Parathon: un outil pour l'annotation de la communication digitale

Sorcha Walsh
Sous la direction d'Aris Xanthos
Projet en Informatique ou Méthodes Mathématiques

Printemps 2021

1 Introduction

Pour le cours "Projet en informatique ou en méthodes mathématiques", j'ai choisi de développer un package Python pouvant permettre d'annoter des documents issus de la communication digitale et en extraire les indices paralinguistiques propres à ce type de communication, tout en les mettant en rapport avec la paralinguistique présente en communication "face to face".
Ce rapport est donc constitué en partie d'une explication de la théorie derrière ce projet, grandement aidé par le travail de Kucharska, 2021. L'autre partie principale est dédiée au package en lui-même. Elle comporte notamment la spécification du package comme décrite au début du projet, et une analyse des résultats obtenus.

2 Histoire et contextualisation

La communication humaine va bien au-delà des simples mots échangés. Un certain nombre d'indices non verbaux et verbaux servent à enrichir notre communication et ont une fonction importante dans les relations interpersonnelles. Il n'est donc pas surprenant que les indices paralinguistiques dans la communication assistée par ordinateur aient été étudiés en détail dès les années 80, dans l'article de Carey, 1980 intitulé "Paralanguage in Computer-Mediated Communication". Dans cet article, Carey explore les différentes façons dont le paralangage était exprimé dans les communications par courrier électronique à l'époque. Il décrit les catégories suivantes :

- Orthographe vocale: peut exprimer des caractéristiques dialectiques ("how you doin' ?") ou le tempo ("Weeeeell...")
- Substituts lexicaux: petits apartés comme dans : "J'aime l'idée, mais encore une fois, c'était la mienne (dit-elle en rougissant)". Carey inclut également ce qu'il appelle les "ségrégations vocales", qui, d'un point de vue sociolinguistique, peuvent être appelées des sons de conversation non lexicaux ou, d'un point de vue grammatical, des interjections. Il s'agit, par exemple, de "mmh", "uh huh", "uhhhh" et ainsi de suite.
- Matrices spatiales: Carey note que certains utilisateurs traitent la page "comme une toile" (68), remplissant la zone avec des lettres si nécessaire.
- Manipulation des marqueurs grammaticaux. La ponctuation, les majuscules et ainsi de suite ont été, dans l'étude de Carey, utilisées à des fins très diverses, notamment pour souligner l'accentuation, le tempo, le rythme et la modification du ton.
- Caractéristiques négatives. Carey considérait qu'un manque de soin dans la rédaction d'un message (laisser des fautes d'orthographe ou des coquilles, ne pas utiliser les majuscules comme prévu) pouvait contribuer à un ton informel.

Au fil du temps et des recherches, tant du point de vue des CMC que de la linguistique, les définitions des indices de communication non verbale sont devenues plus spécifiques. Le chapitre du livre "Nonverbal Communication", 2012, par exemple, classe les catégories de communication non verbale comme suit.

1. La kinésique concerne le mouvement. Elle comprend les gestes, la posture, les mouvements du corps, le contact visuel et les expressions faciales.
2. L'haptique concerne la communication par le toucher. Par exemple, les poignées de main ou les étreintes.
3. Vocalique (ou Prosodie). L'aspect vocalique de la communication, telle que l'utilise "Nonverbal Communication", 2012, est un autre terme pour le paralangage. Il s'agit des aspects vocaux qui vont au-delà des mots ou de la prononciation : le volume, l'intonation, la vitesse, l'utilisation de compléments verbaux (que Carey qualifie de "substituts lexicaux"), la qualité vocale et la hauteur du son.

Code	Type	Commentaire
TK	Tactile Kinesics	Interaction physique et haptique non verbale.
VKG	Visual Kinesics Gestures	Mouvement d'une partie ou l'ensemble du corps.
VKF	Visual Kinesics Facial	Expressions faciales
VS	Vocalisation	Sons, pas nécessairement des mots.
VQ	Voice Quality	Indique la façon dont le ou les mots doivent être prononcés.
A	Artifact	Les éléments de présentation et de formatage d'un message
NoFTF	No FTF equivalent	Élements sans équivalent face-to-face

Table 1: Tableau répertoriant les types de cue FTF

4. La proxémique concerne les relations spatiales entre les personnes, en particulier l'espace personnel.

5. La chronémique est un terme qui décrit la façon dont le temps affecte la communication.

Parmi ces catégories, toutes ne sont pas pertinentes pour le CMC. Par exemple, la proxémique n'a pas de véritable équivalent en CMC. De plus, si certaines, comme la chronémique, sont très pertinentes et faciles à étudier dans le cadre du CMC, elles ne sont pas pertinentes dans le cadre de ce projet.

