Intégrer les émoticônes (et interjections) à des outils de traitement automatiques de corpus numériques : modélisation, enjeux, projets

Pierre Halté

Intégrer les émoticônes (et interjections) à des outils de traitement automatiques de corpus numériques : modélisation, enjeux, projets

Pierre Halté
Université François Rabelais - Tours
CREM / LLL

Introduction

La naissance de la communication par t'chat, dans les années 1970, a provoqué l'apparition de marques destinées à remplir, à l'écrit, les fonctions qu'occupent certains gestes et mimiques faciales à l'oral : les émoticônes[1]. Ce sont, à l'origine, de petits pictogrammes pixellisés, représentant des visages souriant ou tristes ; puis, dans les années 1980, apparaissent des émoticônes constituées de signes de ponctuation, comme « :-) », qu'on lit en penchant la tête vers la gauche – même si des versions « orientales », à lire de face comme : « ^_^ », sont aujourd'hui très présentes dans les corpus. Aujourd'hui et depuis la fin des années 2000, la banque de pictogrammes « emoji », apparue au Japon au sein de la téléphonie mobile, fait fureur : elle est constituée de pictogrammes dessinés, représentant des visages, des gestes, des objets… Pour notre part, nous réservons le terme d'émoticônes aux pictogrammes servant à indiquer les émotions du locuteur – nous reviendrons sur le cadrage théorique autour de la notion d'index ou d'indice. Les autres, ceux qui servent simplement à représenter des objets, seront nommés « pictogrammes purement iconiques». La banque emoji contient donc des émoticônes et des pictogrammes purement iconiques.

D'autres marques sont très présentes dans les corpus de t'chat, et remplissent les mêmes fonctions d'index d'émotion : les interjections, dont certaines sont apparues avec les nouvelles formes de communication médiatisée par ordinateur (les interjections acronymiques comme « lol », « mdr », « omg », etc.).

Ces marques sont très présentes dans les corpus, et les enjeux qui entourent leur étude sont nombreux : en effet, elles concernent une dimension importante du langage, qui relève de l'inscription du sujet parlant dans son propre discours.

Notre propos est d'abord de monter quels intérêts peut présenter l'étude de ces marques dans les corpus numériques, en termes de faisabilité tout d'abord, mais aussi en termes d'intérêt théorique et de la diversité des domaines qu'elles peuvent concerner. Une fois ces intérêts établis, nous proposerons un modèle d'analyse de ces marques, permettant de calculer, à partir de certains critères sémiotiques et pragmatiques, quels effets elles provoquent. Nous présenterons enfin le travail que nous avons accompli au sein du projet ANR APPEL[2], aux côtés de collègues informaticiens, pour intégrer notre modèle aux outils de traitement des corpus.

1 – Intérêts de ces marques pour les analyses de corpus numériques

1.1 – Des marques « numériques »

La première chose à souligner concernant ces marques est évidente : elles relèvent de l'écrit en général, pour les interjections, et de l'écrit en contexte numérique, spécifiquement, pour les émoticônes, et sont influencées, comme le montre Yus (2011), par ce contexte numérique. Ceci a des implications fortes en terme de recherche : puisqu'elles sont codées, intégrées dans les textes produits en contexte numérique, il est possible de les y rechercher facilement et automatiquement. Pour les interjections, la tâche est en effet assez aisée : il s'agit de repérer automatiquement des combinaisons de lettres données, comme on le fait pour n'importe quelle recherche textuelle au sein de divers logiciels (traitement de texte et autre). Pour les émoticônes, la tâche s'avère plus ardue : certaines

[1] Voir les travaux de Dear (2002) sur le réseau PLATO.

[2] « Analyse pluridisciplinaire du pétitionnement en ligne », coordonné par Contamin, J.G., Leblanc, J.M., Paye, O.

émoticônes apparaissent sous forme de pictogrammes dessinés, comme les emojis : elles sont alors codées en unicode, généralement, ce qui rend leur repérage automatique assez facile. Les émoticônes constituées de combinaisons de signes typographiques, elles, posent plus de problèmes (voir par exemple Pak et Paroubek 2011). Nous y reviendrons dans la dernière partie de cet article. Retenons pour l'instant qu'avec les émoticônes et les interjections, nous avons des marques qui apparaissent dans les corpus numériques et donc sur lesquelles nous pouvons faire des recherches automatiques facilement.

1.2 – Des marques (plus ou moins) figées

Le second intérêt de ces marques est qu'elles sont figées. Formellement, elles sont assez stables, même si les interjections peuvent être étirées (comme dans « looool » par exemple), et même s'il existe de nombreuses variétés d'émoticônes de sourire. Elles sont invariables, ne s'accordent pas avec les autres éléments de la phrase, ne se fléchissent pas, etc. Il suffit donc, pour pouvoir commencer à les repérer automatiquement, d'inventorier toutes les formes existantes (ou, tout du moins, les plus fréquentes) et des les classer dans différentes catégories. Voici un exemple de variations d'émoticônes trouvées dans notre corpus de commentaires de pétition en ligne, sur lequel nous travaillons au sein de l'ANR APPEL :

