La variation géographique et sociale dans le français d'internet: émojis et émoticônes en France et au Québec

Marie Flesch

Travaux neuchâtelois de linguistique, 2023, 78, 19-40

La variation géographique et sociale dans le français d'internet: émojis et émoticônes en France et au Québec

Marie FLESCH
LLF (CNRS-université Paris Cité) et ATILF (CNRS-université de Lorraine)
marie.flesch@gmail.com

This article explores geographical and gender variation in the use of emoticons and emojis in French, in two sets of corpora: two corpora of tweets geolocalized in France and in Québec, and two corpora of comments posted on a French and a Québec Reddit forum. Results show that emojis are more frequent in the Quebec Reddit corpus than in the France Reddit corpus, and that emoticons are more frequent in the French corpora. They also show that emojis are more frequent than emoticons on Twitter, but not on Reddit. The effect of gender is significant in all corpora, with women using more emoticons and emojis than men. Finally, an analysis of the French Twitter corpus, which is annotated with age, reveals no effect of age on the frequency of emojis.

1. Introduction

Parce qu'ils font partie des phénomènes les plus visibles de la communication médiée par ordinateur, les émoticônes et émojis ont été abondamment étudiés, sur diverses plateformes: les listes de diffusion (Rezabeck & Cochenour 1995), la messagerie instantanée (Baron 2004; Garrison et al. 2011), les emails (Skovholt et al. 2014), les chats rooms (Del-Teso-Craviotto 2008), les blogs (Cassell et al. 2006; Kavanagh 2016), Twitter (Coats 2017; Magué et al. 2020; Pavalanathan & Eisenstein 2016; Schnoebelen J. 2012), Reddit (Flesch 2020; Tsou 2016), ou encore Facebook (Oleszkiewicz et al. 2017; Settanni & Marengo 2015). Certains travaux se sont intéressés à la fonction des émoticônes et émojis (Magué et al. 2020; Spina 2019; Tagg 2012), et d'autres à leurs liens avec des variables sociales. De nombreuses études ont montré qu'émoticônes et émojis sont utilisés plus fréquemment par les femmes que par les hommes, (Baron 2004; Chen et al. 2017; Coats 2017; Flesch 2020; Fullwood et al. 2013; Holtgraves 2011; Spina et al. 2013; Witmer & Katzman 1997). La majorité de ces travaux portent sur la langue anglaise; les travaux réalisés sur le français (Alloing & Pierre 2021; Halté 2017; Magué et al. 2020, par exemple) n'ont pas exploré l'effet de variables sociales comme le genre. Cette étude quantitative se propose d'explorer la variation sociale et géographique dans l'utilisation des émojis et émoticônes en français. Elle s'appuie sur deux ensembles de corpus: deux corpus réalisés à partir des forums de Reddit r/France et r/Quebec, et deux corpus de tweets géolocalisés en France et au Québec. Dans un premier temps, l'article propose un tour d'horizon des recherches sur les émoticônes et les émojis et décrit les deux plateformes étudiées, Reddit et Twitter (rebaptisé X en juillet 2023). Il détaille ensuite la construction et l'annotation des quatre corpus, ainsi que les techniques utilisées

pour en extraire émoticônes et émojis. Après avoir présenté des résultats généraux sur la fréquence des émoticônes et émojis dans les corpus, les analyses quantitatives, qui s'appuient principalement sur des modèles de régression, examinent les liens entre la fréquence des émoticônes et émojis et plusieurs variables: le type de plateforme (Reddit ou Twitter), l'identité de genre, le pays et enfin l'âge, uniquement pour les émojis et pour le corpus de tweets français, plus richement annoté que les autres corpus.

