affective sciences, pragmatics.

## 1 Introduction

L'analyse des émotions dans le traitement automatique des langues vise à développer des modèles computationnels capables de discerner les émotions humaines dans le texte. Récemment, les modèles de langue ont largement été utilisés pour résoudre diverses tâches en traitement automatique des langues, dont l'analyse des émotions (Devlin *et al.*, 2019; Brown *et al.*, 2020). Ce domaine de recherche fait face à plusieurs limites. Tout d'abord, les différentes façons de conceptualiser les émotions amènent à différents schémas d'annotation et jeux de données (Klinger, 2023). En conséquence, la capacité de généralisation des modèles est limitée et il est souvent impossible de comparer les études. Pour remédier à ces limites, il a été proposé d'unifier certains schémas d'annotation en se basant sur la proximité sémantique des catégories d'émotions (Bostan & Klinger, 2018), de trouver automatiquement les catégories d'émotions d'intérêts à partir des données (De Bruyne *et al.*, 2020) ou d'obtenir des plongements émotionnels indépendants des schémas d'annotation (Buechel *et al.*, 2021). En s'inspirant de débats récents en sciences cognitives (Scherer, 2022), nous pensons qu'il serait possible de construire un schéma d'annotation unifiant différentes perspectives sur l'émotion.

De plus, les *benchmarks* existants évaluent certains aspects de la compréhension émotionnelle, mais sans prendre en compte toute sa complexité (Campagnano *et al.*, 2022; Zhang *et al.*, 2023; Paech, 2024). Par exemple, Paech (2024) propose d'évaluer la compréhension émotionnelle des modèles de langue à travers la prédiction de l'intensité des émotions dans des scènes de conflits. Ce type d'évaluation est trop limité : les *benchmarks* devraient refléter autant que possible la richesse de la compréhension émotionnelle chez les humains, une richesse documentée ces dernières décennies dans différentes branches des sciences affectives (Green, 2007; Wharton, 2016; Scarantino, 2017; Barrett *et al.*, 2019; Bonard & Deonna, 2023).

Un autre domaine de recherche connexe se concentre sur la théorie de l'esprit des modèles de langue, c'est-à-dire leur capacité à attribuer correctement des états mentaux aux autres. Cette littérature est prometteuse selon nous en cela qu'elle relie les développements récents des modèles de langue aux théories et aux méthodes empiriques des sciences cognitives (Bonard, 2024, section 5). Notamment, plusieurs tâches et *benchmarks* ont été développés pour mesurer la capacité des modèles de langue à réussir différentes versions de la tâche de la fausse croyance (*False Belief Task*) (Trott *et al.*, 2022; Aru *et al.*, 2023; Gandhi *et al.*, 2023; Holterman & van Deemter, 2023; Kosinski, 2023; Mitchell & Krakauer, 2023; Shapira *et al.*, 2023; Stojnić *et al.*, 2023; Ullman, 2023).

*Les auteurs ont contribué à parts égales et apparaissent par ordre alphabétique.

Cependant, la théorie de l'esprit et, plus généralement, les capacités de raisonnement social vont au-delà de la capacité à réussir la tâche de la fausse croyance (Apperly & Butterfill, 2009; Langley *et al.*, 2022; Ma *et al.*, 2023). La capacité à interpréter correctement les émotions exprimées ne peut s'y réduire. Le degré de possession de cette compétence émotionnelle par les modèles de langue mérite d'être étudiée en soi.

D'une façon générale, la recherche portant sur les modèles de langue pour l'analyse des émotions bénéficierait d'un apport de la recherche en sciences cognitives. Notamment, nous pensons que cela peut mener à de meilleures manières d'annoter les émotions exprimées dans le texte, mais aussi à une meilleure évaluation de la compréhension émotionnelle des modèles de langue en développant de nouveaux *benchmarks*. Nous présentons un panorama général sur les théories psychologiques des émotions (section 2) et sur les manières d'annoter les émotions dans le traitement automatique des langues (section 3). Puis, en s'inspirant de certaines théories psychologiques et linguistiques (section 4), nous proposons des directions de recherche pour remédier à certaines limites actuelles de l'analyse des émotions (section 5).

**Contributions.** Pour améliorer l'analyse automatique des émotions, nous proposons d'intégrer différentes théories des sciences cognitives avec le TAL. Nous expliquons pourquoi et comment l'analyse des émotions devrait utiliser des théories en psychologie des émotions – en particulier le cadre intégré – ainsi que des théories en pragmatique cognitive – en particulier l'analyse du détective. Cela conduit à l'élaboration d'un nouveau schéma d'annotation et à une meilleure évaluation des modèles de langue.

## 2 Les théories des émotions dans les sciences cognitives

Cette section présente les trois principales théories des émotions en psychologie afin de fournir un arrière-plan quant à notre projet de mieux connecter l'analyse des émotions en traitement automatique des langues avec les sciences cognitives.

**La théorie des émotions de base.** La théorie des émotions de base est certainement la plus influente aujourd'hui. Inspirée par les recherches de Darwin sur les émotions (Darwin, 1872), elle postule un certain nombre d'émotions discrètes et fondamentales qui sont universelles et innées chez les humains en raison de leurs origines évolutives. Les émotions sont comprises comme des « programmes » psycho-physiologiques qui ont été sélectionnés pour aider à surmonter les défis évolutifs récurrents (Cosmides & Tooby, 2000). Une version importante de cette théorie est celle de Paul Ekman (Ekman, 1999), qui a cherché à montrer, comme le prévoyait Darwin, que certaines émotions sont associées aux mêmes expressions faciales à travers toutes les cultures - Ekman a utilisé la liste d'émotions proposées par Darwin (Darwin, 1872) : colère, peur, surprise, dégoût, bonheur et tristesse. Il a notamment mené des études auprès d'individus n'ayant pas été exposés à la culture occidentale, indiquant qu'ils pouvaient correctement identifier les expressions faciales pour ces six émotions sur des photographies (Ekman & Friesen, 1971). Des tentatives ont également été faites pour soutenir la théorie des émotions de base en identifiant des signatures physiologiques et neurologiques des émotions de base (Moors, 2022, p. 129–131).

Il convient de noter qu'Ekman n'a pas précisé le nombre exact d'émotions de base. Outre les six émotions énumérées, les candidats comprennent l'amusement, le mépris, la gêne, la culpabilité, la fierté et la honte (Ekman, 1999). D'autres versions de la théorie des émotions de base proposent différentes listes (Tomkins, 1962; Izard, 1992; Panksepp, 1998; Plutchik, 2001).

**Le constructivisme psychologique.** Le constructivisme psychologique est l'alternative la plus influente à la théorie des émotions de base aujourd'hui. Il rejette l'idée qu'il existe des émotions discrètes et fondamentales partagées universellement par les humains et postule au contraire que les types d'émotions tels que la colère, la peur et la joie sont construits à travers l'interaction de facteurs biologiques, psychologiques et socioculturels. Parmi les adeptes de la première heure figurent Schachter & Singer (1962). Cette théorie est aujourd'hui principalement associée à James Russell et Lisa Feldman Barrett (Russell, 1980, 2003; Barrett, 2006; Russell, 2009; Barrett, 2017). Les constructivistes psychologiques se concentrent sur les sentiments subjectifs associés aux émotions qui sont interprétés comme un continuum sans barrière catégorique. Les sentiments sont généralement représentés dans un espace bidimensionnel avec un axe de valence (sentiments agréables-désagréables) et un axe d'excitation (sentiments d'activation-désactivation). L'impression qu'il existe des émotions distinctes est considérée comme une construction sociale : différentes formes d'acculturation suscitent différentes façons de conceptualiser ou d'étiqueter nos sentiments corporels en types d'émotions distincts.