fx Smiley

	A	B	C	D	E	F
1	Smiley	Name	Meaning	Pre-category	Type	Forme
2	"=("	eqeyesfrown		negative	emoticon	=(
3	"=/"	eqeyesslant		negative	emoticon	=/
4	":("	frown		negative	emoticon	:(
5	":'("	frownapos		negative	emoticon	:'(
6	":-("	frownnose		negative	emoticon	:-(
7	":/[^/]"	slant		negative	emoticon	:/
8	":-/"	slantnose		negative	emoticon	:-/
9	":D"	bigsmile		positive	emoticon	:D
10	":-D"	bigsmilenose		positive	emoticon	:-D
11	":))"	doublesmile		positive	emoticon	:))
12	"=D"	eqeyesbigsmile		positive	emoticon	=D
13	"=]"	eqeyesbrac		positive	emoticon	=]
14	"=)"	eqeyessmile		positive	emoticon	=)
15	"(:"	rsmile		positive	emoticon	(:
16	":)"	smile		positive	emoticon	:)
17	":')"	smileapos		positive	emoticon	:')
18	":]"	smilebrac		positive	emoticon	:]
19	":-)"	smilenose		positive	emoticon	:-)
20	"=P"	eqeyestongue		relation	emoticon	=P
21	":P"	tongue		relation	emoticon	:P
22	":-P"	tonguenose		relation	emoticon	:-P
23	"D:"	dworry			emoticon	D:
24	":O"	omouth		surprise	emoticon	:O
25	";P"	winktongue		relation	emoticon	;P
26	"XD"	xeyesbigsmile		positive	emoticon	XD
27	";)"	wink		relation	emoticon	;)
28	";D"	winkbigsmile		positive	emoticon	;D

1 – Exemples d'émoticônes trouvées dans le corpus APPEL

Il existe quelques variations pour chaque type d'émoticône, mais rien de comparable avec les différentes flexions possibles d'un verbe, par exemple.

1.3 – Des indices fiables de subjectivité

Enfin, dernier et principal intérêt de ces marques, elles sont des indices fiables de l'émotion exprimée par le locuteur, et personne d'autre. Elles sont en effet nécessairement attachées, pour des raisons sémiotiques développées dans la partie suivante, au locuteur dans la situation d'énonciation. Elles indiquent son émotion *hic et nunc*, au moment où il la ressent. Il est impossible de trouver des émoticônes au discours rapporté (voir Halté 2017), et pour les interjections, s'il est possible d'en trouver rapportées au discours direct, le cas est rare, et de plus facilement identifiable. Considérer ces marques plutôt que d'autres est un avantage indéniable. Dans beaucoup de travaux consacrés aux études des émotions dans les corpus numériques, les recherches portent autour d'éléments lexicaux qui réfèrent à un objet de façon symbolique (au sens Peircien). Ainsi, organiser une recherche sur les émotions en utilisant le mot « tristesse » est rarement fiable : ce mot n'indique pas nécessairement la tristesse exprimée par le locuteur, il peut être employé pour référer à une réalité existante hors de la situation d'énonciation. Ce n'est pas le cas pour l'émoticône et l'interjections, qui, toutes deux, renvoient nécessairement à l'émotion du locuteur, ici et maintenant, au moment de l'énonciation. Ce sont donc des indices très fiables de subjectivité qui, à ce titre, devraient être inclus, aux côtés d'autres marques (y compris lexicales, bien sûr), dans les recherches de ce type.

2 – Modélisation sémiotique et pragmatique des interjections et émoticônes

2.1 – Des marques indexicales

Comparons ces trois énoncés :

(1) Mon chien est malade ☹
(2) Snif… Mon chien est malade.
(3) Je suis triste que mon chien soit malade.

Dans (3), « Je suis triste » est un énoncé prédicatif classique qui *dit*, ou *décrit*, l'émotion du locuteur. Le sens de « je suis triste » est propositionnel, ou, dans la sémiose peircienne, « symbolique » : « je suis triste » renvoie, selon des règles de représentation logiques, à un état de fait. On peut lier ce caractère propositionnel à des caractéristiques logiques comme la vériconditionnalité : on peut juger cet énoncé en termes de conditions de vérité, et, par exemple, le réfuter.
Dans (1) et (2), l'émoticône de tristesse et l'interjection onomatopéique « snif » *montrent* l'émotion du locuteur, au moment où il l'éprouve, tout comme le ferait à l'oral un geste ou une mimique faciale. Ce sont des signes indexicaux : dans la sémiose peircienne (voir Everaert-Desmedt 1990), l'*index* est un signe qui est relié à son objet de façon contigüe, immédiate, dans la situation d'énonciation, et qui le révèle ou qui pointe vers lui. Ces signes ont pour caractéristique, entre autres, d'échapper aux règles de la vériconditionnalité. On ne peut pas les juger en termes de vrais ou de faux ; ceci s'illustre par le fait que, par exemple, on ne puisse pas réfuter une émoticône, ni une interjection.
Les interjections et les émoticônes sont des index d'émotion, ou plus généralement de l'attitude du locuteur. Elles servent à indiquer, *hic et nunc*, son émotion. C'est la caractéristique sémiotique qui nous intéressera principalement dans cet article. Néanmoins, nous pouvons, par souci de définition, détailler plus précisément leur fonctionnement sémiotique, toujours en nous appuyant sur la sémiose peircienne.
Les émoticônes sont des *icônes* de mimiques faciales, de gestes, ou d'objets divers. L'icône est un signe qui est formellement similaire à l'objet auquel il renvoie (que ce soit dans sa matérialité sonore