1.1 Émoticônes et émojis

Utilisées pour la première en 1982 par l'informaticien américain Scott Fahlman (Original Board Thread in which:-) was proposed, s. d.), les émoticônes sont composées de plusieurs caractères, qui peuvent être des signes de ponctuation, des lettres, des chiffres ou des symboles. Les émojis, quant à eux, sont nés dans la téléphonie mobile japonaise dans les années 1990, et ont été intégrés à Unicode en 2010. La dernière version d'Unicode, sortie en 2022, compte 3664 émojis (Emoji Counts v15.0, s. d.). Aujourd'hui, émoticônes et émojis coexistent sur les diverses plateformes, mais il semble que l'arrivée des émojis ait entrainé une baisse de la fréquence des émoticônes sur Twitter (Pavalanathan & Eisenstein, 2016). Les émoticônes et les émojis ont été décrits comme des éléments paralinguistiques qui remplacent les ressources de la communication en face à face (Provine et al. 2007). Ils sont donc souvent considérés comme des marqueurs d'émotion et d'affect (Rezabeck & Cochenour 1995). Émoticônes et émojis ont cependant plusieurs autres fonctions. Ils ont tout d'abord une fonction sociale: ils indiquent la familiarité et l'empathie, et renforcent les liens sociaux (Baron 2004; Vandergriff 2013). Ce sont également des marqueurs pragmatiques, qui fournissent des informations permettant d'interpréter les messages (Thompson & Filik 2016): ils explicitent l'attitude des personnes sur ce qu'elles écrivent (Tagg 2012). Ils peuvent être des marqueurs d'atténuation (Dresner & Herring 2010; Kavanagh 2016) ou au contraire intensifier un énoncé (Derks et al. 2008). En plus de ces fonctions attachées aux émoticônes, les émojis peuvent avoir un rôle lexical, en remplaçant par exemple un verbe ("je t'❤ ") ou un nom ("y a de la 📡") (Na'aman et al. 2017). Ils peuvent aussi être tout simplement utilisés pour attirer l'attention par leurs formes et leurs couleurs (Yus 2011). Émoticônes et émojis ont par ailleurs une fonction structurelle: ils délimitent les tours de parole et ont été comparés à la ponctuation (Provine et al. 2007). Sur Twitter, par exemple, ils sont le plus fréquemment utilisés en fin de tweets (Magué et al. 2020; Spina 2019). Ainsi, pour Halté (2017), les émoticônes "sont interprétées comme visant des contenus propositionnels situés à leur gauche".

Les émoticônes et émojis sont aussi liés à des variables sociales, à commencer par le genre. Ils seraient un marqueur de féminité et un "acte genré" (Del-Teso-Craviotto 2008: 259). De nombreuses études montrent que les femmes les utilisent plus fréquemment (Baron 2004; Chen et al. 2017; Coats 2017; Flesch

2020; Fullwood et al. 2013; Holtgraves 2011; Spina et al. 2013; Witmer & Katzman 1997). Certaines ont trouvé que les femmes utilisent davantage d'émoticônes et d'émojis "positifs" que les hommes (Koch et al. 2022; Oleszkiewicz et al. 2017). Les études qui se sont intéressées aux liens entre âge et utilisation des émojis et émoticônes ont des résultats contrastés: sur Facebook, plusieurs ont trouvé que les internautes les plus jeunes utilisent davantage d'émoticônes que les plus âgé·e.s (Oleszkiewicz et al. 2017; Schwartz et al. 2013; Settanni & Marengo 2015). Dans un corpus de WhatsApp, Koch et al. (2022) ont trouvé que les personnes les plus âgées préfèrent les émojis "objets" et "personnes", tandis que les plus jeunes ont tendance à utiliser davantage d'émojis représentant des visages. D'autres études n'ont pas mis en évidence de différences significatives (Fullwood et al. 2015; Pérez-Sabater 2013). Il y également des différences culturelles dans l'utilisation des émojis, à la fois dans leur fonction sémantique (Sampietro et al. 2022) et dans leur fréquence (Guntuku et al. 2019).

1.2. Reddit

Créé aux États-Unis en 2005, Reddit est un site internet qui accueille des centaines de milliers de forums portant sur toutes sortes de thèmes. Ces forums, qui sont appelés "subreddits" et dont le nom est toujours précédé par le préfixe "r/", sont conçus comme des communautés. Pour participer aux discussions, les internautes doivent créer un compte sur Reddit et choisir un pseudonyme. Contrairement aux profils sur les réseaux sociaux, les profils des Redditeurs et Redditrices ne contiennent généralement pas d'informations sociodémographiques. Très populaire aux États-Unis, Reddit comptait en janvier 2021 plus de 50 millions d'utilisateurs et d'utilisatrices actives par jour, et 13 milliards de commentaires (Reddit - Press, 2021). Même si la plupart des subreddits sont en anglais, de nombreux subreddits existent dans d'autres langues, dont le français.

1.3 Twitter

Twitter est un réseau social de microblogage qui permet à ses membres d'envoyer des messages, ou tweets, de 280 caractères maximum. Il compte 433 millions d'utilisatrices et utilisateurs actifs par mois à l'échelle mondiale (Réseaux sociaux les plus utilisés 2023, 2023). Le réseau social a 10 millions de membres actifs par mois en France, et en a 7.9 millions au Canada (Countries with most Twitter users 2022, 2023).