**La théorie de l'évaluation (*appraisal*).** La troisième théorie psychologique majeure de l'émotion est la théorie de l'évaluation (*appraisal*), dont la version empirique a été initiée par Magda Arnold (Arnold, 1960). Elle a été développée pour expliquer l'absence de correspondance bijective, une-à-une, entre les types d'émotions et les types de stimuli émotionnels, c'est-à-dire le fait que le même type de stimuli peut déclencher des émotions différentes et que des types de stimuli différents peuvent déclencher le même type d'émotion. Pour expliquer ce phénomène, des évaluations (*appraisals*) sont postulées comme médiatrices entre les stimuli et les réactions émotionnelles.

Les évaluations en question sont des catégorisations cognitives (inconscientes, rapides et souvent erronées) de la pertinence des stimuli par rapport aux préoccupations de la personne et de la manière dont elle doit y réagir. La théorie de l'évaluation postule que, par exemple, Simon a peur de la souris dans la cuisine parce qu'il l'évalue comme une menace imminente pour sa sécurité, tandis que Sylvie, au contraire, est en colère qu'il y ait une souris dans la cuisine parce qu'elle l'évalue comme un intrus à chasser. Ainsi, chaque type d'émotion peut être analysé par le type d'évaluation qui lui est associée. Ainsi, Lazarus (1991) propose *le danger imminent* pour la peur, *l'offense dégradante* pour la colère, *la perte irrévocable* pour la tristesse et *le progrès vers un but* pour la joie.

Dans les années 1980, des adeptes de cette théorie ont proposé d'analyser les évaluations comme des régions dans un espace multidimensionnel (Moors *et al.*, 2013). Ces dimensions d'évaluation comprennent généralement : (a) la pertinence du stimulus par rapport aux objectifs de l'individu, (b) la capacité de l'individu à faire face à la situation, (c) l'urgence de la réponse nécessaire, (d) la cause de l'événement déclencheur (moi, quelqu'un d'autre, intentionnelle ou non) et (e) la compatibilité avec les normes personnelles de l'individu. Par exemple, la peur est déclenchée par l'évaluation d'un stimulus comme étant (a) fortement contraire aux objectifs, (b) difficile à gérer et (c) nécessitant une réponse urgente.

**Cadre intégré pour les théories des émotions.** Bien que les trois théories examinées soient habituellement considérées comme rivales, il a été défendu qu'il fallait au contraire les intégrer dans un cadre commun (Scherer & Moors, 2019; Bonard, 2021b; Scherer, 2022). En effet, on peut affirmer que les trois théories diffèrent avant tout vis-à-vis de l'objet de leur enquête, leur axe de recherche et les aspects de l'émotion sur lesquels elles mettent l'accent. La théorie des émotions de base se concentre sur les traits universels hérités de l'évolution et en particulier sur leurs expressions physiologiques et motrices (réactions corporelles). Le constructivisme psychologique se concentre sur les dimensions du ressenti et la façon dont les individus les catégorisent (sentiment subjectif). La théorie de l'évaluation se concentre sur le déclenchement émotionnel (processus d'évaluation) et les tendances à l'action qui en découlent. Nous pensons qu'un cadre intégrant les différents éléments étudiés par ces théories est possible et souhaitable. Ce que nous appelons « le cadre intégré pour les théories des émotions » propose de le faire en postulant que les épisodes émotionnels paradigmatiques sont faits de changements synchronisés et causalement interconnectés dans quatre composantes : (1) processus d'évaluation (*appraisal*), (2) tendances à l'action, (3) changements corporels (expressions motrices et réponses physiologiques), (4) sentiments subjectifs. Pour une discussion d'un tel cadre intégré, voir Scherer (2022).

## 3 L'analyse des émotions dans le texte

**L'émotion est une catégorie.** L'analyse des émotions dans le texte s'appuie sur les théories de l'émotion de base pour définir les différentes catégories d'émotion à associer aux unités textuelles (un empan de texte, une phrase ou un document). Par exemple, la phrase « J'adore la philosophie. » pourrait être associée automatiquement à l'émotion discrète *joie*. Plusieurs schémas d'annotation se concentrent sur des sous-ensembles de catégories alors que d'autres considèrent un plus large ensemble, pouvant atteindre plus de 28 catégories différentes (Demszky *et al.*, 2020; Bostan & Klinger, 2018).

**L'émotion est une valeur continue ayant un sens affectif.** Au lieu de représenter l'émotion par une catégorie, certains schémas d'annotation considèrent que l'émotion est un point dans un espace multidimensionnel et associent à des unités textuelles des valeurs continues (Buechel & Hahn, 2017). Ces dimensions portent un sens affectif. Deux dimensions sont dominantes dans la littérature et proviennent des théories du constructivisme psychologique qui considèrent qu'une émotion peut être caractérisée par son degré d'*agréabilité* et son degré d'*activation physiologique*. Ainsi, la phrase « Sa voix m'apaise. » pourrait être associée automatiquement à deux valeurs continues : un degré d'*agréabilité* de 4 sur 5 et un degré d'*activation physiologique* de 1 sur 5.

**L'émotion est une valeur continue ayant un sens cognitif.** Ces dimensions peuvent aussi porter un sens cognitif. Récemment, une nouvelle ligne de recherche propose d'incorporer les théories psychologiques de l'évaluation cognitive dans les modèles d'analyse des émotions (Hofmann *et al.*, 2020; Troiano *et al.*, 2022; Zhan *et al.*, 2023). Depuis cette perspective, les émotions sont causées par des évènements évalués selon plusieurs dimensions cognitives. Par exemple, la phrase « J'ai reçu un cadeau surprise. » pourrait être associée automatiquement à plusieurs valeurs continues : l'évènement est *soudain* (4 sur 5), *contraire aux normes sociales* (0 sur 5) et la personne a le *contrôle* sur l'évènement (0 sur 5).