ou graphique). Les émoticônes représentent donc, graphiquement, ces mimiques, gestes ou objets. Ces représentations graphiques sont conventionnelles et suivent des règles combinatoires et formelles : certaines sont formées de combinaisons de signes typographiques, elles ont une taille globale qui permet de les intégrer au texte, etc. Qu'elles relèvent du dessin (comme les emojis) ou de la combinaison de signes typographiques, ce sont des pictogrammes tels que les définit Vaillant (1999). Leur parcours sémiotique est donc complexe : ce sont des icônes utilisées comme index. Certaines émoticônes font même appel aux trois potentialités sémiotiques du triangle peircien : l'émoticône de cœur, par exemple, « <3 », ne peut *indiquer* l'amour que parce qu'elle est *l'icône* d'un objet qui *symbolise* l'amour. Note : nous n'appellerons pas « émoticône » un pictogramme qui n'indique pas une émotion mais qui représente simplement un objet.
Les interjections sont elles aussi des index de l'émotion du locuteur. Selon les mêmes critères sémiotiques, on peut les classer en deux groupes.
Les interjections *primaires* (ou onomatopéiques) sont des icônes sonores (onomatopées) actualisées en index d'émotion (« ouf ! », « Aïe ! », « Argh ! »). Elles sont à ce titre très proches des émoticônes. Elles représentent, dans leur forme sonore, un bruit, pour servir d'index d'émotion.
Les interjections *secondaires* (ou dérivées) sont des mots employés à l'origine de façon symbolique, pour référer à un élément extérieur à leur énonciation, qui au fil du temps se sont figés dans des emplois indexicaux par un processus de délocutivité décrit notamment par Anscombres (1985a et 1985b). Pour certaines, les deux emplois sont encore en concurrence dans la langue : par exemple le mot « chic », dans son emploi symbolique, sert à qualifier un nom (« Ce manteau est très chic »), alors qu'employé comme interjection il sert à manifester la joie du locuteur (« Chic ! Mon chien est guéri ! »). On trouve, en contexte numérique, nombre d'interjections secondaires, qui sont souvent acronymiques : « lol », « mdr », « omg », etc. Ainsi, un locuteur francophone n'a plus besoin, aujourd'hui, de savoir à quoi renvoie littéralement « lol » (« laughing out loud ») pour l'utiliser comme index de la joie.

Les interjections et les émoticônes ont donc un point commun : elles servent toutes deux à inscrire le sujet parlant dans son discours, y compris à l'écrit, en indiquant ses émotions. Pour nous, c'est l'occasion de déterminer le **premier critère** de notre modèle de fonctionnement de ces marques : nous pouvons classer les émoticônes et les interjections selon le type d'émotion qu'elles indiquent. Nous avons distingué quatre grandes catégories (voir Halté 2013) – qui demanderont à être confirmées et légitimées par les résultats que nous obtiendrons en les confrontant aux grands corpus dont nous parlerons plus loin.

- Les émoticônes et interjections **positives** indiquent des émotions positives et, en ce qui concerne l'aspect iconique des émoticônes, présentent des bouches aux coins relevé, des yeux rieurs, etc. : :), « lol », « mdr », etc.
- Les émoticônes et interjections **négatives** indiquent des émotions négatives et en ce qui concerne l'aspect iconique des émoticônes, présentent des bouches tordues, des yeux tombants, des larmes, etc. : :(, :S, :/, « merde ! », « hélas ! », etc.
- Les émoticônes et interjections de **surprise** indiquent la surprise et le choc, et en ce qui concerne l'aspect iconique des émoticônes, présentent des bouches et des yeux arrondis : :O, O_o, « omg ! », « wtf ! », etc.
- Les émoticônes et interjections **relationnelles** ou **d'adresse** codent spécifiquement l'adresse à l'interlocuteur : elles n'ont de sens qu'en relation avec l'autre. On y trouve par exemple l'émoticône de clin d'œil, établissant une connivence, ou l'émoticônes de tirage de langue servant à provoquer autrui, etc. : ;), :P, « Ouste ! », « Hey ! », etc.

Notons que ces catégories représentent la signification conventionnelle des émoticônes et interjections ; cette signification peut participer à la construction d'un sens qui peut lui être opposé (en fonction des éléments avec lesquels l'émoticône interagit), par exemple dans les cas de sarcasme ou d'ironie.
Parce qu'elles sont des index d'émotions, les émoticônes et les interjections sont aussi des modalisateurs.

2.2 – Des modalisateurs à visées multiples

Les modalisateurs[3] sont des marques indexicales dont la fonction spécifique est de modifier l'interprétation littérale d'un énoncé propositionnel. On peut dire, suivant les travaux de Bally (1944), qu'elles relèvent du *modus* : l'ensemble des marques non propositionnelles, *montrant* la position du sujet parlant par rapport à un *dictum* (ce qui est *dit* – la composante propositionnelle de l'énoncé) pour en orienter l'interprétation et fabriquer, en contact avec lui, le sens final de l'énoncé. Cette interaction est évidente lorsque l'on se livre à quelques petits tests de substitution. Comparons par exemple :
(1) Mon chien est malade ☹
(4) Mon chien est malade ☺

Ou encore :

(5) Youpi ! Mon chien est malade.
(6) Hélas ! Mon chien est malade.

Dans tous ces exemples, l'énoncé propositionnel, référant de façon logique et véricondi tionnelle à un état de fait, reste stable. Il s'agit avec « Mon chien est malade » de représenter un état de fait. Par contre, les marques modales, elles, varient. A chaque fois, l'interprétation finale de l'énoncé change complètement. C'est le cas le plus classique de ce qu'on appelle aujourd'hui la modalisation : les marques du modus *visent* le contenu propositionnel pour en modifier l'interprétation littérale.