2. Méthodologie

2.1 Les corpus

2.1.1 Reddit

Les corpus ont été créés à partir des subreddits r/France et r/Quebec. Avec plus de 1.1 million de membres, r/France est le plus grand forum francophone de Reddit. Décrit, au moment de la collecte de nos données, comme "Le subreddit pour ce qui concerne la France et les Français", r/France est fortement axé sur l'actualité et la politique, mais on y trouve également des discussions sur la science, l'écologie, la cuisine ou encore les relations. r/Quebec, quant à lui, compte 246 000 membres; ses thématiques sont similaires à celles de r/France. Nous avons utilisé pmaw, une fonction wrapper pour l'API Pushshift (Podolak 2021), pour extraire tous les commentaires publiés sur r/France et r/Quebec entre septembre 2021 et septembre 2022. Depuis mai 2023, date à laquelle Reddit a restreint l'accès à son interface de programmation, il n'est plus possible d'utiliser cette technique pour recueillir les données du site. Nous avons traité le corpus en supprimant les URL, les commentaires écrits dans d'autres langues que le français et les citations, qui sont des commentaires ou des textes issus du web cités au sein d'un commentaire. Nous avons également supprimé les internautes qui ont publié des commentaires sur les deux subreddits, afin que les corpus soient les plus représentatifs possibles du français de France et du Québec.

Comme les profils des Redditeurs et des Redditrices ne contiennent généralement pas d'informations sur leur identité de genre, nous avons annoté le corpus en recherchant des déclarations du type "je suis une femme/un homme", ainsi que les marqueurs du genre grammatical. Pour ces marqueurs, la méthode utilisée a été la suivante: tout d'abord, nous avons recherché la séquence de mots "je suis" et identifié des adjectifs ("je suis content"), adjectifs précédés d'adverbes ("je suis très heureuse") et parfois noms ("je suis le seul") qui la suivent, qui sont fréquents et pour lesquels le genre grammatical est marqué à la fois à l'oral et à l'écrit. Cette condition avait pour but de limiter les erreurs d'annotation liées à d'éventuelles erreurs d'orthographe. Ensuite, nous avons effectué des requêtes basées sur ces résultats dans les corpus, en recherchant à la fois les formes au féminin et au masculin. La liste des requêtes utilisées est fournie en Annexe 1. L'identité de genre des internautes a ensuite été annotée automatiquement en utilisant les résultats des requêtes. Cette méthode a pour défaut de ne pas prendre en compte les personnes non binaires. En revanche, après une inspection manuelle de 500 profils et de leur historique de commentaires sur Reddit (et non dans le corpus uniquement), elle semble générer un faible taux d'erreur: seules deux personnes ont été mégenrées, dont une personne non binaire. Notons cependant que, comme pour l'annotation réalisée pour les tweets, la méthode utilisée repose sur le fait

que nous faisons confiance aux déclarations des internautes, qui peuvent mentir sur leur identité.

Après annotation, il reste 5248 personnes dans le corpus français, soit 7.40% des 70 937 internautes compris dans le corpus initial, et 2329 personnes dans le corpus québécois, soit 7.57% des 30 771 internautes du corpus de départ. Les corpus comprennent ainsi les commentaires d'approximativement 0.48% du 1.1 million de membres de r/France, et 0.95% des 246 000 membres de r/Quebec. Dans chaque corpus, il y a une forte majorité d'hommes: 87.65% dans le corpus de r/France et 89.55% dans le corpus de r/Quebec. Les corpus ont été tokénisés avec le package R Quanteda (Benoit et al. 2018). Après traitement et annotation, il reste près de 70 millions de tokens (ou mots) dans le corpus français, et près de 25 millions dans le corpus québécois (Figure 1). Le nombre médian de tokens par personne est de 4262 (*IQR* = 12504.5) dans le corpus français et de 3877 par personne (*IQR* = 10616) dans le corpus québécois.

	r/France	**r/Quebec**
Commentaires	1 376 316	578 311
Internautes	5248	2329
Femmes	648	333
Hommes	4600	1996
Tokens (femmes)	5 329 973	2 465 103
Tokens (hommes)	63 937 628	22 357 388
Tokens (total)	69 267 601	24 822 491