**L'émotion est constituée de rôles sémantiques.** Une émotion ne peut se réduire à une catégorie ou des valeurs continues ayant un sens affectif ou cognitif. Pour avoir une meilleure compréhension d'un évènement émotionnel, plusieurs approches associent à des empans de texte des rôles sémantiques comme la *cause*, la *cible*, l'*expérienceur·euse* et l'*indice* de l'émotion (Lee *et al.*, 2010; Kim & Klinger, 2018; Bostan *et al.*, 2020; Oberländer *et al.*, 2020; Campagnano *et al.*, 2022; Wegge *et al.*, 2023; Cortal, 2024). Ainsi, au lieu de considérer l'émotion comme causée par un évènement, l'analyse des rôles sémantiques de l'émotion considère que l'émotion *est* un évènement (Klinger, 2023) qu'il faut reconstituer en répondant à la question : « Qui (*expérienceur·euse*) ressent quoi (*indice*) envers qui (*cible*) et pourquoi (*cause*) ? ». Dans cet exemple, chaque empan de texte peut être associé à un rôle sémantique : « Louise (*expérienceuse*) était en colère (*indice*) contre Paul (*cible*), car il ne l'a pas prévenue (*cause*). »

**Première limite : il n'existe pas de schéma d'annotation unifié.** Les divergences dans la définition de l'émotion en psychologie mènent vers des divergences dans la manière d'annoter l'émotion dans le texte. Les différentes théories psychologiques des émotions représentent différentes perspectives sur le phénomène émotionnel. Elles sont loin de se contredire et peuvent même tendre à s'unifier (section 2). Nous pensons que c'est aussi le cas pour les schémas d'annotation dans l'analyse des émotions. Dans la section 5, nous donnons des pistes pour la construction d'un schéma d'annotation unifié, inspiré par les débats récents en sciences cognitives (Scherer, 2022).

**Seconde limite : la verbalisation de l'émotion est peu considérée.** L'analyse des émotions considère rarement le processus de verbalisation de l'émotion. En conséquence, il est difficile d'obtenir des guides d'annotation qui définissent clairement les marqueurs linguistiques à annoter dans le texte. Nous voulons mettre en lumière la théorie linguistique de Raphaël Micheli, qui catégorise un large panel de marqueurs linguistiques en trois modes d'expression de l'émotion (Micheli, 2014) : l'émotion peut être *dite*, *montrée* ou *suggérée* (ou « étayée »). L'émotion peut être exprimée explicitement avec un terme du lexique émotionnel (« Je suis *triste* »), être montrée avec des caractéristiques de l'énoncé comme les interjections

et les ponctuations (« *Ah !* C'est super*!* »), ou être suggérée avec la description d'une situation qui généralement, dans un contexte socioculturel donné, mène à une émotion (« *Elle m'a offert un cadeau* »). La majorité des schémas d'annotation se sont concentrés implicitement sur l'émotion *dite*, en occultant les deux autres modes d'expression. Récemment, les schémas d'annotation basés sur les théories de l'évaluation cognitive s'intéressent implicitement à l'émotion *suggérée*. La théorie de Micheli analyse donc les différents types de signes verbaux qui sont utilisés, chez les humains, pour inférer les émotions exprimées. Par contraste, les théories de la pragmatique cognitive s'intéressent aux mécanismes psychologiques qui sont utilisés pour inférer ce qui est communiqué, dont notamment les émotions exprimées par ces différents types de signes. Dans la prochaine section, nous suggérerons l'hypothèse que les catégories de signes distinguées par Micheli correspondent à différentes sources d'inférences postulées par la pragmatique cognitive.

# 4 Pragmatique cognitive et communication émotionnelle

**Deux analyses de la communication.** La pragmatique cognitive est la branche des sciences cognitives qui s'intéresse à la façon dont les personnes utilisent et interprètent les signes dans la communication. Dans cette branche et d'autres disciplines connexes, il est courant de distinguer deux grandes manières d'analyser la communication : l'analyse du dictionnaire (également appelée « modèle du code », « sémiotique » ou « sémantique ») et l'analyse du détective (également appelée « analyse gricéene », « modèle inférentiel » ou « pragmatique ») (Sperber & Wilson, 1995; Schlenker, 2016; Heintz & Scott-Phillips, 2023).

**Analyse du dictionnaire.** L'analyse du dictionnaire décrit la communication comme suit : les expéditeur·ices *encodent* (intentionnellement ou non) des informations dans un signal que les destinataires *décodent*. De manière vitale, avant l'échange communicatif, les expéditeur·ices et destinataires doivent partager le même *code*. Par « code », on entend ici une association préétablie entre des types de stimuli (symbolisés par « <...> ») et des ensembles d'informations (symbolisés par « [...] »). Par exemple, le code Morse consiste en une association entre <combinaisons de signaux courts et longs> et [lettres] qui doit être partagée pour communiquer avec lui. Les codes peuvent être conventionnels, comme le code Morse, mais aussi comme la sémantique formelle d'une langue : un code fait de règles syntaxiques et lexicales qui associent des <chaînes de mots> à des [significations de phrases] (Heim & Kratzer, 1998). Les codes peuvent également être non conventionnels ou « naturels » (Wharton, 2003; Bonard, 2023a). Par exemple, les abeilles utilisent un code associant leurs <danses> à la [localisation du nectar]. Comme mentionné dans la section 2, Darwin ou Ekman postulent que les humains utilisent un code transmis génétiquement qui associe des types d'<expressions faciales> à des types d'[émotions exprimées].

La principale limite de l'analyse du dictionnaire est que, parfois, les codes *sous-déterminent* le sens : les associations préétablies entre <types de stimuli> et [ensembles d'informations] sont parfois insuffisantes pour rendre compte de l'information communiquée. De manière paradigmatique, les *implicatures conversationnelles* (Grice, 1975) communiquent implicitement des informations au-delà de ce qui est linguistiquement encodé, au-delà de ce qui est déterminé par les règles syntaxiques et lexicales de la langue utilisée. Par exemple (Wilson & Sperber, 2006), si Pierre demande : « Est-ce que Jean t'a remboursé l'argent qu'il te devait ? » et Marie répond : « Il a oublié d'aller à la banque. », Pierre comprendra facilement que Marie veut dire « non » bien que le code pertinent - les règles associant <la grammaire et le lexique français> à [la signification des phrases] - soit insuffisant à lui seul pour en rendre compte, puisque le code ne dit que Jean a oublié d'aller à la banque.

Cette limite de l'analyse du dictionnaire concerne également l'expression verbale des émotions. Pour l'illustrer, revenons à la typologie de Micheli : émotions *dites*, *montrées* et *suggérées* (Micheli, 2013). En ce qui concerne les émotions *dites*, l'analyse du dictionnaire fonctionne assez bien grâce à l'association entre <mots d'émotion> (par exemple, « heureux », « merveilleux », « tristement ») et les [types d'émotion] auxquels ils font référence. Cependant, même les émotions *dites* n'encodent parfois pas tout ce qui est communiqué. Par exemple, « Je suis triste. » est explicite sur le type d'émotion exprimée, mais n'encode pas ce sur quoi porte l'émotion. Néanmoins, dans le contexte pertinent, nous comprenons en général à propos de quoi porte la tristesse en question. L'analyse du dictionnaire s'en sort encore moins bien avec les émotions *montrées*, car celles-ci sont souvent ambiguës. Par exemple, des interjections telles que « Wow ! », « Oulalala ! », « Diantre ! », « Ah ! » et « Oh ! », bien qu'elles montrent de façon évidente qu'une émotion est exprimée, peuvent en fait exprimer une variété d'émotions positives et négatives. De plus, ces interjections n'encodent pas non plus ce sur quoi porte l'émotion en question. Cependant, les destinataires réussissent généralement à inférer ces informations. L'analyse du dictionnaire est encore plus limitée quand il s'agit d'émotions *suggérées*. En fonction de ce que croit ou souhaite la personne exprimant son émotion, une phrase ne faisant que suggérer l'émotion peut en communiquer une multitude indéfinie. Imaginez, par exemple, que quelqu'un dise « Le navire a des voiles noires. ». Dans un certain contexte, cette phrase apparemment dénuée d'affectivité peut en fait transmettre de manière poignante une émotion intense - parce que, disons, elle signifie que le fils de celui qui prononce la phrase est mort, comme dans l'histoire d'Égée et Thésée. Il convient de noter qu'au-delà de l'expression verbale, la plupart, voire tous les types d'expressions émotionnelles, sous-déterminent également ce qui est communiqué par les émotions exprimées. Les expressions faciales ou les indices acoustiques (par exemple, cris, rires, soupirs) communiquent également différentes émotions en fonction des contextes (Aviezer *et al.*, 2008; Teigen, 2008; Vlemincx *et al.*, 2009; Barrett *et al.*, 2011, 2019; Bonard, 2023b). L'analyse du code est donc aussi insuffisante pour ce genre d'expressions émotionnelles.