Les choses se compliquent un peu lorsque l'on observe les corpus numériques de t'chat. L'interprétation des émoticônes et des interjections diffère selon qu'elles visent l'énoncé produit par le locuteur (nous parlons alors d'émoticône ou interjection *monologique*), ou selon qu'elles visent l'énoncé produit par l'interlocuteur (nous parlons alors d'émoticône ou interjection *dialogique*). Comparez par exemple :

(1) Mon chien est malade ☹.

(7) L1 : Mon chien est malade
L2 : ☹

Dans (7) par exemple, l'émoticône négative, produite par L2 en réaction à l'énoncé propositionnel de L1, est une marque de sympathie. Ce n'est évidemment pas le cas dans (1).
C'est le **second critère** de notre modèle : les interjections et émoticônes peuvent avoir une visée *monologique* ou *dialogique*, et, selon cette visée, elles ne marquent pas le même effet pragmatique[4].
Enfin, à la suite des travaux de Colletta (2004) et Perrin (2013) sur la modalisation, le concept de modalisation s'étend aujourd'hui à trois visées possibles pour les modalisateurs, et c'est le **troisième et dernier critère** de notre modèle :

- Visée **de contenu** : l'émoticône ou l'interjection vise le contenu propositionnel pour en modifier l'interprétation. C'est la conception traditionnelle de la modalisation, vue au début de cette sous-partie. Exemple : (1) Mon chien est malade ☹.
- Visée **pragmatique** : l'émoticône ou l'interjection vise non pas un contenu propositionnel mais la relation du locuteur à l'interlocuteur. Exemple tiré d'un corpus d'apprentissage en ligne via t'chat :

[3] On réserve habituellement ce terme à des marques linguistiques, mais nous l'étendons, sans scrupule, aux émoticônes.
[4] Pour une réflexion sur les aspects monologiques et dialogiques des marques modales, voir Perrin (2013).

(8) [18:58:53] E2 : Bonsoir **:)**
[18:59:52] FORMATEUR : Bonsoir E1
[19:00:05] FORMATEUR: Bonsoir E2
Ici, l'émoticône de sourire ne vise pas un contenu. « Bonsoir » est une formule de salutation qui n'a pas de contenu propositionnel. Il ne s'agit pas, ici, avec l'émoticône de sourire, de commenter « bonsoir », de montrer l'amusement du locuteur concernant le fait de dire bonsoir ou le fait que le soir soit bon, etc. Non : il s'agit d'accompagner une formule de salutation d'un sourire. C'est, ici, la relation à l'autre qui est modalisée de façon positive, comme c'est d'ailleurs l'usage dans les salutations en face à face.

- Visée **énonciative** : l'émoticône ou l'interjection vise la relation du locuteur à son énoncé. L'émoticône peut ainsi constituer une réaction à la forme de l'énoncé, mais aussi servir à soutenir l'acte de langage de l'énoncé qu'elle accompagne. Les émoticônes ou interjections à visée énonciative apparaissent notamment après une faute de frappe ou un lapsus.

2.3 – Modèle de fonctionnement pragmatique

À partir de l'étude qualitative de plusieurs corpus (notre corpus de thèse, constitué de cinquante pages de t'chat ; un corpus d'apprentissage en ligne via t'chat, voir Halté 2017 ; et un corpus de commentaires de pétitions en ligne, étudié au sein du projet ANR APPEL), nous pouvons croiser nos trois critères pour déterminer les fonctions pragmatiques des émoticônes et des interjections dans des configurations différentes. Ainsi par exemple, les émoticônes négatives dialogiques à visée de contenu manifestent la sympathie du locuteur.
Voici, synthétisé sous forme de tableau, les résultats provisoires que nous avons obtenus, en croisant ces trois critères, après une étude qualitative de plusieurs corpus (nous avons laissé de côté la visée énonciative, trop peu fréquente dans les corpus) :

Visée / **Type d'interjection ou d'émoticône**	**Monologique**		**Dialogique**	
	Contenu	**Pragmatique**	**Contenu**	**Pragmatique**
Positive	Plaisanterie Atténuation Ironie	Politesse Soutien d'un acte de langage (acquiescement ou requête)	Compréhension d'une plaisanterie Empathie	Politesse Acquiescement
Négative	Tristesse Gêne		Empathie	Désaveu, désapprobation
Surprise	Marquage de la surprise		Marquage de la surprise	Demande de reformulation
Relationnelle	Connivence	Provocation	Connivence réussie	Marquage de la présence du

	Ironie Taquinerie Support de requête			locuteur

2 – Fonctions pragmatiques des émoticônes (récapitulatif)

L'enjeu est évidemment de pouvoir intégrer ce modèle à des outils de traitement automatique de vastes corpus numériques, afin d'obtenir des résultats exploitables concernant les effets pragmatiques qui s'y trouvent. Le travail commence et est loin d'être aisé.

