Enjeux liés à la détection de l'ironie

Samuel Laperle

Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles
Lille, France, 28 juin au 2 juillet 2021

Enjeux liés à la détection de l'ironie

Samuel Laperle[1]
(1) Université du Québec à Montréal, Montréal, Canada
laperle.samuel@courrier.uqam.ca

RÉSUMÉ

L'ironie verbale est un type de discours difficile à détecter automatiquement. En créant des ponts entre les recherches en linguistique et en informatique sur cette question, il est possible de souligner des caractéristiques importantes permettant de faciliter ce type de tâche. Dans cet article, il sera question du rapport entre la définition de ce phénomène et son adéquation avec l'élaboration de corpus d'entraînement..

ABSTRACT

Challenges of automatic irony detection

Verbal irony is a type of speech that is difficult to detect. By building bridges between linguistics and computer science research on this question, it is possible to highlight important characteristics that facilitate this type of task. This paper will focus on the relationship between the definition of this type of discourse and its adequacy with the development of training corpus.).

MOTS-CLÉS : Ironie, détection automatique, linguistique.

KEYWORDS: Irony, automatic detection, linguistics.

1 Introduction

La question de la détection de l'ironie verbale par des systèmes informatiques est un enjeu économique et marketing ayant généralement comme objectif d'améliorer les algorithmes d'analyse de sentiment qui permettent de traiter d'énormes quantités de données issues des médias sociaux (Eke *et al.*, 2020; Strapparava *et al.*, 2011). L'ironie est un type de discours compliqué à définir. De ce fait, les algorithmes proposés pour tenter de détecter adéquatement ce type de discours doivent jongler avec des variables difficilement implémentables. Par conséquent, on retrouve une rupture entre le travail linguistique sur la caractérisation de l'ironie verbale et ses traitements computationnels. Ce travail tentera de mettre en relief l'écart présent dans ces deux cadres de recherche concernant cet enjeu.

Concrètement, la question centrale de ce travail est de déterminer au niveau théorique les limitations des méthodes computationnelles de détection de l'ironie verbale. D'abord, il sera question de faire une présentation des différentes définitions de l'ironie verbale proposée en linguistique. Ensuite, certains types de méthodes de détection automatique proposées en TAL seront présentés. Avec ces deux perspectives, il sera possible de les confronter pour déterminer, sur la base des théories linguistiques, les différents enjeux propres à la définition de l'ironie. Par exemple, il sera question de la confusion entre les concepts de sarcasme et d'ironie, de la construction des corpus d'entraînement et de la polarité des énoncés ironiques.

2 Définir l'ironie

Pour Kerbrat-Orecchioni (1978), l'ironie serait un acte illocutoire permettant de se moquer d'une cible. Elle qualifie plus particulièrement l'ironie verbale comme étant « la mise en relation entre deux niveaux sémantiques littéral et figuratif attachés à une même séquence signifiante ». De façon similaire, Grice (1975) caractérise l'ironie comme étant une transgression de la maxime de qualité qui stipule qu'un locuteur ne devrait pas dire ce qu'il croit être faux.

Bien que ces définitions soulèvent des points primordiaux, Wilson & Sperber (1992) soulignent certaines de leurs limitations. Concrètement, elles négligent trois types d'ironie et surgénéralisent sur des types de discours non ironiques, soit respectivement les litotes ironiques, les citations ironiques, les interjections ironiques et les mensonges éhontés. De ce fait, Wilson & Sperber (1992) quant à eux, définissent l'ironie par la théorie de l'écho, qui s'appuie sur la distinction entre mention et usage. Pour eux, l'ironie serait un type de citation indirecte transmettant l'attitude d'un locuteur concernant une cible. Parallèlement, (Clark & Gerrig, 1984) définissent l'ironie par la théorie du faire-semblant. Cette proposition met l'accent sur la relation entre le locuteur et l'interlocuteur plutôt que sur les processus interprétatifs de ce dernier. Pour eux, le locuteur joue un rôle et l'interlocuteur doit être en mesure de déceler la mascarade.

Les propositions de Grice (1975), Wilson & Sperber (1992) et Clark & Gerrig (1984) s'imposent comme des incontournables. Elles négligent néanmoins certains points importants. On n'y retrouve aucune référence à l'aspect évaluatif caractéristique de l'ironie.

Pour Alba-Juez & Attardo (2014), le spectre d'attitudes pouvant être transmises par l'ironie verbale serait large. Par exemple, on peut retrouver des énoncés ironiques qui transmettent une attitude négative comme en (1), une attitude positive comme en (2) ou, même, une attitude neutre comme en (3).