Comment donc les humains désambiguïsent-ils les expressions émotionnelles dans les cas où les codes des expressions émotionnelles sous-déterminent ce qui est communiqué ? Si l'on se fie à la pragmatique cognitive contemporaine, la réponse devrait se trouver dans l'analyse de la communication dite « du détective ».

**L'analyse du détective.** Ce que nous appelons « l'analyse du détective » est constitué d'une famille de théories développées par Paul Grice (Grice, 1957, 1989) et ses héritier·ères (pour un compte rendu, voir Bonard (2021a), chapitre 1 et appendice). Notez que bien que notre présentation vise à rester équilibrée entre différentes théories, il n'existe pas de version universellement acceptée de cette analyse.

Comme mentionné, l'analyse du détective a été développée pour rendre compte des implicatures conversationnelles, des cas où ce qui est communiqué va au-delà de ce qui est transmis par le sens littéral des mots utilisés, comme dans l'exemple de Pierre et Marie ci-dessus. Pour ce faire, l'analyse du détective conceptualise l'interprétation linguistique comme un type de raisonnement *abductif* - c'est-à-dire une inférence qui cherche la conclusion la plus simple et la plus probable en fonction des données probantes disponibles. L'analyse décrit trois sources principales de données probantes :

1. *Les codes*, par exemple les règles syntaxiques et lexicales de l'anglais ou encore les codes des expressions émotionnelles verbales et non verbales. Comme nous l'avons vu avec la typologie de Micheli (Micheli, 2013), les expressions utilisant des émotions dites (par exemple, « Je suis triste. ») et montrées (par exemple, « Wow ! ») sont partiellement comprises grâce à de tels codes, bien que ces derniers soient trop ambigus pour rendre compte de tout ce qui est communiqué ;
2. *Les attentes pragmatiques*, c'est-à-dire les attentes concernant la façon dont les gens sont censés se comporter dans des contextes donnés et en particulier en fonction du type de signal qu'ils ont reçu. Par exemple, dans les conversations, on s'attend à ce que soient dites des choses *pertinentes* quant à la question discutée (voir les maximes de conversation de Grice (Grice, 1975)). Pour cette raison, bien que ce qui est littéralement encodé dans la réponse de Marie soit que Jean a oublié d'aller à la banque, Pierre s'attendra néanmoins à ce que cela soit pertinent à la question qu'il a posée. De même, nous nous attendons à ce que les expressions émotionnelles de quelqu'un portent sur quelque chose qui lui importe particulièrement (Wharton *et al.*, 2021; Bonard, 2022). Par exemple, si quelqu'un dit « Diantre ! » après avoir reçu un compliment étonnamment gentil, nous nous attendons à ce que le compliment importe particulièrement à la personne et interpréterons l'interjection en fonction de cela ;
3. *Les connaissances partagées* (en anglais *common ground*), c'est-à-dire l'information que les participants à l'échange présument partager (Stalnaker, 2002). Par exemple, Marie et Pierre présument qu'une banque est un endroit où l'on peut retirer de l'argent. De même, on présume généralement que recevoir un compliment est une chose que l'on recherche, surtout s'il est agréablement surprenant – bien que cela ne fasse pas toujours partie des connaissances partagées, par exemple si l'on sait que le compliment vient de l'ennemi juré de la personne complimentée. C'est aussi les connaissances partagées qui nous permettent de comprendre qu'Égée peut exprimer un profond désespoir avec la phrase « Le navire a des voiles noires. ».

En se basant sur ces trois sources de données probantes, l'analyse du détective postule ensuite que l'interprète utilise ses capacités de « lecture de l'esprit » (en anglais *mindreading*, aussi appelée « théorie de l'esprit », « mentalisation » ou « cognition sociale ») pour inférer l'information la plus probable qui est implicitement communiquée. Par exemple, Pierre déduit que Marie voulait dire « non, il ne m'a pas rendu mon argent » et nous inférons que la personne qui dit « Diantre ! » après avoir reçu le compliment est probablement contente (sauf si le compliment vient de son ennemi juré). Enfin, l'analyse du détective précise que l'information ainsi inférée est ajoutée aux connaissances partagées des personnes participant à l'échange, de sorte qu'elle puisse devenir une nouvelle source de données probantes dans la suite de l'échange ou les échanges suivants.

Il est intéressant de noter que l'analyse du détective prédit que la capacité à inférer correctement ce qui est communiqué par les expressions émotionnelles dépend fortement de nos capacités de « lecture de l'esprit ». En corroboration de cette prédiction, les personnes autistes ou les enfants peuvent avoir du mal à inférer correctement le sens implicite, par exemple dans les implicatures conversationnelles (Foppolo & Mazzaggio, 2024) ou dans les expressions utilisant des émotions suggérées (Blanc & Quenette, 2017; Etienne *et al.*, 2022).

## 5 Les directions de recherche pour l'analyse des émotions

**Vers un schéma d'annotation unifié.** Pour améliorer la compréhension émotionnelle des modèles, il est souhaitable de les entraîner sur des données annotées avec un schéma rendant compte fidèlement d'une situation émotionnelle. Un tel schéma devrait intégrer différentes perspectives sur le phénomène émotionnel pour permettre de meilleures comparaisons entre les études ainsi qu'augmenter les performances et la généralisation des modèles.

**Les tentatives d'unification.** Plusieurs études récentes essayent d'unir différentes manières d'annoter l'émotion dans le texte. Campagnano *et al.* (2022) proposent un nouveau schéma d'annotation qui unifie plusieurs schémas sur les rôles sémantiques des émotions. Pour choisir un ensemble de catégories partagées, les différentes émotions discrètes des schémas ont été converties vers les émotions de base de la théorie de Plutchik (Plutchik, 2001). Klinger (2023) explore les divergences et les points communs entre l'analyse des rôles sémantiques de l'émotion et les approches basées sur l'évaluation cognitive. L'étude identifie plusieurs directions de recherche, comme l'utilisation des variables de l'évaluation cognitive pour améliorer la tâche de détection des causes de l'émotion, ou l'analyse des évaluations cognitives spécifiques aux expérienceur·euses (Wegge *et al.*, 2023). Ces études montrent que l'unification des schémas permet le transfert de connaissance entre différentes tâches, ce qui augmente les performances et la généralisation des modèles.