2.1 Typologie

Une typologie développée en collaboration avec Wioletta Kucharska et Jacinto Fernandez sera utilisée. Cette typologie permet de classer chaque indice dans trois catégories. Il y a une catégorie "face-to-face", qui identifie *grosso modo* le type de communication paralinguistique auquel l'indice en question correspond. Ces catégories sont répertoriés dans le tableau 1. Dans cette liste, la catégorie "Artifact" n'est pas utilisé dans ce projet. Il y a également deux catégories CMC (de communication digitale) pour chaque élément: une principale (cmc_main) et une secondaire (cmc_sub). Ces sont énumérés au tableau 2.
Cependant, pas toutes ces catégories ont été utilisés dans le cadre de ce projet. Notamment, étant donné les similarités entre les catégories "onom" (ontomatopée) et "voc_seg" (vocal segregates), ces deux catégories sont les deux comptés sous "voc_seg". La catégorie "combin", une combinaison de signes de ponctuation désignant la censure, a également été supprimé car plus très utilisé - aucun exemple de cet indice n'a été trouvé dans l'échantillon du corpus "Whats-up Switzerland" tagué selon cette typologie par Fernandez, 2021. La catégorie "EMO" (emoji other) a aussi été supprimé par souci de temps. Le

problème de détecter ces emojis sans également en détecter d'autres qui, eux, auraient une autre catégorisation, me paraissait trop complexe pour l'envergure du projet et le temps à disposition. Cette fonctionnalité pourrait cependant être implémenté dans une version future du package.
Une fois ces catégories établies, il s'agissait de répertorier et classer chaque type d'indice selon ces trois catégories. Ceci s'est surtout avéré être laborieux pour les emojis, qui sont très nombreux. Des expressions régulières permettant de détecter ces indices ont ensuit été établies.

Code	Main	Sub
VSP	Vocal spelling	LET_REP, CAP, ONOM, VOC_SEG, ACRO
PUN	Punctuation	PERIOD, ELLI, COMMA, QUOT, PAREN, EXCLAM, QUEST, SLASH, DASH, CORRECT, WHISPER, CENSOR, COMBIN
EMJ	Emoji	EMF, EMB, EMO
EMT	Emoticon	No subcategories

Table 2: Catégories CMC principales

3 Spécification

3.1 But du package

Le but de ce package est de créer un package Python capable de détecter et identifier des indices paratextuelles dans des corpus de textes extraits de CMC (computer mediated communication).

3.2 Besoins

Les besoins identifiées sont:

- Détecter dans un corpus de texte des éléments pouvant correspondre à des indices paratextuelles.
  - Pour ce faire, segmenter le texte de façon judicieuse: il faudrait en plus de la division en mots (qui devrait s'avérer simple), isoler les éléments de ponctuation. En plus, il arrive que plusieurs mots font partie de la même indice. Ex: "boo hoo", *fist bump*, etc.

  - Il arrive également que l'agencement spatiale qui permet de donner des informations paralinguistiques soit d'un type qui pourrait rendre ce procédé plus complexe. Ex: "p l e a s e".

- Identifier de quel type d'indice paralinguistique il s'agit.
- Pouvoir, au besoin, assigner plusieurs types à un élément.
- Avoir comme output un fichier tagué en format tabulé. Si possible, l'utilisateur devrait pouvoir choisir entre un format XML ou CSV.

3.3 Contraintes fonctionnelles

- Le package prend comme input des file object textuels.
- Le package utilise comme mécanismes de fonctionnement d'autres packages Python: NumPy, SpaCy, RegEx...
- Les expressions régulières seront le mécanisme de fonctionnement principal.
- Il sera aussi utile pour le package de pouvoir vérifier si un mot est écrit de façon correcte ou pas.
- Le package fonctionne pour l'anglais. Cependant, certains indices pouvant être détectés de façon purement formelle sont détectés pour toutes les langues.
- Le package est conçu de façon qui permet l'ajout éventuel d'autres fonctionnalités, par exemple:
  - Le support pour d'autres langues
  - La sélection d'un ou plusieurs types d'indice à détecter

4 Analyse du cahier de spécification

4.1 But

Le but, étant d'avoir à la fin du projet, un package Python qui puisse détecter et classifier des indices paralinguistiques, a été atteint. Le résultat final est effectivement en mesure d'accomplir cette tâche.