1. Quelle belle partie ! (exclamé suite à une défaite)
2. Quelle triste partie ! (exclamé suite à une victoire)
3. C'était une partie. (exclamé suite à une partie s'étant rapidement terminée)

De plus, un énoncé peut transmettre différentes attitudes à différentes cibles à la fois. Prouvant ce point, Alba-Juez & Attardo (2014) proposent l'exemple (4) où une actrice doutant de son talent dirait à son ami :

4. A : Je suis un échec. Je ne réussirai jamais à percer dans le monde du théâtre. Je suis une actrice médiocre.

Pour qu'après la réception d'un prix soulignant son talent, cet ami lui réponde (5) :

5. Félicitations, mon amie ! Tu es une actrice médiocre. Je ne sais pas comment ils ont pu te donner ce prix !

D'un côté, on note que le locuteur en (5) transmet une évaluation positive du talent d'actrice de son amie tout en émettant une appréciation négative du jugement négatif qu'elle s'auto-imposait. Ces variations dépendent des cibles visées par le locuteur et du contexte de savoir partagé entre les individus.

Ce contexte joue aussi un rôle primordial dans la production et l'interprétation d'un énoncé ironique. Cette capacité ne dépend pas seulement de la relation entre un mot ou une expression avec l'ensemble d'une situation ou d'un texte (Martini *et al.*, 2018). Elle dépend aussi de notre faculté à nous mettre à la place d'autrui (Nilsen *et al.*, 2011). Les objectifs communicationnels varient et s'adaptent aux informations que nous collectons à travers nos conversations. Ainsi, comme le rapporte Gibbs (2000), nous sommes plus prompts à utiliser l'ironie dans des contextes sociaux où nous connaissons bien nos interlocuteurs, car ces derniers arriveraient plus facilement à déterminer adéquatement les attitudes véritables que nous tentons de communiquer.

Ces différents travaux soulignent des caractéristiques importantes de l'ironie verbale que devront prendre en compte les algorithmes de détection automatique de ce type de discours.

3 Détection l'ironie

3.1 Méthodes à base de règles

Les informations langagières présentes sur Internet étant textuelles, la plupart des algorithmes fonctionnent en se basant principalement sur la présence d'ironie verbale exprimée sous cette modalité. Par exemple, Kreuz & Caucci (2007) proposent d'évaluer les expressions considérées comme étant stéréotypiques de l'ironie verbale pour les ajouter à des algorithmes de détection. Pour ce faire, des participants devaient lire des énoncés en anglais préalablement collectés par les chercheurs et évaluer leur niveau d'ironie. En collectant ainsi environ 100 énoncés, ils notent certains traits lexicaux plus fréquents lors de l'expression de ce type de discours comme des interjections, des expressions convenues (thanks alot, good job), des questions rhétoriques et de la répétition. Allant dans le même sens, Bouazizi & Ohtsuki (2015) proposent d'utiliser comme marqueurs lexicaux de l'ironie la présence de mots peu communs en termes de fréquence ou la présence d'énoncés, eux aussi, considérées comme prototypiques de ce type de discours (P.ex. : « love [pronoun] when » ou « [pronoun] be [adverb] funny »). Ils justifient ce choix en soulignant que l'ironie peut être utilisée comme une façon d'éviter de donner une réponse claire à une question. Se faisant, le locuteur emploie des phrases plus longues, plus complexes et, ainsi, utilise des expressions moins fréquentes. Ce raisonnement fait écho à une des raisons possibles derrière l'utilisation de l'ironie proposée par Jorgensen (1996). Pour ce dernier, ce type de discours pourrait être utilisé pour transmettre une critique sans qu'elle soit perçue comme étant trop directe ou négative.

Comme Attardo (2000) le mentionne, l'ironie est un type de discours pragmatique. De ce fait, se baser sur des informations exclusivement lexicales peut s'avérer problématique. Conséquemment, Joshi *et al.* (2015) ont mis au point un algorithme de détection basé sur la présence d'incongruité textuelle explicite ou implicite. Dans le premier cas, on peut soupçonner qu'un énoncé soit sarcastique en se basant sur un lexique (Lingpipe SA system (Alias-I, 2014)) s'il comporte deux mots de valences différentes comme dans l'énoncé « j'aime être malade », où « aimer » porte une valence positive et