**À la recherche d'un cadre commun.** Ce que nous avons appelé plus haut « le cadre intégré pour les théories de l'émotion »

(section 2) vise à réconcilier les principales théories de l'émotion en sciences cognitives (Scherer, 2022). Il constitue selon nous un bon candidat pour fournir un cadre commun aux schémas d'annotation. Pour rappel, ce modèle considère qu'une émotion est constituée de changements synchronisés dans différents composants : le processus d'évaluation, les tendances à l'action, les changements corporels (expressions motrices et réponses physiologiques) et les sentiments subjectifs. La recherche en analyse des émotions doit s'inspirer des récents débats en psychologie des émotions pour faire dialoguer les schémas d'annotation existants sur une base théorique solide et, idéalement, construire un schéma d'annotation unifié.

**L'émotion est constituée de plusieurs composantes en interaction.** Un schéma d'annotation unifié pourrait clarifier certaines zones d'ombre existantes dans l'analyse des émotions, comme l'absence de définitions claires des rôles sémantiques liés à l'émotion, comme l'expérienceur·euse, la cause et la cible. Il pourrait aussi permettre de mieux situer les schémas existants. Par exemple, l'annotation des émotions discrètes et des dimensions affectives met l'accent sur le sentiment subjectif, alors que l'annotation des dimensions cognitives met l'accent sur l'évaluation cognitive. Peu de schémas rendent compte des réponses physiologiques, des expressions motrices et des tendances à l'action. Plus généralement, peu de schémas considèrent la totalité des composantes. Kim & Klinger (2019) analysent la communication des émotions dans des fictions à travers des descriptions de sensations subjectives, de postures, d'expressions faciales et de relations spatiales entre les personnages. Casel *et al.* (2021) associent à des empans de texte des catégories correspondant aux composantes de Scherer. Cortal *et al.* (2022, 2023) structurent des récits narratifs émotionnels selon des composantes similaires à celles de Scherer. Chaque empan de texte correspond à des comportements observables, des pensées, des ressentis physiques ou des évaluations cognitives. À notre connaissance, il n'existe pas de schémas d'annotation qui essaient de capturer l'interaction entre les composantes. Généralement, l'analyse des émotions se concentre peu sur le caractère dynamique de l'émotion et la synchronisation des diverses composantes.

**Améliorer la clarté des guides d'annotation.** Nous soulignons que peu d'études justifient psychologiquement le choix des différents objets à détecter dans le texte. L'analyse des émotions a besoin de développer une approche systématique pour comparer les guides d'annotation entre eux et ainsi comprendre précisément comment l'émotion est capturée par les différents schémas d'annotation. Avec des guides d'annotation clairs, il sera plus facile pour les équipes de recherche de se concentrer sur les points de convergence entre les schémas. Ainsi, ces schémas devront s'inspirer des théories en psychologie des émotions (section 2) mais aussi des théories linguistiques (sections 3 et 4) pour identifier les marqueurs linguistiques qui verbalisent l'émotion.

**Vers une meilleure évaluation de la compréhension émotionnelle.** Récemment, les *benchmarks* sur les émotions évaluent des modèles de langue sur certains aspects de la compréhension émotionnelle (Wang *et al.*, 2023; Paech, 2024), sans prendre en compte toute sa richesse (Scherer, 2007; Mayer *et al.*, 2008; O'Connor *et al.*, 2019). Par exemple, Paech (2024) évalue la compréhension émotionnelle à travers la prédiction de l'intensité de plusieurs émotions dans des scènes de conflits. Il existe aussi des *benchmarks* qui évaluent des modèles sur des tâches connexes, comme l'analyse du sentiment (Zhang *et al.*, 2023) et la théorie de l'esprit (Zhou *et al.*, 2023; Ma *et al.*, 2023; Kim *et al.*, 2023; Gandhi *et al.*, 2023). Ainsi, il n'existe aucun *benchmark* qui propose d'évaluer spécifiquement plusieurs aspects du phénomène émotionnel. Il est donc difficile de savoir si les modèles actuels sont performants pour la compréhension émotionnelle.

Cette limite s'ajoute au fait qu'il est difficile de déterminer clairement les propriétés de la compréhension émotionnelle à évaluer. Nous pensons qu'il faudrait s'inspirer de la communication émotionnelle chez les humains pour évaluer les modèles de langue, et notamment des travaux en psycholinguistique. Ainsi, avant dix ans, les émotions de base (par exemple, la joie ou la tristesse) sont mieux retenues que les émotions complexes (par exemple, la fierté ou la culpabilité) (Davidson *et al.*, 2001; Creissen & Blanc, 2017). De six à dix ans, les émotions *dites* sont mieux comprises que les émotions *suggérées* (Blanc, 2010; Creissen & Blanc, 2017). Un autre exemple d'études pertinentes concerne la plus ou moins grande facilité qu'ont les personnes sur le spectre de l'autisme à comprendre différents types d'expressions émotionnelles (Foppolo & Mazzaggio, 2024). Ces études montrent que, pour les humains, différents types d'émotions et différents modes d'expressions émotionnels sont plus ou moins difficiles à interpréter. Il serait souhaitable que les *benchmarks* évaluent les modèles de langue de manière à refléter la plus ou moins grande difficulté des tâches pour les humains. Un tel projet bénéficierait certainement des recherches en pragmatique cognitive (section 4) sachant, par exemple, que des personnes souffrant de troubles de la communication ont du mal à comprendre les implicatures conversationnelles (Foppolo & Mazzaggio, 2024), ce qui indique que les différentes sources de données probantes distinguées par l'analyse du détective impliquent différents degrés de difficultés.

Nous pensons que le concept d'émotion doit être adressé à travers sa relation avec la compréhension du texte, c'est-à-dire la capacité qu'à un·e lecteur·rice à construire une représentation mentale d'une situation dans un texte (Zwaan & Radvansky, 1998). Ainsi, il faudrait aller au-delà des conceptualisations courantes de l'émotion en traitement automatique des langues (section 3) pour prendre en compte la diversité des marqueurs linguistiques employés pour verbaliser l'émotion (section 3) ainsi que les différents types d'émotion (basique ou complexe) issus des travaux en psycholinguistiques. Etienne *et al.* (2022) ont proposé un schéma d'annotation inspiré par les études précédentes qui considère les modes d'expression de l'émotion et les types d'émotion. De futurs *benchmarks* évaluant les capacités des modèles de langue à analyser les émotions devraient prendre en compte de tels schémas d'annotation qui, comme nous l'avons recommandé, cherchent à solidement se baser sur les recherches pertinentes en sciences cognitives.

# 6 Conclusion

Pour remédier à certaines limites dans l'analyse des émotions, nous avons proposé d'exploiter les recherches en sciences cognitives sur les émotions et la communication. Nous avons expliqué pourquoi et comment l'analyse des émotions devrait utiliser des théories en psychologie des émotions – en particulier le cadre intégré – ainsi que des théories en pragmatique cognitive – en particulier l'analyse du détective. Ces recherches ouvrent la voie à la construction de nouveaux schémas d'annotation et d'un possible *benchmark* pour la compréhension émotionnelle, considérant différentes facettes de l'émotion et de la communication chez l'humain.