Figure 1: Composition des deux corpus de Reddit

2.1.2 Twitter

Les tweets ont été recueillis avec le package R rtweet (Kearney 2019) entre avril 2022 et février 2023, en indiquant des coordonnées GPS correspondant à des zones situées en France métropolitaine et au Québec, et en ciblant les tweets en langue française[1]. Les comptes vérifiés, appartenant souvent à des personnalités ou à des organisations, n'ont pas été inclus. Nous avons supprimé les doublons et enlevé les mentions (@Beyonce), les hashtags (#france) et les URLs des tweets. Nous avons également éliminé un maximum de tweets mis en ligne par des bots en utilisant la métadonnée "source" fournie par l'API de Twitter, en conservant uniquement les tweets mis en ligne depuis les applications de Twitter pour Android, iPhone, iPad et pour le web. Les profils Twitter sont souvent riches en informations; comme beaucoup indiquent un prénom, il est courant de procéder à une annotation automatique basée sur une

[1] Depuis les changements de ses conditions d'utilisation en avril 2023, Twitter (aujourd'hui X) n'offre plus d'accès gratuit à ses données. Par conséquent, le package rtweet ne permet plus d'extraire de tweets.

liste officielle des prénoms d'un pays donné (Coats 2017; Mislove et al. 2011, par exemple). Cette méthode a toutefois pour inconvénient de générer de nombreuses erreurs; dans leur étude, Hu et Kearney (2021) ont par exemple noté que les résultats de l'annotation automatique et de l'annotation humaine sont concordants dans 71.43% des cas. Nous avons opté, à la place, pour une annotation semi manuelle.

Dans un premier temps, l'âge des internautes a été extrait à l'aide d'expressions régulières, fournies en Annexe 2. L'identité de genre des internautes a ensuite été annotée manuellement à partir de plusieurs indices, quand ils étaient présents: l'utilisation du genre grammatical, le prénom, et les pronoms par lesquels les personnes souhaitent être désignées ("he/they" pour une personne non binaire, "she/her" pour une femme, par exemple; voir Annexe 2). Ce processus en deux temps a uniquement été réalisé pour le corpus français, car l'annotation automatique de l'âge a produit peu de résultats dans le corpus québécois: elle a retenu seulement 214 profils, contre plus de 8000 dans le corpus français. Pour le corpus québécois, nous avons donc procédé autrement: nous avons créé un échantillon aléatoire de 1000 profils, puis annoté l'identité de genre des internautes manuellement, en utilisant les mêmes indices que pour l'annotation du corpus français. Après annotation, il reste 51 206 tweets dans le corpus français et 31 012 dans le corpus québécois (Figure 2). Ces tweets ont été écrits par 2861 personnes en France, soit 40.87% des profils examinés, et par 526 personnes au Québec, soit 52.6% des profils examinés. Comme dans les corpus de Reddit, il y a une majorité d'hommes dans les corpus (58.79% en France, 64.26% au Québec). Le corpus français compte également 91 personnes non binaires, ou 3.18% des internautes. L'âge moyen des utilisatrices et utilisateurs du corpus français est de 24.99 ans (*SD* = 9.30), et l'âge médian est de 23 ans (*IQR* = 6). La majorité des tweets ont été publiés depuis les applications de Twitter pour iPhone, Android et iPad (80.67% des tweets québécois et 81.96% des tweets français). 19.33% des tweets québécois et 18.04% des tweets français ont été mis en ligne depuis l'application web de Twitter. Le corpus de tweets français contient 887 811 tokens, avec une médiane de 67 tokens par personne (*IQR* = 237). Dans le corpus québécois, il y a 732 265 tokens; le nombre médian de tokens par personne est de 102 (*IQR* = 486.75).

	France	**Québec**
Tweets	51 206	31 012
Internautes	2861	526
Femmes	1088	188
Hommes	1682	338
Personnes non binaires	91	-
Tokens (femmes)	262 083	211 162
Tokens (hommes)	578 063	521 103
Tokens (non binaires)	47 665	-
Tokens (total)	887 811	732 265

Figure 2: Composition des corpus de tweets

2.2 Extraction des émoticônes et émojis

2.2.1 Émoticônes

Nous avons utilisé la liste d'émoticônes de Wikipédia ("List of Emoticons" 2020) pour effectuer les requêtes dans les corpus. Cette liste se compose de 129 émoticônes. Pour Reddit, nous n'avons pas intégré l'émoticône cœur (<3); en effet, la nécessité d'enlever les commentaires contenant des citations d'autres commentaires, introduites par des chevrons, a entrainé une suppression de tous les chevrons du corpus. Les réduplications, comme:)))))), ont été considérées comme leurs formes simples (:), dans ce cas).