« malade » en porte une qui est négative. Cette catégorisation ne peut pas s'appliquer dans des cas d'incongruité implicite comme dans une expression du genre « J'aime tellement ce repas que je l'ai donné à mon chien ». Dans cette dernière, les éléments lexicaux « donner à son chien » ne sont pas explicitement négatifs et ne créent pas d'incongruité avec l'expression positive « j'aime ». Pour rendre compte de ce type d'expressions convenues, les chercheurs proposent de créer un système à base de règles impliquant divers types de phrases typiques portant une polarité implicite. En plus de ces

deux paramètres, l'architecture des chercheurs tient en compte des traits lexicaux soit des unigrames, la présence de lettres majuscules, d'émoticônes ou de rires et de ponctuations particulières comme des points d'exclamation excessifs. Pour tester ce système, ils ont utilisé trois bases de données. La première, Tweet-A, contient 5208 tweets non sarcastiques et 4170 tweets sarcastiques collectés grâce à la présence de #sarcasm. Le deuxième, Tweet-B, contient 2278 tweets non sarcastiques et 506 tweets sarcastiques. Ces derniers sont tirés du travail préalable de Riloff *et al.* (2013). Le troisième jeu de données, Discussion-A, provient d'un corpus préalablement créé par Walker *et al.* (2012). Au total, il contient 1502 messages littéraux et 752 messages sarcastiques. Parmi ceux-ci, Joshi *et al.* (2015) en gardent 752 littéraux et 752 sarcastiques. Ces messages sont issus de forums de discussion en ligne.

3.2 Méthodes à base d'apprentissage machine

en termes d'apprentissage machine, Poria *et al.* (2016) proposent un modèle basé sur un réseau neuronal convolutif (RNC) pré entraîné pour extraire des traits concernant les sentiments, les émotions et la personnalité du locuteur. Concrètement, selon Poria *et al.* (2016), grâce à ce type de réseau, il est possible de former un vecteur englobant l'ensemble de traits locaux d'un énoncé lui permettant de créer une représentation adéquate du contexte lexical. Ils ont testé leur algorithme sur trois jeux de données. Le premier, créé par Ptáček *et al.* (2014), contient un nombre balancé de tweets sarcastiques (50 000) et non sarcastiques (50 000) en anglais. Le deuxième, aussi créé par Ptáček *et al.* (2014), contient un nombre déséquilibré de tweets sarcastiques (25 000) et de tweets non sarcastiques(75 000) en anglais. Le troisième, issu du site internet The Sarcasm Detector, [1] contient 20 000 tweets sarcastiques et 100 000 tweets non sarcastiques en anglais. Parmi ceux-ci, Poria et al. (2016) en ont collecté de façon aléatoire 10 000 sarcastiques et 20 000 non sarcastiques. Sur l'ensemble de ces jeux de données, ils arrivent à des F1 supérieurs à 0.9 de détection de l'ironie verbale.

De leur côté, Ghosh & Veale (2016) utilisent une conjonction de différents types de réseaux neuronaux, soit un réseau de neurones composé d'un RNC, suivi d'un réseau de neurones récurrent (RNR), d'un long-short term memory (LSTM) et d'un réseau de neurones profond (RNP). La première couche de cette architecture est celle des données langagières contenues dans un tweet qui est vectorisé. Ensuite, le résultat de ce traitement passe par une couche du RNC qui permettrait d'extraire des séquences de mots importants pour la détection en éliminant les variations de fréquences. Ce processus fournit à la couche du LSTM les données adéquates. Ce dernier serait en mesure de créer une représentation sémantique grâce à la présence d'un module temporel permettant d'entreposer des informations contextuelles. Finalement, les données sont traitées par le RNP. Pour tester leur modèle, ils ont utilisé deux jeux de données préalablement créés par Tsur *et al.* (2010) et Riloff *et al.* (2013). Le premier jeu de données comprend 471 critiques en anglais de produit en vente sur le site Amazon classées comme étant sarcastique et de 5020 critiques non sarcastiques. Le second jeu de données est composé de 693 tweets sarcastiques et 2 307 tweets non sarcastiques en anglais L'ensemble de ces couches arrive à un score de précision de 0.919, un score de rappel de 0.923 et un score-f de 0.921. Malgré ces bons résultats, Ghosh & Veale (2016) notent tout de même que leur système de détection n'est pas en mesure de classifier une expression ironique comme

« Thank God it's Monday ! » bien qu'il soit en mesure de détecter l'ironie dans une expression « I just love Mondays ! ».