# Références


APPERLY I. A. & BUTTERFILL S. A. (2009). Do humans have two systems to track beliefs and belief-like states ? *Psychological review*, **116**(4), 953. Publisher : American Psychological Association.

ARNOLD M. B. (1960). *Emotion and Personality*. New York : Columbia University Press.

ARU J., LABASH A., CORCOLL O. & VICENTE R. (2023). Mind the gap : challenges of deep learning approaches to Theory of Mind. *Artificial Intelligence Review*, **56**(9), 9141–9156. DOI : 10.1007/s10462-023-10401-x.

AVIEZER H., HASSIN R. R., RYAN J., GRADY C., SUSSKIND J., ANDERSON A., MOSCOVITCH M. & BENTIN S. (2008). Angry, disgusted, or afraid ? Studies on the malleability of emotion perception. *Psychological science*, **19**(7), 724–732. Publisher : SAGE Publications Sage CA : Los Angeles, CA.

BARRETT L. F. (2006). Solving the Emotion Paradox : Categorization and the Experience of Emotion. *Personality and Social Psychology Review*, **10**(1), 20–46. DOI : 10.1207/s15327957pspr1001_2.

BARRETT L. F. (2017). *How Emotions Are Made : The Secret Life of the Brain*. Boston & New York : Houghton Mifflin Harcourt.

BARRETT L. F., ADOLPHS R., MARSELLA S., MARTINEZ A. M. & POLLAK S. D. (2019). Emotional expressions reconsidered : Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*. Publisher : SAGE PublicationsSage CA : Los Angeles, CA, DOI : 10.1177/1529100619832930.

BARRETT L. F., MESQUITA B. & GENDRON M. (2011). Context in emotion perception. *Current Directions in Psychological Science*, **20**(5), 286–290. Publisher : Sage Publications Sage CA : Los Angeles, CA.

BLANC N. (2010). La compréhension des contes entre 5 et 7 ans : Quelle représentation des informations émotionnelles ? [The comprehension of the tales between 5 and 7 year-olds : Which representation of emotional information ?]. *Canadian Journal of Experimental Psychology / Revue canadienne de psychologie expérimentale*, **64**(4), 256–265. DOI : 10.1037/a0021283.

BLANC N. & QUENETTE G. (2017). La production d'inférences émotionnelles entre 8 et 10 ans : quelle méthodologie pour quels résultats ? *Enfance*, **4**(4), 503–511. Publisher : NecPlus.

BONARD C. (2021a). *Meaning and emotion : The extended Gricean model and what emotional signs mean*. Doctoral dissertation, University of Geneva and University of Antwerp.

BONARD C. (2021b). Émotions et sensibilité aux valeurs : quatre conceptions philosophiques contemporaines. *Revue de métaphysique et de morale*, **110**(2), 209–229. Place : Paris cedex 14 Publisher : Presses Universitaires de France, DOI : 10.3917/rmm.212.0209.

BONARD C. (2022). Beyond ostension : Introducing the expressive principle of relevance. *Journal of Pragmatics*, **187**, 13–23. DOI : 10.1016/j.pragma.2021.10.024.

BONARD C. (2023a). Natural meaning, probabilistic meaning, and the interpretation of emotional signs. *Synthese*, **201**(5), 167. Publisher : Springer, DOI : https ://doi.org/10.1007/s11229-023-04144-z.

BONARD C. (2023b). Underdeterminacy without ostension : A blind spot in the prevailing models of communication. *Mind & Language*. DOI : https ://doi.org/10.1111/mila.12481.

BONARD C. (2024). Can AI and humans genuinely communicate ? In A. STRASSER, Éd., *Anna's AI Anthology. How to live with smart machines ?* Berlin : Xenemoi.

BONARD C. & DEONNA J. (2023). Emotion and language in philosophy. In G. L. SCHIEWER, J. ALTARRIBA & B. C. NG, Éds., *Language and emotion : An international handbook*, volume 1, p. 54–72. Berlin : de Gruyter.

BOSTAN L. A. M., KIM E. & KLINGER R. (2020). GoodNewsEveryone : A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 1554–1566, Marseille, France : European Language Resources Association.

BOSTAN L.-A.-M. & KLINGER R. (2018). An analysis of annotated corpora for emotion classification in text. In E. M. BENDER, L. DERCZYNSKI & P. ISABELLE, Éds., *Proceedings of the 27th International Conference on Computational Linguistics*, p. 2104–2119, Santa Fe, New Mexico, USA : Association for Computational Linguistics.

BROWN T., MANN B., RYDER N., SUBBIAH M., KAPLAN J. D., DHARIWAL P., NEELAKANTAN A., SHYAM P., SASTRY G., ASKELL A., AGARWAL S., HERBERT-VOSS A., KRUEGER G., HENIGHAN T., CHILD R., RAMESH A., ZIEGLER D., WU J., WINTER C., HESSE C., CHEN M., SIGLER E., LITWIN M., GRAY S., CHESS B., CLARK J., BERNER C., MCCANDLISH S., RADFORD A., SUTSKEVER I. & AMODEI D. (2020). Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, p. 1877–1901 : Curran Associates, Inc.

BUECHEL S. & HAHN U. (2017). EmoBank : Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In M. LAPATA, P. BLUNSOM & A. KOLLER, Éds., *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 2, Short Papers*, p. 578–585, Valencia, Spain : Association for Computational Linguistics.

BUECHEL S., MODERSOHN L. & HAHN U. (2021). Towards label-agnostic emotion embeddings. In M.-F. MOENS, X. HUANG, L. SPECIA & S. W.-T. YIH, Éds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 9231–9249, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.728.

CAMPAGNANO C., CONIA S. & NAVIGLI R. (2022). SRL4E – Semantic Role Labeling for Emotions : A unified evaluation framework. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 4586–4601, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.314.

CASEL F., HEINDL A. & KLINGER R. (2021). Emotion recognition under consideration of the emotion component process model. In K. EVANG, L. KALLMEYER, R. OSSWALD, J. WASZCZUK & T. ZESCH, Éds., *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, p. 49–61, Düsseldorf, Germany : KONVENS 2021 Organizers.

CORTAL G. (2024). Sequence-to-sequence language models for character and emotion detection in dream narratives. In N. CALZOLARI, M.-Y. KAN, V. HOSTE, A. LENCI, S. SAKTI & N. XUE, Éds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, p. 14717–14728, Torino, Italia : ELRA and ICCL.

CORTAL G., FINKEL A., PAROUBEK P. & YE L. (2022). Natural language processing for cognitive analysis of emotions. In *Semantics, Memory, and Emotion 2022*, Paris, France. HAL : hal-03805702.

CORTAL G., FINKEL A., PAROUBEK P. & YE L. (2023). Emotion recognition based on psychological components in guided narratives for emotion regulation. In S. DEGAETANO-ORTLIEB, A. KAZANTSEVA, N. REITER & S. SZPAKOWICZ, Éds., *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, p. 72–81, Dubrovnik, Croatia : Association for Computational Linguistics. DOI : 10.18653/v1/2023.latechclfl-1.8.