4.2 Besoins

Par souci de temps et de moyens, la solution de ségmentation choisie est la plus simple. Il s'agit d'une tokenisation "standard", qui sépare les textes en mots et en emojis. Bien qu'une indice s'étend souvent sur plusieurs tokens, il s'est avéré impossible de prendre cela en compte pour ce projet. Il s'agit d'ailleurs d'une limite structurelle dans la conception du package: ceci ne pourra pas facilement être résolu par la suite à moins de repenser de façon radicale la structure actuelle du code. C'est la principale faiblesse que j'identifie actuellement dans ce projet: à cause de cela, l'identificateur passe à côté de de nombreuses indices, que ce soit l'écriture espacée qui peut indiquer l'intensité ou encore la répétition d'emojis, une technique très utilisée pour intensifier celles-ci.
L'identification du type d'indice paralinguistique et l'assignation de plusieurs types à un indice, par contre, fonctionne très bien. La façon de faire choisie pour implémenter cela augmente de façon considérable la complexité d'execution de l'analyse, et donc le temps pris, mais la rapidité d'execution n'est pas un but pour ce projet.
Le choix entre le format XML et CSV fonctionne également bien. Il y a simplement deux fonctions différentes qui peuvent être appelés en cas de besoin. Dans un avenir hypothétique pour ce package, il serait peut-être envisageable d'intégrer le JSON à la panoplie de formats offers.
Parmi les autres fonctionnalités supplémentaires déjà offertes, il y a les colonnes de position, qui permettent une compatibilité avec, entre autres logiciels pour le traitement de données textuelles, Orange Textable. Il est également possible de préciser, en plus du langage, un "mode", permettant de détecter les options de formattage de WhatsApp. D'autres fonctionnalités pourraient être ajoutés au "mode", par exemple le nettoyage des données Whatsapp (enlever les dates et les heures d'envoi des messages typiquement).

4.3 Contraintes fonctionnelles

Globalement, ces contraintes sont respectées. La conception du package rend possible non seulement une analyse neutre de l'input, mais aussi l'ajout facile et non-invasive d'autres langues. Les indices propres à une langue sont détectées grâce à un dictionnaire JSON, avec en clé une expression régulière et en valeur une liste comprenant les trois catégories pertinentes, ainsi qu'un quatrième élément optionnel: un *regex flag* qui permet de dire que, par exemple, un tel regex devrait être insensible à la casse. Cet élément, bien que pas réfléchi en

amont, s'est avéré être particulièrement important: l'utilisation de majuscules est très important dans la communication digitale, mais certaines des expressions régulières utilisées sont, de par leur nature, obligées d'être insensibles à la casse.
En termes d'autres fonctionnalités, la sélection de types d'indices à détecter pourrait effectivement être implementé, mais cela ne devrait pas non plus être une grande priorité.

5 Résultats

Plusieurs facteurs font qu'il est difficile d'obtenir une évaluation quantitative de ce package. Notamment, le projet de Jacinto Fernandez se portait sur le corpus "Whats-Up Switzerland", où les emojis ont été retranscrits de façon textuelle, par exemple "emojiQCryingFace". Étant donné que ce package était conçu dans une optique de vouloir être générale, et que devoir refaire le même travail de catégorisation des emojis me paraissait laborieux, j'ai donc décidé de ne pas intégrer cette fonctionnalité dans la version finale du projet. Il est également important de noter que j'ai assimilé dans mon projet les "vocal segregates" et les onomatopés. Il s'agit certainement d'une différence, mais elle n'est pas significative: le fichier tagué manuellement comporte seulement 15 exemples d'onomatopées contre 510 exemples de "vocal segregates". La question de la pertinence de la catégorie des onomatopoées se pose donc.
La version faite par Parathon détecte 1599 indices, contre 3636 dans la version taguée manuellement. Cependant, un grand nombre (environ le tiers, 1377 sur 3636) des indices utilisées dans cet échantillon correspondent à des emojis. La somme des emojis et les indices détectées revient à 2976. 81% des indices sont donc détectées par Parathon, alors qu'il s'agissait uniquement de langages pas encore présents dans le dictionnaire.
Étant donné la taille de l'échantillon, une comparaison manuelle plus précise est impossible. De plus, le format rend l'utilisation des outils tels que les fonctions precision et recall de NLTK impossible ou du moins très peu juste.
J'ai donc manuellement tagué une petite conversation en anglais que j'ai eue avec une connaissance, qui n'est cependant pas très longue (70 indices au total). J'ai ensuite comparé ma version et celle faite par le package, et j'ai pu remarquer que nos résultats étaient identiques.
À vue d'oeil, il paraît donc que ce package permet d'obtenir des résultats signi-