1. thesarcasmdetector.com

4 Limitations

4.1 Problèmes de définition : ironie ou sarcasme

Un des enjeux centraux relatifs à la détection automatique de l'ironie se situe au niveau de la définition même de ce concept. La majorité des travaux concernant ce sujet semble éviter d'aborder cette question. Paradoxalement, la façon dont on choisit d'appréhender ce problème influence nécessairement les méthodes employées pour arriver à une détection adéquate de ce type de discours. Lorsqu'on en trouve, elles sont essentiellement superficielles. Par exemple, Carvalho *et al.* (2009) caractérisent l'ironie verbale ainsi : *as the rhetorical process of intentionally using words or expressions for uttering a meaning different (usually the opposite) from the one they have when used literally*. Dans le même sens, Van Hee (2017) propose de décrire l'ironie comme "*an evaluative expression whose polarity (i.e. positive, negative) is inverted between the literal and the intended evaluation, resulting in an incongruence between the literal evaluation and its context*". Ces définitions ressemblent à celles proposées par Kerbrat-Orecchioni (1978) et Grice (1975). Par conséquent, percevoir l'ironie verbale comme étant une négation du sens littéral ou comme un rapport de polarité inverse néglige un ensemble important d'énoncés ironiques s'exprimant autrement.

Par ailleurs, dans la littérature sur la détection automatique de l'ironie, on retrouve fréquemment une juxtaposition de ce terme avec celui désignant le sarcasme. On considère parfois ces deux concepts comme étant interchangeables. Par exemple, Davidov *et al.* (2010) écrivent : *sarcasm (also known as verbal irony) is a sophisticated form of speech act in which the speakers convey their message in an implicit way*". D'autres fois, on présente l'ironie comme étant une supracatégorie pouvant contenir des énoncés sarcastiques (sans expliquer comment) comme dans cet exemple (Farías *et al.*, 2016) : "*irony is here considered an umbrella term that also covers sarcasm*".

Au niveau de cette distinction difficile à faire, Van Hee (2017) rapporte que "*researchers do not differentiate between irony and sarcasm [because they observed] a shift in meaning between the two terms. Over time, the term 'sarcasm' seems to have gradually replaced what was previously designed by 'irony' (Nunberg, 2001)*".

Les points apportés par (Van Hee, 2017) sont valides. Toutefois, ils font fi de tout un pan de la littérature qualifiant le sarcasme, qui possède une connotation plus négative que le terme « ironie » (Kreuz & Caucci, 2007). De plus, on n'évalue pas comment ces problèmes affectent indéniablement les processus de détection. Il existe un rapport évident entre la définition du phénomène que l'on souhaite détecter et la façon prévue pour y arriver. De ce fait, une attention particulière devrait être accordée à cette question.

Goddard (2018) souligne ce type de problème circulaire au niveau de la définition terminologique d'un concept :

> (i) One starts with ordinary English words, poorly defined or undefined, then (ii) "technicalizes" them and extends their range, often making some formal adjustments along the way (iii) Subsequently [..], different scholars begin to employ the terms, often using them in slightly different ways from the original authors. (iv) Scholarly debate begins about what the new terms mean or should mean.

S'il est indéniable qu'« ironie » et « sarcasme » sont liés, il est faux de dire que ces deux termes désignent la même chose. En effet, par exemple, en termes d'usage, on pourra dire d'une situation

qu'elle est ironique, mais jamais sarcastique. Allant dans ce sens, Sulis *et al.* (2016) ont évalué un corpus contenant 10 000 tweets pour évaluer les différences caractéristiques entre les tweets contenant l'expression #irony, #sarcasm et #not. Parmi ces divergences notables, ils rapportent qu'au niveau affectif, les tweets contenant l'expression #irony contiennent moins de mots associés à la joie et l'anticipation que les tweets contenant l'expression #sarcasm. Au contraire, le #irony serait plus associé à des sentiments comme la colère, la tristesse et la peur. En utilisant des lexiques (Affective Norms for English Words, Dictionary of Affective Language) recensant des valeurs comme le niveau d'imagerie, d'activation générale et émotive et de dominance, Sulis *et al.* (2016) soulignent aussi que les énoncés comportant le mot clé #irony seraient plus subtils que ceux contenant #sarcasm. Se faisant, il est inadéquat d'utiliser les termes sarcasme et ironie de façon interchangeable. Si ces concepts sont difficiles à distinguer, on remarque néanmoins des différences concrètes au niveau de leurs usages.

4.2 La polarité d'un énoncé

Comme mentionné plus haut, la polarité d'un énoncé ironique n'est pas systématiquement négative. Alba-Juez & Attardo (2014) ont démontré les différents cas de figure où un énoncé ironique peut avoir une polarité positive, négative et neutre. De plus, parfois, un même énoncé peut être positif envers une cible tout en étant négatif envers une autre. Ainsi, des algorithmes comme celui de Joshi *et al.* (2015) présenté plus haut négligeront nécessairement des énoncés ironiques comme celui présenté en (5) dans la section 2.