3 – Intégration du modèle aux logiciels : bilan, exemples et projets

3.1 – État des lieux

Nous avons commencé cette intégration au sein du projet ANR APPEL, dédié à l'analyse d'un corpus de pétitions en ligne. En collaboration avec Philippe Gambette, maître de conférence en informatique à l'université de Marne la Vallée et développeur d'*Expora*, logiciel d'extraction de corpus à partir d'une base de données, nous travaillons dans un premier temps au repérage automatique des émoticônes dans notre corpus. La tâche pose quelques problèmes. D'abord, en ce qui concerne les émoticônes constituées de signes typographiques, ces derniers interviennent très souvent dans les règles syntaxiques des requêtes dont on se sert pour effectuer des recherches sur les corpus. Par exemple, la parenthèse, les deux points, l'accent circonflexe, ou une combinaison des trois, peuvent être utilisés dans certains langages informatiques pour exclure une variable de notre recherche ou effectuer une recherche aux paramètres spécifiques. Cela rend périlleux l'utilisation de certaines requêtes. Il a donc été nécessaire de coder des outils d'extraction spécifiques de certaines combinaisons de signes typographiques. Ensuite, nombre de ces signes typographiques apparaissent les uns à côté des autres dans les corpus sans pour autant constituer des émoticônes. C'est le cas notamment au sein des adresses *URL*. Sur de très vastes corpus, comme celui du projet APPEL, cela augmente le bruit de façon considérable. Il nous faut donc exclure certaines combinaisons en fonction des premiers résultats que nous avons obtenus. Enfin, il faut aussi tenir compte des variations qui peuvent altérer les formes des émoticônes : nombre de bouches, parfois doublées ou triplées pour manifester l'intensité de l'émotion indiquée, etc.

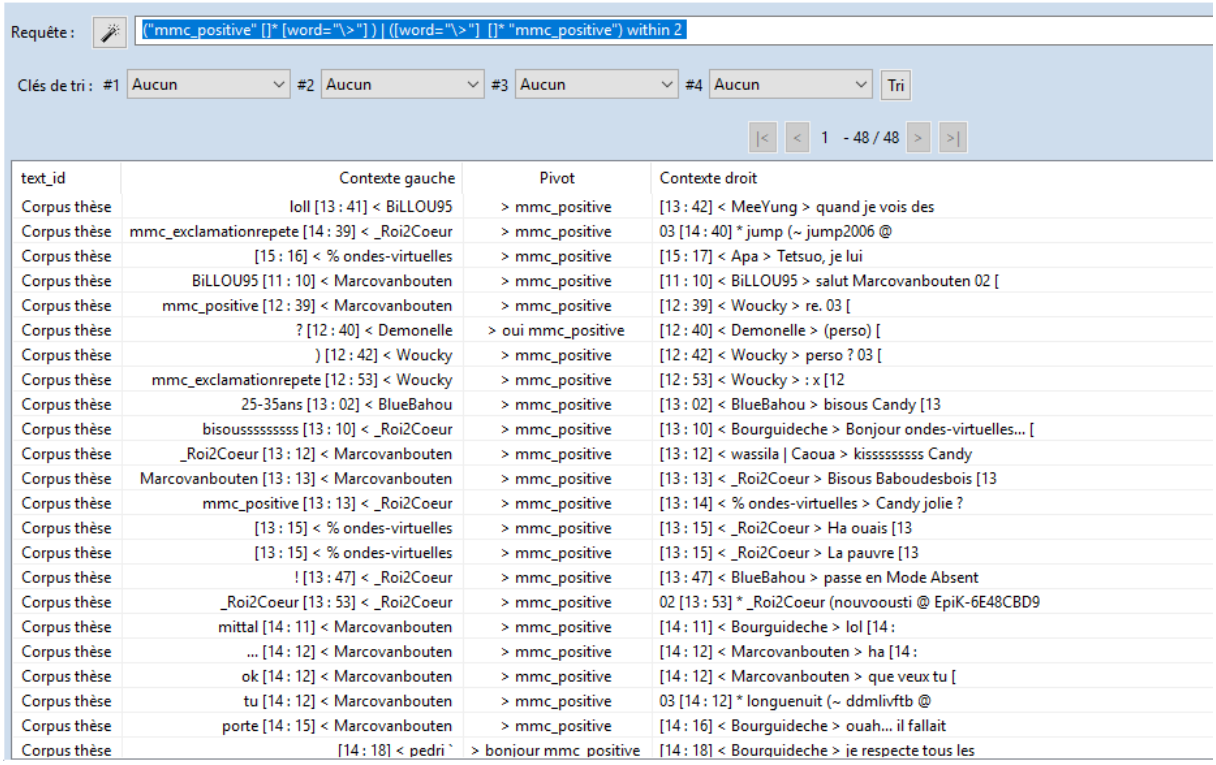
Une fois ce premier travail effectué, nous nous sommes attachés à classer les émoticônes que nous avons trouvées dans le corpus selon les quatre catégories proposées (notre premier critère) et à intégrer ces catégories dans *Expora*. Nous avons aussi nommé chacune des émoticônes observées. Nous pouvons donc actuellement automatiquement récupérer tous les énoncés contenant des émoticônes positives du corpus, par exemple, ou toutes les émoticônes de sourire seulement, ou encore toutes les émoticônes de clin d'œil, etc. Après extraction *via Expora*, nous obtenons un fichier .txt dans lequel les émoticônes sont remplacées par leur nom de catégorie ou leur nom sous forme textuelle. L'idée est de pouvoir utiliser ce fichier .txt avec des logiciels de TAL / Textométrie, comme *TXM* ou *TextObserver*, qui ne reconnaissent pas les émoticônes. Il est donc nécessaire de transformer ces dernières en texte. A partir de là, nous pouvons utiliser ces logiciels pour analyser notre corpus, et obtenir des résultats qui pour l'instant nous servent à affiner nos catégories et à tester quelques hypothèses. Nous pouvons par exemple constater que dans l'entourage proche des émoticônes positives, les éléments lexicaux qui apparaissent sont eux aussi connotés positivement ; nous pouvons faire des analyses factorielles en prenant en compte différentes émoticônes, etc.