2.2.2 Émojis

Pour extraire les émojis du corpus, nous avons utilisé le package R emoji (Hvitfeldt 2022), qui contient une liste de 4702 émojis et leurs métadonnées, dont la catégorie à laquelle ils appartiennent ("animaux et nature", "smileys et émotions", "personnes et corps", "symboles", etc.) Le package contient un nombre d'émojis supérieur aux 3664 que compte Unicode en 2023, parce que certains émojis du package sont des combinaisons d'émojis, qui permettent par exemple d'indiquer le genre d'un émoji (👩‍🔬 ♀).

2.3 Analyses statistiques

Nous fournissons ci-dessous des statistiques descriptives et des analyses inférentielles. Les fréquences relatives des émojis et des émoticônes sont données pour 100 000 tokens lorsqu'elles décrivent un corpus entier, et pour 1000 tokens quand elles s'appliquent aux individus. Pour les analyses inférentielles, nous avons opté pour des méthodes non paramétriques adaptées à la surdispersion des données de corpus: le test de Wilcoxon-Mann-Whitney, la régression binomiale négative et la régression "zero-inflated". Les analyses ont été réalisées avec R, version 4.1.1 (R Core Team 2021).

3. Résultats

3.1 Résultats généraux

3.1.1 Émojis

Les requêtes ont permis d'identifier 29 311 émojis dans le corpus français de Reddit. Un examen des émojis les plus fréquents sur r/France a révélé une utilisation extrêmement fréquente (10 307 émojis) des émojis "carré vert", "carré rouge", "carré bleu" et "disque jaune", utilisés sous forme de séquences. Celles-ci représentent les résultats du jeu Sutom (*SUTOM*, s. d.), qui sont souvent partagés par les membres de r/France sur les fils de discussion "forum libre" (VifEspoirPirez 2022). Nous avons éliminé ces émojis, car ils ne reflètent pas un usage "classique" des émojis. Il reste donc 18 864 émojis, dont 818 émojis différents, dans le corpus français. Nous n'avons pas rencontré ce phénomène dans le corpus québécois, qui contient 16 695 émojis, dont 670 émojis différents. Les émojis ont une fréquence relative de 27.23 par 100 000 tokens dans le corpus français, et de 67.26 dans le corpus québécois. Ils sont donc 2.47 fois plus fréquents dans le corpus québécois que dans le corpus français. Dans le corpus québécois, 2.05% des commentaires contiennent au moins un émoji, contre 0.95% dans le corpus français.

Pour Twitter, nous avons recensé 874 émojis différents dans le corpus français et 521 dans le corpus québécois. En tout, il y a 22 732 émojis dans le corpus de tweets français et 17 746 émojis dans le corpus québécois. La fréquence relative des émojis est assez similaire dans les deux corpus: elle est de 2560.45 par 100 000 tokens pour la France, et de 2423.44 pour le Québec. Dans les deux corpus, plus d'un quart des tweets contient au moins un émoji: 25.07% pour la France et 28.55% pour le Québec. La figure 3 présente les dix émojis les plus fréquents dans les corpus. L'émoji "pleurs de joie" 😂 est le plus fréquent dans les corpus de Reddit: il représente respectivement 9.18% et 14.84% de tous les émojis identifiés dans les corpus français et québécois. Les émojis 😅 , 🤣 , 😉, 🤔 et 😭 sont également présents dans les deux listes. Dans les corpus de Twitter, l'émoji "visage qui pleure de joie" est moins fréquent que 😭 en France, et que 🤣 au Québec. Outre ces trois émojis, seuls deux autres émojis sont présents dans les deux listes: ♡ et 😅. Dans les corpus de Reddit, les dix émojis les plus fréquents représentent 38.45% de tous les émojis pour r/France, et 43.34% de tous les émojis sur r/Quebec. La proportion est plus élevée dans les corpus de Twitter: elle est de 68,37% pour la France et de 49.56% pour le Québec.