Les énoncés ironiques neutres sont particulièrement problématiques de par leur nature. Alba-Juez & Attardo (2014) les définissent comme étant des énoncés qui ne visent, ne critiquent et qui ne vantent personne ni rien de particulier. Bien que ce type d'énoncé comporte une certaine valeur évaluative, cette dernière est loin d'être positive ou négative. La motivation derrière ce type d'énoncé serait avant tout de faire preuve d'humour. Dans l'exemple (6), il est difficile de catégoriser adéquatement l'attitude voulant être transmise.

6. Ce courriel est plus long qu'à l'habitude parce que je n'avais pas le temps d'en écrire un plus court.

Ainsi, pour Alba-Juez & Attardo (2014), l'ironie ne dépend pas de son caractère évaluatif, mais plutôt de l'inférence contradictoire qui en découle. Se faisant, en (6), c'est de la contradiction entre la longueur du courriel et l'excuse en général qui permet l'émergence d'une interprétation ironique de l'énoncé.

4.3 Corpus

Généralement, les corpus d'énoncés ironiques se construisent de deux façons (Farías *et al.*, 2016). D'un côté, on retrouve les énoncés ironiques explicitement indiqués comme tels par les locuteurs les produisant. Concrètement, sur Twitter, on aurait des tweets contenant des hashtags comme #sarcasm, #sarcastic ou #nottrue. Derrière l'élaboration de ce type de corpus, on prend pour acquis que le locuteur serait le mieux placé pour savoir si un énoncé qu'il produit est ironique ou non. De l'autre côté, on trouve des corpus basés sur des accords interjuges. La qualité de ces derniers dépend évidemment de facteurs internes aux juges ayant catégorisé les éléments constitutifs du corpus.

Idéalement, les corpus construits grâce à la présence de hashtag passent par la suite par une équipe d'annotateurs jugeant si les tweets sont réellement perçus comme étant sarcastiques. Ce deuxième tri dépend fortement de la définition de sarcasme ou d'ironie utilisée par l'équipe de chercheurs.

Au-delà de ces aspects, un autre problème dans les études mentionnées plus haut est l'inadéquation entre les définitions de l'ironie que certains chercheurs proposent et les corpus d'énoncés ironiques utilisés.

Par exemple, Kreuz & Caucci (2007), en tentant de vérifier s'il existe des termes lexicaux permettant de caractériser l'ironie, ne proposent pas de définir ce terme. Néanmoins, ils construisent tout de même un corpus d'analyse en sélectionnant sur Google Books des énoncés suivis de l'expression "said sarcastically" en prenant soin de retirer cette expression avant de les présenter à leurs participants. Donc, on ne sait pas sur quels critères se basent ces derniers pour émettre leurs jugements. De plus, comme le mentionnent Kreuz & Caucci (2007), il est possible qu'un auteur utilise l'expression "said sarcastically" comme synonyme de "said jokingly" ou "said angrily", entraînant nécessairement des biais importants concernant le type d'énoncé qu'on retrouve dans un tel corpus.

Dans le même ordre d'idée, dans le travail de Bouazizi & Ohtsuki (2015), on définit le sarcasme comme étant "*a special form of irony by which the person conveys implicit information, usually the opposite of what is said, within the message he transmits*". Cependant, on ne sait pas comment cette conceptualisation se traduit au niveau de l'élaboration des corpus utilisés. Dans ce cas-ci, ces derniers sont composés de tweets comportant l'expression "#sarcasm" pour, par la suite, être revérifiés par les chercheurs. Outre leur définition du sarcasme mentionné plus haut, il nous est donc impossible de savoir sur quelles bases sont fondés ces critères de sélection.

Cette relation entre la définition du concept étudié, l'élaboration d'un algorithme de détection automatique et le choix des éléments constituant les jeux de données permettant d'extraire des traits ou de tester leur justesse est particulière. Si l'élaboration de l'algorithme de détection de l'ironie dépend de la caractérisation de ce type de discours par les chercheurs, l'élaboration du corpus test doit nécessairement refléter cet aspect. Dans le cas contraire, on peut s'attendre à la présence de faux positifs ou de faux négatifs.

4.4 Problèmes concernant les algorithmes d'apprentissage machine

Les approches se basant sur les techniques d'apprentissage machine nous offrent des résultats variables concernant la détection automatique de l'ironie. Comme pour les méthodes à base de règles, elles réussissent généralement bien dans les conditions tests. Néanmoins, de par la nature de ces méthodes, il nous est impossible de savoir concrètement ce que ces algorithmes "apprennent". Aussi, on sait peu de choses sur la possibilité de généraliser ces apprentissages à un milieu plus écologique. À ce sujet, Wallace (2015) dit :

> Current machine learning methods rely too heavily on shallow, unstructured, syntactic modelling of text to consistently discern ironic intent. Irony detection is an interesting machine learning problem because, in contrast to most text classification tasks, it requires a semantics that cannot be inferred directly from word counts over documents alone.