Type d'indice	Référence	Test
TK	0	0
VKG	198	0
VKF	1533	66
VS	877	990
VQ	880	443
A	176	0
NoF2F	55	14
Total	3619	1513

Table 3: Comparaison du nombre d'indice par catégorie "face to face" dans les deux corpus

ficatifs, qui semblent être juste. La précision s'améliore lors de l'utilisation des dictionnaires spécifiques. Cependant, le manque d'une vraie analyse statistique fait qu'il est difficile d'affirmer cela de façon définitive. Il faudrait, peut-être même avant d'ajouter d'autres fonctionnalités, effectuer une analyse plus approfondie de l'efficacité actuelle de ce package pour la détection d'indices paralinguistiques.

Dans le tableau 3, il faut noter que le nombre d'indices répertoriés ne correspond pas au nombre d'indices au total. Ceci est simplement à cause de la possibilité précedemment évoqué pour un token de représenter différentes indices en même temps. Ceci dit, il y a tout de même moins d'indices répertoriés pour les deux corpus dans ce tableau qu'auparavant. Ceci s'explique simplement par le fait que, en l'absence d'autres outils, ces indices ont dû être comptés avec un outil peu fin: une recherche dans le texte faite dans Notepad ++. Les valeurs sont donc surtout indicatives. Néanmoins, on remarque de façon très nette l'absence de détection d'emojis dans le corpus de test. Quelques indices visuelles sont là (certainement représentés par des emoticones) mais un nombre moindre comparé à ce qui est réellement présent.

Un autre élément intéressant, et qui pourrait indiquer de faux positifs, est la sur-représentation de vocalisations dans le corpus de test. En examinant de plus près les corpus, il ne semblerait pas qu'il y ait un seul élément qui donne lieu à cette sur-représentation. Une analyse plus fine serait réquise.

Une partie de la raison pour laquelle Parathon détecte très peu d'éléments VQ est parce que certains emojis, dans le corpus de référence, étaient considérés comme pouvant indiquer la qualité du voix. De même, la présence des artefacts ("A") dans le corpus de référence est surtout du aux emojis.

6 Conclusion

Bien que ce package permet déjà d'annoter de façon simple et efficace des corpus, il est difficile d'évaluer correctement son efficacité en l'absence de données annotées permettant réellement de tester sa performance. Néanmoins, les buts principaux du projet ont été largements atteints, et l'architecture mise en place permet d'envisager de nombreux changements sans pour autant modifier le fonctionnement de base. L'évolution dans la communication digitale depuis que Carey écrivait son premier article sont évidents jusqu'à dans la façon dont la catégorisation a été faite, et l'échec de l'évaluation du package sur un certain corpus à cause de son incapacité à catégoriser les emojis sous leur forme textuelle.

Parmi les évolutions possibles de ce package, un ajout évident est le support pour d'autres langues. Ceci est naturellement très simple, nécessitant que l'ajout d'un dictionnaire dans le dossier "dictionaries". Le support pour le multilinguisme est envisageable, mais ceci nécessiterait un léger modification dans le code de base, et également de réfléchir de façon très prudente à ces dictionnaires pour ne pas avoir de conflit. Il serait peut-être même plus pertinent de créer pour chaque combinaison de langues un dictionnaire, ce qui permettrait peut-être également de détecter certains éléments propres au "franglais" par exemple.

Pour conclure, bien que ce package n'ait pas pu être évalué de façon très fine et nécessite encore sans doute quelques perfectionnements, je suis d'avis personnelle qu'il fonctionne déjà bien, et constitue tout au moins une bonne base de travail pour l'avenir.

References

Carey, J. (1980). Paralanguage in computer mediated communication. *Proceedings of the 18th annual meeting on Association for Computational Linguistics*, 67–69. https://doi.org/10.3115/981436.981458

Fernandez, J. (2021). *Constitution d'un corpus de référence multilingue pour les indices paralinguistiques dans les chats WhatsApp.* (Projet en Informatique). Université de Lausanne. Lausanne.

Kucharska, W. (2021). *Analysis of nonverbal cues based on swiss WhatsApp corpus. the semiotics of emotion expression* (Mémoire de master). Université de Lausanne. Lausanne.

Nonverbal communication. (2012). In *A primer on communication studies* (p. 80).