COSMIDES L. & TOOBY J. (2000). Evolutionary psychology and the emotions. In M. LEWIS & J. M. HAVILAND-JONES, Éds., *Handbook of emotions*, p. 91–115. New York : Guilford Press, 2nd édition. Publisher : Citeseer.

CREISSEN S. & BLANC N. (2017). Quelle représentation des différentes facettes de la dimension émotionnelle d'une histoire entre l'âge de 6 et 10 ans ? Apports d'une étude multimédia. *Psychologie Française*, **62**(3), 263–277. DOI : 10.1016/j.psfr.2015.07.006.

DARWIN C. (1872). *The expression of the emotions in man and animals*. London : John Murray.

DAVIDSON D., LUO Z. & BURDEN M. J. (2001). Children's recall of emotional behaviours, emotional labels, and nonemotional behaviours : Does emotion enhance memory ? *Cognition and Emotion*, **15**(1), 1–26. DOI : 10.1080/0269993004200105.

DE BRUYNE L., DE CLERCQ O. & HOSTE V. (2020). An emotional mess ! deciding on a framework for building a Dutch emotion-annotated corpus. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 1643–1651, Marseille, France : European Language Resources Association.

DEMSZKY D., MOVSHOVITZ-ATTIAS D., KO J., COWEN A., NEMADE G. & RAVI S. (2020). GoEmotions : A dataset of fine-grained emotions. In D. JURAFSKY, J. CHAI, N. SCHLUTER & J. TETREAULT, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 4040–4054, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.372.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In J. BURSTEIN, C. DORAN & T. SOLORIO, Éds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

EKMAN P. (1999). Basic emotions. In T. DALGLEISH & M. J. POWER, Éds., *Handbook of cognition and emotion*, p. 45–60. Chichester : John Wiley & Sons Ltd.

EKMAN P. & FRIESEN W. V. (1971). Constants across cultures in the face and emotion. *Journal of personality and social psychology*, **17 2**, 124–9.

ETIENNE A., BATTISTELLI D. & LECORVÉ G. (2022). A (psycho-)linguistically motivated scheme for annotating and exploring emotions in a genre-diverse corpus. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 603–612, Marseille, France : European Language Resources Association.

FOPPOLO F. & MAZZAGGIO G. (2024). Conversational Implicature and Communication Disorders. In M. J. BALL, N. MÜLLER & E. SPENCER, Éds., *The Handbook of Clinical Linguistics, Second Edition*, p. 15–27. Wiley, 1 édition. DOI : 10.1002/9781119875949.ch2.

GANDHI K., FRÄNKEN J.-P., GERSTENBERG T. & GOODMAN N. D. (2023). Understanding Social Reasoning in Language Models with Language Models. DOI : 10.48550/arXiv.2306.15448.

GREEN M. (2007). *Self-expression*. Oxford : Oxford University Press.

GRICE H. P. (1957). Meaning. *The Philosophical Review*, **66**(3), 377–388.

GRICE H. P. (1975). Logic and conversation. In *Speech acts*, p. 41–58. Leiden : Brill.

GRICE H. P. (1989). *Studies in the way of words*. Cambridge (MA) : Harvard University Press.

HEIM I. & KRATZER A. (1998). *Semantics in generative grammar*. Hoboken : Wiley. Google-Books-ID : jAvR2DB3pPIC.

HEINTZ C. & SCOTT-PHILLIPS T. (2023). Expression unleashed : The evolutionary & cognitive foundations of human communication. *Behavioral and Brain Sciences*, **46**, E1. type : article, DOI : 10.31234/osf.io/mcv5b.

HOFMANN J., TROIANO E., SASSENBERG K. & KLINGER R. (2020). Appraisal theories for emotion classification in text. In D. SCOTT, N. BEL & C. ZONG, Éds., *Proceedings of the 28th International Conference on Computational Linguistics*, p. 125–138, Barcelona, Spain (Online) : International Committee on Computational Linguistics. DOI : 10.18653/v1/2020.coling-main.11.

HOLTERMAN B. & VAN DEEMTER K. (2023). Does ChatGPT have Theory of Mind ? arXiv :2305.14020 [cs].

IZARD C. E. (1992). Basic Emotions, Relations Among Emotions, and Emotion-Cognition Relations. *Psychological Review*, **99**(3), 561–565.

KIM E. & KLINGER R. (2018). Who Feels What and Why ? Annotation of a Literature Corpus with Semantic Roles of Emotions. In *Proceedings of the 27th International Conference on Computational Linguistics*, p. 1345–1359, Santa Fe, New Mexico, USA : Association for Computational Linguistics.

KIM E. & KLINGER R. (2019). An analysis of emotion communication channels in fan-fiction : Towards emotional storytelling. In F. FERRARO, T.-H. K. HUANG, S. M. LUKIN & M. MITCHELL, Éds., *Proceedings of the Second Workshop on Storytelling*, p. 56–64, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/W19-3406.

KIM H., SCLAR M., ZHOU X., BRAS R. L., KIM G., CHOI Y. & SAP M. (2023). FANToM : A Benchmark for Stress-testing Machine Theory of Mind in Interactions. DOI : 10.48550/arXiv.2310.15421.

KLINGER R. (2023). Where are We in Event-centric Emotion Analysis ? Bridging Emotion Role Labeling and Appraisal-based Approaches. In Y. ELAZAR, A. ETTINGER, N. KASSNER, S. RUDER & N. A. SMITH, Éds., *Proceedings of the Big Picture Workshop*, p. 1–17, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.bigpicture-1.1.

KOSINSKI M. (2023). Theory of Mind Might Have Spontaneously Emerged in Large Language Models. arXiv :2302.02083 [cs], DOI : 10.48550/arXiv.2302.02083.

LANGLEY C., CIRSTEA B. I., CUZZOLIN F. & SAHAKIAN B. J. (2022). Theory of Mind and Preference Learning at the Interface of Cognitive Science, Neuroscience, and AI : A Review. *Frontiers in Artificial Intelligence*, **5**.

LAZARUS R. S. (1991). Progress on a cognitive-motivational-relational theory of emotion. *American psychologist*, **46**(8), 819.

LEE S. Y. M., CHEN Y. & HUANG C.-R. (2010). A text-driven rule-based system for emotion cause detection. In D. INKPEN & C. STRAPPARAVA, Éds., *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, p. 45–53, Los Angeles, CA : Association for Computational Linguistics.

MA Z., SANSOM J., PENG R. & CHAI J. (2023). Towards A Holistic Landscape of Situated Theory of Mind in Large Language Models. DOI : 10.48550/arXiv.2310.19619.

MAYER J. D., ROBERTS R. D. & BARSADE S. G. (2008). Human Abilities : Emotional Intelligence. *Annual Review of Psychology*, **59**(1), 507–536. DOI : 10.1146/annurev.psych.59.103006.093646.

MICHELI R. (2013). Esquisse d'une typologie des différents modes de sémiotisation verbale de l'émotion. *Semen*, (35). DOI : 10.4000/semen.9795.