En effet, récemment, dans une tâche partagée axée sur la détection de sarcasme, Ghosh *et al.* (2020) évaluent différents modèles proposés. Dans cette tâche, les participants devaient proposer un algorithme qui serait en mesure de déterminer, en ayant accès au contexte conversationnel nécessaire,

la présence de sarcasme dans un énoncé. Les meilleurs résultats ont été obtenus par le participant «miroblog» (Lee *et al.*, 2020). L'architecture proposée par ce dernier comprend un classificateur composé de BERT suivi d'un BiLSTM et d'un NeXtVLAD . De plus, pour les données non étiquetées , ils ont utilisé un système basé sur le niveau de confiance associé à la prédiction d'une phrase issu de BERT. Ainsi, ils arrivent à des scores de précision de 0.932, de rappel de 0.936 et F1 de 0.931 sur des corpus issus de Twitter. De ce fait, ils dépassent le score F1 du deuxième meilleur modèle de 8,4%.

Malgré cette performance intéressante, comme défis supplémentaires découlant de ce type de tâche,Ghosh *et al.* (2020) soulignent des éléments qui font écho aux sections précédentes de ce travail :

«*However, we still notice that instances with subtle humor or positive sentiment are missed by the best-performing models even if they are pretrained on a very large-scale corpora*».

5 Objectifs futurs

L'objectif de ce travail était de présenter les différents défis propres à la réalisation d'algorithmes de détection automatique de l'ironie. Dans des travaux futurs, il serait intéressant de vérifier empiriquement la dynamique des limitations théoriques présentées plus haut et de déterminer quels types d'ironies verbales semblent les plus difficilement détectables par les modèles computationnels existants. Pour ce faire, il est nécessaire d'avoir une définition concrète de ce type de discours. S'il existe peu de consensus dans la littérature linguistique à ce sujet, Beals (1995) propose la définition suivante qui semble la plus englobante des différents types d'ironie existants :

> Ironie verbale : l'utilisation d'une expression verbale pour faire semblant que quelque chose est vrai tout en soulignant quelque chose d'extrêmement faux.

« L'utilisation d'une expression verbale » fait référence à la proposition théorique de Wilson & Sperber (1992). Pour (Beals, 1995) un énoncé ironique n'est pas nécessairement qu'une mention, mais peut aussi être un usage direct d'une expression. Par « pour faire semblant que quelque chose est vrai », Beals (1995) critique la théorie du faire semblant proposé par Clark & Gerrig (1984). Selon elle, affirmer qu'un locuteur, en ironisant, jouerait un rôle s'avère trop large et non représentatif de sa relation avec l'interlocuteur. Cette description se rapproche plus de celle d'une caricature que de la production d'un énoncé ironique. Elle maintient néanmoins que le locuteur ne croit pas le propos qu'il énonce. Par « souligner quelque chose d'extrêmement faux », Beals (1995) propose de regrouper les énoncés qui sont inappropriés, non pertinents et non véridiques sous cette catégorie. L'adverbe « extrêmement » souligne le rapport souvent humoristique de l'ironie verbale.

À partir de cette définition, elle est en mesure de proposer une trentaine de types d'ironie verbale. Parmi celles-ci, on retrouve des cas classiques d'opposition de sens faisant écho à la perspective gricéenne mentionnée plus haut. Toutefois, on retrouve aussi des cas d'ironie verbale plus subtils. Par exemple, en (8), l'ironie prend forme dans une fausse causalité négligeant volontairement la cause réelle. Cet énoncé est ironique si on sait que le locuteur qui émet cette phrase est quelqu'un d'infâme dans ses relations de couple et que son travail ne joue aucun rôle dans sa situation matrimoniale. On retrouve aussi des cas où l'ironie peut prendre naissance dans le choix des mots utilisés. En (9), c'est le fait de désigner une librairie comme étant des vendeurs de drogue qui est ironique. De plus, Beals (1995) mentionne des types d'ironie qui se manifestent sous forme de question où la réponse est évidente, comme en (10).

7. Je me retrouve célibataire parce que je travaille beaucoup trop.
8. Depuis que je suis sobre, je vais à la libraire du Square pour avoir mon fix.
9. Qui voudrait réellement avoir des conditions de vie décentes en ayant accès à un salaire minimum adéquat ?