Les autres éléments du modèle restent encore à intégrer, mais leur modélisation est beaucoup plus difficile. Il nous faut avant déterminer les facteurs qui nous permettraient de pointer automatiquement vers une visée monologique ou dialogique (c'est assez facile pour ce critère : les émoticônes

dialogiques n'ont pas de texte à leur gauche), et vers une visée de contenu / pragmatique / énonciative (ce qui est beaucoup plus difficile).

3.2 – Exemples d'applications

Avec les éléments du modèle que nous avons intégrés à nos recherches informatisées, nous pouvons, dans un premier temps, quantifier les émoticônes et les classer automatiquement selon leur type. Voici par exemple une recherche de cooccurrences portant sur les émoticônes positives de notre corpus de thèse (le corpus APPEL est confidentiel), sur *TXM* :


Requête : ("mmc_positive" []* [word="\>"]) | ([word="\>"] []* "mmc_positive") within 2

Clés de tri : #1 Aucun #2 Aucun #3 Aucun #4 Aucun Tri

|< < 1 - 48 / 48 > >|

text_id	Contexte gauche	Pivot	Contexte droit
Corpus thèse	lolll [13 : 41] < BiLLOU95	> mmc_positive	[13 : 42] < MeeYung > quand je vois des
Corpus thèse	mmc_exclamationrepete [14 : 39] < _Roi2Coeur	> mmc_positive	03 [14 : 40] * jump (~ jump2006 @
Corpus thèse	[15 : 16] < % ondes-virtuelles	> mmc_positive	[15 : 17] < Apa > Tetsuo, je lui
Corpus thèse	BiLLOU95 [11 : 10] < Marcovanbouten	> mmc_positive	[11 : 10] < BiLLOU95 > salut Marcovanbouten 02 [
Corpus thèse	mmc_positive [12 : 39] < Marcovanbouten	> mmc_positive	[12 : 39] < Woucky > re. 03 [
Corpus thèse	? [12 : 40] < Demonelle	> oui mmc_positive	[12 : 40] < Demonelle > (perso) [
Corpus thèse	) [12 : 42] < Woucky	> mmc_positive	[12 : 42] < Woucky > perso ? 03 [
Corpus thèse	mmc_exclamationrepete [12 : 53] < Woucky	> mmc_positive	[12 : 53] < Woucky > : x [12
Corpus thèse	25-35ans [13 : 02] < BlueBahou	> mmc_positive	[13 : 02] < BlueBahou > bisous Candy [13
Corpus thèse	bisousssssssss [13 : 10] < _Roi2Coeur	> mmc_positive	[13 : 10] < Bourguideche > Bonjour ondes-virtuelles... [
Corpus thèse	_Roi2Coeur [13 : 12] < Marcovanbouten	> mmc_positive	[13 : 12] < wassila | Caoua > kisssssssss Candy
Corpus thèse	Marcovanbouten [13 : 13] < Marcovanbouten	> mmc_positive	[13 : 13] < _Roi2Coeur > Bisous Baboudesbois [13
Corpus thèse	mmc_positive [13 : 13] < _Roi2Coeur	> mmc_positive	[13 : 14] < % ondes-virtuelles > Candy jolie ?
Corpus thèse	[13 : 15] < % ondes-virtuelles	> mmc_positive	[13 : 15] < _Roi2Coeur > Ha ouais [13
Corpus thèse	[13 : 15] < % ondes-virtuelles	> mmc_positive	[13 : 15] < _Roi2Coeur > La pauvre [13
Corpus thèse	! [13 : 47] < _Roi2Coeur	> mmc_positive	[13 : 47] < BlueBahou > passe en Mode Absent
Corpus thèse	_Roi2Coeur [13 : 53] < _Roi2Coeur	> mmc_positive	02 [13 : 53] * _Roi2Coeur (nouvoousti @ EpiK-6E48CBD9
Corpus thèse	mittal [14 : 11] < Marcovanbouten	> mmc_positive	[14 : 11] < Bourguideche > lol [14 :
Corpus thèse	... [14 : 12] < Marcovanbouten	> mmc_positive	[14 : 12] < Marcovanbouten > ha [14 :
Corpus thèse	ok [14 : 12] < Marcovanbouten	> mmc_positive	[14 : 12] < Marcovanbouten > que veux tu [
Corpus thèse	tu [14 : 12] < Marcovanbouten	> mmc_positive	03 [14 : 12] * longuenuit (~ ddmlivftb @
Corpus thèse	porte [14 : 15] < Marcovanbouten	> mmc_positive	[14 : 16] < Bourguideche > ouah... il fallait
Corpus thèse	[14 : 18] < pedri `	> bonjour mmc_positive	[14 : 18] < Bourguideche > je respecte tous les

3 – Capture d'écran de TXM : recherche du cotexte immédiat des émoticônes positives

Nous pouvons donc, *via* une recherche de cooccurrences, déterminer assez facilement, sur des petits corpus, si leur visée est monologique ou dialogique. Ainsi, par exemple, dans les trois petits corpus d'interactions didactiques via t'chat que nous avons étudié pour la revue *ELA* (Halté 2017), nous avons encadré tous les pseudonymes de chevrons (comme dans la capture d'écran ci-dessus). Par conséquent, toute émoticône suivant directement un chevron fermé est nécessairement dialogique : elle suit en fait le pseudonyme du locuteur qui la produit, ce qui indique qu'elle est produite en tout début de tour de parole, et donc qu'elle constitue une réaction à l'énoncé précédent.