MICHELI R. (2014). *Les émotions dans les discours*. De Boeck Supérieur. DOI : 10.3917/dbu.mchel.2014.01.

MITCHELL M. & KRAKAUER D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, **120**(13), e2215907120. Publisher : Proceedings of the National Academy of Sciences, DOI : 10.1073/pnas.2215907120.

MOORS A. (2022). *Demystifying emotions : A Typology of theories in psychology and philosophy*. Cambridge, cambridge university press édition.

MOORS A., ELLSWORTH P. C., SCHERER K. R. & FRIJDA N. H. (2013). Appraisal theories of emotion : state of the art and future development. *Emotion Review*, **5**(2), 119–124. Publisher : Sage Publications Sage UK : London, England.

OBERLÄNDER L., REICH K. & KLINGER R. (2020). Experiencers, Stimuli, or Targets : Which Semantic Roles Enable Machine Learning to Infer the Emotions ? *arXiv :2011.01599 [cs]*.

O'CONNOR P. J., HILL A., KAYA M. & MARTIN B. (2019). The measurement of emotional intelligence : A critical review of the literature and recommendations for researchers and practitioners. *Frontiers in psychology*, **10**, 1116. Publisher : Frontiers.

PAECH S. J. (2024). EQ-Bench : An Emotional Intelligence Benchmark for Large Language Models. DOI : 10.48550/arXiv.2312.06281.

PANKSEPP J. (1998). *Affective neuroscience : the foundations of human and animal emotions*. New York : Oxford University Press.

PLUTCHIK R. (2001). The Nature of Emotions : Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, **89**(4), 344–350.

RUSSELL J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, **39**(6), 1161. Publisher : American Psychological Association.

RUSSELL J. A. (2003). Core affect and the psychological construction of emotion. *Psychological review*, **110**(1), 145. Publisher : American Psychological Association.

RUSSELL J. A. (2009). Emotion, core affect, and psychological construction. *Cognition and Emotion*, **23**(7), 1259–1283. DOI : 10.1080/02699930902809375.

SCARANTINO A. (2017). How to do things with emotional expressions : The theory of affective pragmatics. *Psychological Inquiry*, **28**(2-3), 165–185. Publisher : Taylor & Francis.

SCHACHTER S. & SINGER J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological review*, **69**(5), 379. Publisher : American Psychological Association.

SCHERER K. R. (2007). Componential emotion theory can inform models of emotional competence. Publisher : Oxford University Press.

SCHERER K. R. (2022). Theory convergence in emotion science is timely and realistic. *Cognition and Emotion*, **36**(2), 154–170. DOI : 10.1080/02699931.2021.1973378.

SCHERER K. R. & MOORS A. (2019). The emotion process : event appraisal and component differentiation. *Annual Review of Psychology*, **70**, 719–745. Publisher : Annual Reviews.

SCHLENKER P. (2016). The semantics-pragmatics interface. In M. ALONI & P. DEKKER, Éds., *The Cambridge Handbook of Formal Semantics*, p. 664–727. Cambridge : Cambridge University Press.

SHAPIRA N., LEVY M., ALAVI S. H., ZHOU X., CHOI Y., GOLDBERG Y., SAP M. & SHWARTZ V. (2023). Clever Hans or Neural Theory of Mind ? Stress Testing Social Reasoning in Large Language Models. arXiv :2305.14763 [cs], DOI : 10.48550/arXiv.2305.14763.

SPERBER D. & WILSON D. (1995). *Relevance : Communication and cognition*. Oxford and Cambridge (MA) : Blackwell, 2nd edition édition.

STALNAKER R. (2002). Common ground. *Linguistics and philosophy*, **25**(5/6), 701–721.

STOJNIĆ G., GANDHI K., YASUDA S., LAKE B. M. & DILLON M. R. (2023). Commonsense psychology in human infants and machines. *Cognition*, **235**, 105406. DOI : https ://doi.org/10.1016/j.cognition.2023.105406.

TEIGEN K. H. (2008). Is a sigh "just a sigh" ? Sighs as emotional signals and responses to a difficult task. *Scandinavian journal of Psychology*, **49**(1), 49–57. Publisher : Wiley Online Library.

TOMKINS S. (1962). *Affect imagery consciousness*, volume Volume I : The positive affects. New York : Springer.

TROIANO E., OBERLÄNDER L. & KLINGER R. (2022). Dimensional Modeling of Emotions in Text with Appraisal Theories : Corpus Creation, Annotation Reliability, and Prediction. *Computational Linguistics*, p. 1–71. DOI : 10.1162/coli_a_00461.

TROTT S., JONES C., CHANG T., MICHAELOV J. & BERGEN B. (2022). Do Large Language Models know what humans know ? *arXiv preprint arXiv :2209.01515*.

ULLMAN T. (2023). Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks. arXiv :2302.08399 [cs].

VLEMINCX E., VAN DIEST I., DE PEUTER S., BRESSELEERS J., BOGAERTS K., FANNES S., LI W. & VAN DEN BERGH O. (2009). Why do you sigh ? Sigh rate during induced stress and relief. *Psychophysiology*, **46**(5), 1005–1013. Publisher : Wiley Online Library.

WANG X., LI X., YIN Z., WU Y. & LIU J. (2023). Emotional intelligence of Large Language Models. *Journal of Pacific Rim Psychology*, **17**, 18344909231213958. DOI : 10.1177/18344909231213958.

WEGGE M., TROIANO E., OBERLÄNDER L. & KLINGER R. (2023). Experiencer-Specific Emotion and Appraisal Prediction. DOI : 10.48550/arXiv.2210.12078.

WHARTON T. (2003). Natural pragmatics and natural codes. *Mind & language*, **18**(5), 447–477. Publisher : Wiley Online Library.

WHARTON T. (2016). That bloody so-and-so has retired : Expressives revisited. *Lingua*, **175**, 20–35. Publisher : Elsevier.

WHARTON T., BONARD C., DUKES D., SANDER D. & OSWALD S. (2021). Relevance and emotion. *Journal of Pragmatics*, **181**, 259–269.

WILSON D. & SPERBER D. (2006). Relevance theory. In L. HORN, Éd., *The Handbook of pragmatics*. Oxford : Blackwell.

ZHAN H., ONG D. & LI J. J. (2023). Evaluating Subjective Cognitive Appraisals of Emotions from Large Language Models. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2023*, p. 14418–14446, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-emnlp.962.

ZHANG W., DENG Y., LIU B., PAN S. J. & BING L. (2023). Sentiment Analysis in the Era of Large Language Models : A Reality Check. DOI : 10.48550/arXiv.2305.15005.

ZHOU P., MADAAN A., POTHARAJU S. P., GUPTA A., MCKEE K. R., HOLTZMAN A., PUJARA J., REN X., MISHRA S., NEMATZADEH A., UPADHYAY S. & FARUQUI M. (2023). How FaR Are Large Language Models From Agents with Theory-of-Mind ? DOI : 10.48550/arXiv.2310.03051.

ZWAAN R. A. & RADVANSKY G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, **123**(2), 162–185. DOI : 10.1037/0033-2909.123.2.162.