Par ailleurs, comme le souligne Beals (1995), l'ironie ne s'exprime pas que verbalement. On retrouve des situations ironiques, comme en (11), qui découlent de la mise en relation de deux événements distincts.

10. Le patron de segway est mort en conduisant son segway en bas d'une falaise.

Avec une définition claire, faisant idéalement consensus, et une catégorisation exhaustive de l'ironie verbale, il est possible de déterminer quels sont les types les plus difficilement identifiables par les systèmes de détection automatique. Ainsi, il est possible d'avoir de meilleurs résultats.

De plus, il sera nécessaire d'évaluer le niveau d'erreur acceptable émis par les algorithmes de détection automatique de l'ironie. Il nous est actuellement impossible de savoir à quel point ces derniers sont meilleurs ou moins bons que les humains pour effectuer cette tâche. Les corpus d'entraînement étant annotés par ces derniers, on s'attend généralement à ce que les algorithmes puissent réussir à détecter efficacement les énoncés qui sont ironiques. Néanmoins, certaines phrases peuvent apparaître beaucoup plus ambiguës que d'autres. Par exemple, (Van Hee, 2017) rapporte le tweet en (7) où, sans la présence du hashtag, on se retrouverait face à une absence d'indice permettant à nous et aux algorithmes de bien détecter la présence d'ironie.

11. There is a thing called an all nightery and apparently, I wanna pull one #not.

Il existe un ensemble de stratégie discursive pour faire comprendre à notre interlocuteur qu'un énoncé produit sera ironique. Ces dernières dépendent des modalités utilisées pour transmettre ce type de discours. Dans une conversation, à l'oral, il sera possible de faire varier sa prosodie pour bien se faire comprendre (Bryant, 2010) ou, même, de faire certaines expressions faciales particulières pour désambiguïser son propos (Deliens *et al.*, 2018). Il est aussi important de noter que nous serions moins portés à utiliser l'ironie lorsque nous connaissons mal notre interlocuteur (Gibbs, 2000). Cohn-Gordon & Bergen (manuscrit) proposent même que l'ironie soit une façon de consolider des informations partagées entre deux personnes. De ce fait, un locuteur se doit de laisser le plus de clés possible à son interlocuteur pour que ce dernier soit capable de bien décoder son propos. L'espace d'informations partagées entre ces derniers est beaucoup plus dynamique à l'oral qu'à l'écrit. Au travers de cette dernière, un locuteur pourra employer diverses stratégies aux objectifs similaires comme l'utilisation particulière d'émojis, de majuscules et/ou de ponctuations. Ces traits sont utilisés dans la plupart des algorithmes de détection (Carvalho *et al.*, 2009; Karoui, 2017) et offrent généralement de bons résultats.

6 Conclusion

Les avancées techniques au niveau de la détection automatique de l'ironie dépendent des recherches linguistiques sur ce type de phénomène. Leur progrès nécessite la prise en considération d'une caractérisation claire du phénomène et d'une adéquation entre cette dernière et les corpus d'entraînement

utilisés. En déterminant les types d'ironie les plus difficilement identifiables par les systèmes de classifications et en caractérisant leurs manifestations, il sera possible dans le futur d'adapter ces derniers en conséquence et d'augmenter leurs performances.

Références

ALBA-JUEZ L. & ATTARDO S. (2014). The evaluative palette of verbal irony. *Evaluation in context*, **242**.

ALIAS-I (2014). Lingpipe natural language toolkit.

ATTARDO S. (2000). Irony as relevant inappropriateness. *Journal of pragmatics*, **32**(6), 793–826.

BEALS K. P. (1995). *A linguistic analysis of verbal irony*. Thèse de doctorat, University of Chicago, Department of Linguistics.

BOUAZIZI M. & OHTSUKI T. (2015). Sarcasm detection in twitter :" all your products are incredibly amazing ! ! !"-are they really ? In *2015 IEEE Global Communications Conference (GLOBECOM)*, p. 1–6 : IEEE.

BRYANT G. A. (2010). Prosodic contrasts in ironic speech. *Discourse Processes*, **47**(7), 545–566.

CARVALHO P., SARMENTO L., SILVA M. J. & DE OLIVEIRA E. (2009). Clues for detecting irony in user-generated contents : oh... ! ! it's" so easy" ;-. In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, p. 53–56.

CLARK H. H. & GERRIG R. J. (1984). On the pretense theory of irony.

COHN-GORDON R. & BERGEN L. (manuscrit). Verbal irony, pretense, and the common ground.

DAVIDOV D., TSUR O. & RAPPOPORT A. (2010). Semi-supervised recognition of sarcasm in twitter and amazon. In *Proceedings of the fourteenth conference on computational natural language learning*, p. 107–116.