[19:07:19] <FORMATEUR> pas grave. on fera mieux la prochaine fois ;)
[19:07:36] <Amandine> :)

Nous pouvons donc assez facilement dresser un tableau récapitulatif des émoticônes employées dans un corpus, ainsi que des usages spécifiques de leurs producteurs :

Forme d'émoticône	Corpus 1	Corpus 2	Corpus 3	Total	Formateurs	Étudiants
:)	13	18	9	40	9	31
^^	1	0	6	7	0	7
:D	2	0	0	2	0	2
:(	0	1	0	1	1	1
:/	1	0	6	7	0	7
;)	15	7	1	23	11	12
:p	1	0	0	1	0	1
Total	33	26	22	81	21	61
Type d'émoticône						
Émoticônes positives	16	18	15	49	9	40
Émoticônes négatives	1	1	6	8	1	7
Émoticônes de relation	16	7	1	24	11	13

Émoticônes à visée dialogique / monologique	Corpus 1	Corpus 2	Corpus 3	Total	Formateurs	Étudiants
Visée monologique	23	21	19	63	18	45
Visée dialogique	10	5	3	18	2	16

4 – Exemple d'étude quantitative des émoticônes dans un corpus d'apprentissage en ligne

Ces éléments permettent de faire des interprétations et d'aller les vérifier dans le corpus, grâce à la possibilité de « retour au texte » que fournit *TXM*. On peut donc étudier des émoticônes en particulier, des catégories complètes d'émoticônes, ou encore des rapports entre les émoticônes et leur cotexte plus ou moins proche, les usages selon les utilisateurs, etc. On peut aussi étudier la proximité et la fréquence d'apparition de certaines émoticônes en fonction d'un élément lexical, ce qui permettrait d'articuler une recherche concernant l'expression des émotions des locuteurs autour d'un thème donné. Ces pistes demandent à être explorées dans des recherches futures, portant sur des corpus beaucoup plus vastes – c'est le travail à venir au sein du projet APPEL.

3.3 – Projets futurs

C'est la première direction qu'empruntera le travail envisagé par la suite : exploiter les outils conçus par Philippe Gambette pour analyser les émoticônes présentes dans notre corpus de commentaires de pétitions en ligne, au sein du projet ANR APPEL. Des recherches concernant les catégories d'émoticônes, notamment, devraient nous permettre de vérifier que nos catégories d'émoticônes et d'interjections fonctionnent, et, le cas échéant, de les affiner ou de les préciser. Nous pouvons aussi effectuer des recherches par type de pétition, ou en croisant nos résultats avec certaines métadonnées que nous devons obtenir (âge des signataires, genre, etc.). Les applications et les interprétations peuvent toucher de nombreux champs : politique, marketing, sociologique, linguistique…
La seconde direction à explorer concerne l'intégration du second et du troisième critères (visée monologique/dialogique et visée de contenu/pragmatique/énonciative) de notre modèle aux outils de recherche. Les difficultés sont beaucoup plus nombreuses, surtout pour le troisième critère. En effet, en ce qui concerne le second critère, la visée monologique ou dialogique d'une émoticône peut assez facilement être identifiée automatiquement : il suffit de trouver un moyen, dans le corpus, de rechercher les émoticônes qui apparaissent au tout début du tour de parole du locuteur (ce que nous avons fait avec les chevrons dans les exemples précédents). Ces émoticônes sont nécessairement dialogiques, pour des raisons de portée explorées dans un autre article (Halté, à paraître en 2017) : l'émoticône porte toujours sur son cotexte gauche. Une émoticône apparaissant en début de tour de parole porte donc sur l'énoncé produit juste avant par l'interlocuteur.
Pour le troisième critère, par contre, les choses se compliquent sérieusement. Il nous faudrait être capables de déterminer, automatiquement, qu'une émoticône a une visée de contenu, une visée pragmatique, ou énonciative. Les moyens de déterminer cela reposent pour l'instant sur l'interprétation humaine. Quelques pistes, néanmoins, peuvent être envisagées. Par exemple, toute émoticône accompagnant une formule de politesse est à visée pragmatique. Les recherches de fréquences de cooccurrences permettraient peut être aussi de déterminer si une émoticônes est à visée de contenu : des éléments lexicaux que la doxa interprèterait comme positifs, en contact proche avec

des émoticônes positives, par exemple, tendraient à faire penser que ces émoticônes visent bien le contenu sémantique de ce qui est énoncé et pas autre chose. Il faudra évidemment tester cette hypothèse pour juger de sa fiabilité. La visée énonciative, elle, est beaucoup plus difficile à déceler automatiquement, mais c'est heureusement la moins fréquente. Reste aussi le problème de visées multiples, qu'on trouve souvent dans les corpus. Les émoticônes de clin d'œil, notamment, posent ce problème : elles visent très souvent à la fois le contenu propositionnel de l'énoncé qui précède, mais aussi la relation à l'interlocuteur ! En effet, la connivence, qui relève de la visée pragmatique, ne peut s'établir qu'autour d'un contenu propositionnel. Par exemple, dans le corpus étudié pour la revue ELA, on trouve ceci :