DELIENS G., ANTONIOU K., CLIN E., OSTASHCHENKO E. & KISSINE M. (2018). Context, facial expression and prosody in irony processing. *Journal of memory and language*, **99**, 35–48.

EKE C. I., NORMAN A. A., SHUIB L. & NWEKE H. F. (2020). Sarcasm identification in textual data : systematic review, research challenges and open directions. *Artificial Intelligence Review*, **53**(6), 4215–4258.

FARÍAS D. I. H., PATTI V. & ROSSO P. (2016). Irony detection in twitter : The role of affective content. *ACM Transactions on Internet Technology (TOIT)*, **16**(3), 1–24.

GHOSH A. & VEALE T. (2016). Fracking sarcasm using neural network. In *Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis*, p. 161–169.

GHOSH D., VAJPAYEE A. & MURESAN S. (2020). A report on the 2020 sarcasm detection shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, p. 1–11.

GIBBS R. W. (2000). Irony in talk among friends. *Metaphor and symbol*, **15**(1-2), 5–27.

GODDARD C. (2018). "joking, kidding, teasing" : Slippery categories for cross-cultural comparison but key words for understanding anglo conversational humor. *Intercultural Pragmatics*, **15**(4), 487–514.

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

JORGENSEN J. (1996). The functions of sarcastic irony in speech. *Journal of pragmatics*, **26**(5), 613–634.

JOSHI A., SHARMA V. & BHATTACHARYYA P. (2015). Harnessing context incongruity for sarcasm detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2 : Short Papers)*, p. 757–762.

KAROUI J. (2017). *Détection automatique de l'ironie dans les contenus générés par les utilisateurs*. Thèse de doctorat, Université de Toulouse 3 Paul Sabatier ; Faculté des Sciences Economiques et ....

KERBRAT-ORECCHIONI C. (1978). *L'ironie*. Presses universitaires de Lyon.

KREUZ R. & CAUCCI G. (2007). Lexical influences on the perception of sarcasm. In *Proceedings of the Workshop on computational approaches to Figurative Language*, p. 1–4.

LEE H., YU Y. & KIM G. (2020). Augmenting data for sarcasm detection with unlabeled conversation context. In *Proceedings of the Second Workshop on Figurative Language Processing*, p. 12–17.

MARTINI A. T., FARRUKH M. & GE H. (2018). Recognition of ironic sentences in twitter using attention-based lstm. *International Journal of Advanced Computer Science and Applications*, **9**(8).

NILSEN E. S., GLENWRIGHT M. & HUYDER V. (2011). Children and adults understand that verbal irony interpretation depends on listener knowledge. *Journal of Cognition and Development*, **12**(3), 374–409.

NUNBERG G. (2001). *The Way We Talk Now : Commentaries on Language and Culture from NPR's" Fresh Air"*. Houghton Mifflin Harcourt.

PORIA S., CAMBRIA E., HAZARIKA D. & VIJ P. (2016). A deeper look into sarcastic tweets using deep convolutional neural networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics : Technical Papers*, p. 1601–1612.

PTÁČEK T., HABERNAL I. & HONG J. (2014). Sarcasm detection on czech and english twitter. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics : Technical papers*, p. 213–223.

RILOFF E., QADIR A., SURVE P., DE SILVA L., GILBERT N. & HUANG R. (2013). Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, p. 704–714.

STRAPPARAVA C., STOCK O. & MIHALCEA R. (2011). Computational humour. In *Emotion-oriented systems*, p. 609–634. Springer.

SULIS E., FARÍAS D. I. H., ROSSO P., PATTI V. & RUFFO G. (2016). Figurative messages and affect in twitter : Differences between# irony,# sarcasm and# not. *Knowledge-Based Systems*, **108**, 132–143.

TSUR O., DAVIDOV D. & RAPPOPORT A. (2010). Icwsm—a great catchy name : Semi-supervised recognition of sarcastic sentences in online product reviews. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 4.

VAN HEE C. (2017). *Can machines sense irony ? : exploring automatic irony detection on social media*. Thèse de doctorat, Ghent University.

WALKER M. A., TREE J. E. F., ANAND P., ABBOTT R. & KING J. (2012). A corpus for research on deliberation and debate. In *LREC*, volume 12, p. 812–817 : Istanbul, Turkey.

WALLACE B. C. (2015). Computational irony : A survey and new perspectives. *Artificial intelligence review*, **43**(4), 467–483.

WILSON D. & SPERBER D. (1992). On verbal irony. *Lingua*, **87**(1), 53–76.