[19:08:57] FORMATEUR: |Vous verrez que j'attends beaucoup de nos échanges
[19:09:11] FORMATEUR: et qu'il est difficile de préparer à manger en même temps temps **;)**

Ici, l'émoticône de clin d'œil vise à la fois le contenu propositionnel (la représentation d'un état de fait : il est difficile de manger et de faire cours en même temps) ET la relation à autrui (visée pragmatique – interprétation soutenue en partie par le contexte : le formateur s'adresse ici à des jeunes mères de famille). On pourrait gloser le tout ainsi : « nous savons tous les deux qu'il est difficile etc. ». Il en va de même avec certaines émoticônes de sourire qui vise à la fois un contenu et la relation à l'interlocuteur. Une partie des travaux à venir doit donc être consacrée à l'affinement du modèle et aux moyens à mettre en œuvre pour être capable de déterminer automatiquement la visée des émoticônes.
L'enjeu est à la hauteur des difficultés : si nous y parvenons, il deviendra possible d'interpréter, automatiquement, les émoticônes selon les différents effets pragmatiques que nous avons répertoriés dans notre tableau récapitulatif : sarcasme, empathie, provocation, etc. Cela nous donnera un outil supplémentaire pour modéliser et interpréter correctement ces effets, qui posent de gros problèmes actuellement en ce qui concerne la recherche automatique des expressions des émotions dans les corpus – puisqu'il est clair que le seul angle « lexical », fondé sur un sens symbolique des expressions, est insuffisant. Il est indispensable et urgent de s'intéresser au sens indexical des expressions.

Conclusion

Les émoticônes et les interjections sont des marques dont l'étude peut ouvrir de nombreuses portes à qui s'intéresse à l'expression des émotions dans les discours numériques. Elles peuvent être recherchées et étiquetées assez facilement et, croisées avec d'autres données, permettent d'obtenir des résultats interprétables assez vite. Il est possible d'en modéliser le fonctionnement sémiotique et pragmatique afin d'en tirer des conclusions sur les effets de sens qu'elles produisent, qui parfois ne sont pas explicitement liés à la catégorie à laquelle elles appartiennent (une émoticône positive peut bien sûr être sarcastique). Ce modèle, reposant sur trois critères (catégorie, visée monologique/dialogique, visée de contenu/pragmatique/énonciative), est en cours d'intégration au sein d'outils informatique. Dans l'état, il est possible d'obtenir des résultats quantitatifs sur les émoticônes employées et leurs catégories ; nous espérons, à terme, intégrer l'ensemble du modèle aux outils de traitement automatique des corpus numériques. Les applications dépassent les enjeux théoriques des Sciences du Langage et sont très diversifiées : analyses politiques, marketing, sociologiques, etc.

Bibliographie :

Anscombre, J.-C. :

- (1985a) : « De l'énonciation au lexique: mention, citativité, délocutivité », *Langages* 80, Paris, Armand Colin, pp. 9-34.
- (1985b,) : « Onomatopées, délocutivité et autres blablas », *Revue Romane* 20/2, pp. 169-207, Copenhague, Université de Copenhague.

Bally, Ch. (1944) : *Linguistique générale et linguistique française*, 2e édition, Berne : A. Francke.

Colletta, J.-M. (2004) : *Le développement de la parole chez l'enfant âgé de 6 à 11 ans: corps, langage et cognition*, Bruxelles, éditions Mardaga.

Dear, B.L. (2002) : *PLATO emoticons,* accessible uniquement en ligne : http://www.platopeople.com/emoticons.html.

Everaert-Desmedt, N. (1990) : *Le processus interprétatif : introduction à la sémiotique de Ch. S. Peirce*, Liège : Mardaga.

Halté, P.

- (2017) : « Émoticônes et modalisation dans un corpus d'enseignement à distance (via t'chat) », *Études de linguistique appliquée*, n° 184, Paris : Klincksiek, pp. 441-452.
- (à paraître en 2017) : « Émoticônes et modalisation : ancrage énonciatif du locuteur dans un corpus de t'chat », in *L'expression des sentiments : de l'analyse linguistique aux applications*, Actes du colloque « Les sentiments dans les corpus », Poitiers : PUR.
- (à paraître en 2017) : « Positionnement syntaxique des interjections et des émoticônes : du rapport entre portée syntaxique et visée énonciative. », *Cahiers de Praxématique*, Montpellier, Université de Montpellier : Praxiling.
- (2013) : *Les marques modales dans le chat : étude sémiotique et pragmatique des interjections et des émoticônes dans un corpus de conversations synchrones en ligne.* Thèse de doctorat, mise en ligne le 13 décembre 2013. URL : [http://www.theses.fr/2013LORR0308].

Pak, A, Paroubek, P. (2011) : *Twitter as a Corpus for Sentiment Analysis and Opinion Mining*, Database and Expert Systems Applications, DEXA, Toulouse : International Workshops,.

Perrin, L., (2013) : « Les formules monologiques et dialogiques de l'énonciation », in *Les théories énonciatives aujourd'hui : un demi-siècle après Benveniste*, Paris : Ophrys, p. 187-211.

Vaillant, P. (1999) : *Sémiotique des langages d'icônes*, Paris : Honoré Champion.

Yus, F. (2011) : *Cyberpragmatics, Internet-mediated communication in context*, Amsterdam / Philadelphie : John Benjamins Publishing Company.