Dans les corpus de Twitter comme dans les corpus de Reddit, la catégorie d'émojis la plus fréquente est "smileys et émotions" (🤣 ou 🤔), suivie par "personnes et corps" (👏 ou 🙆‍♀️). Ces deux catégories représentent 79.38% des émojis dans le corpus de r/France, 87.29% dans le corpus de r/Quebec,

83.88% dans le corpus de tweets français et 92.51% dans le corpus de tweets québécois.

r/France		r/Quebec		tweets France		tweets Québec	
Émoji	Fréq. rel.	Émoji	Fréq. rel.	Émoji	Fréq. rel.	Émoji	Fréq. rel.
😂	2.50	😂	9.98	😭	253.32	🤣	391.39
😅	1.47	🤣	5.62	😂	209.28	😂	360.80
🤣	1.36	😅	3.39	🤣	102.84	😉	91.77
😁	1.06	🤭	2.34	♡	85.60	♡	62.41
😉	0.94	😉	1.60	😅	65.78	😅	54.08
👍	0.76	🙄	1.38	🥰	56.09	😍	50.94
🤭	0.74	🤡	1.36	💀	49.78	🤪	50.66
♡	0.56	🤦‍♂️	1.20	🤭	49. 00	😭	48.07
😭	0.55	😭	1.15	😍	47.87	🫣	46.29
■	0.52	♡	1.12	🥲	42.69	🤭	44.79

Figure 3: Les dix émojis les plus fréquents dans les corpus, avec fréquence relative par 100 000 tokens

3.1.2 Émoticônes

Nous avons identifié 34 641 émoticônes dans le corpus français de Reddit (22 types différents), et 11 284 dans le corpus québécois (19 types différents). Leur fréquence relative par 100 000 tokens est de 50.01 pour r/France et de 45.46 pour r/Quebec. Dans le corpus français, 2.4% des commentaires contiennent une émoticône, contre 1.85% des commentaires du corpus québécois. Les dix émoticônes les plus fréquentes dans les deux corpus sont identiques (Figure 4). Dans chaque corpus, les trois émoticônes les plus fréquentes, c'est-à-dire:);) et:(, représentent environ 70% des émoticônes identifiées. Les émoticônes "sans nez" sont beaucoup plus fréquentes que les émoticônes "avec nez". Par exemple,:) est 12.17 fois plus fréquente que son équivalent avec nez:-) dans le corpus français, et 16.59 fois plus fréquente que:-) dans le corpus québécois.

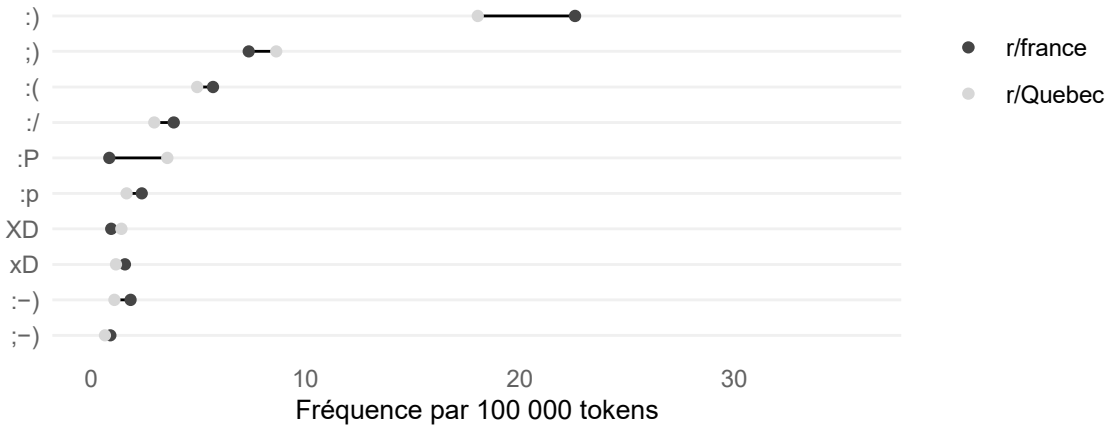

Figure 4: Fréquences relatives des dix émoticônes les plus fréquentes dans les deux corpus de Reddit

Sur Twitter, nous avons identifié 840 émoticônes dans le corpus français (15 types différents), et 341 dans le corpus québécois (10 types différents). Leur fréquence relative est de 94.61 par 100 000 tokens dans le corpus de tweets français et de 46.57 dans le corpus de tweets québécois. Comme sur Reddit,:) arrive en tête dans les deux corpus; cette émoticône représente 37.85% des émoticônes du corpus français et 36.36% des émoticônes du corpus québécois (Figure 5).

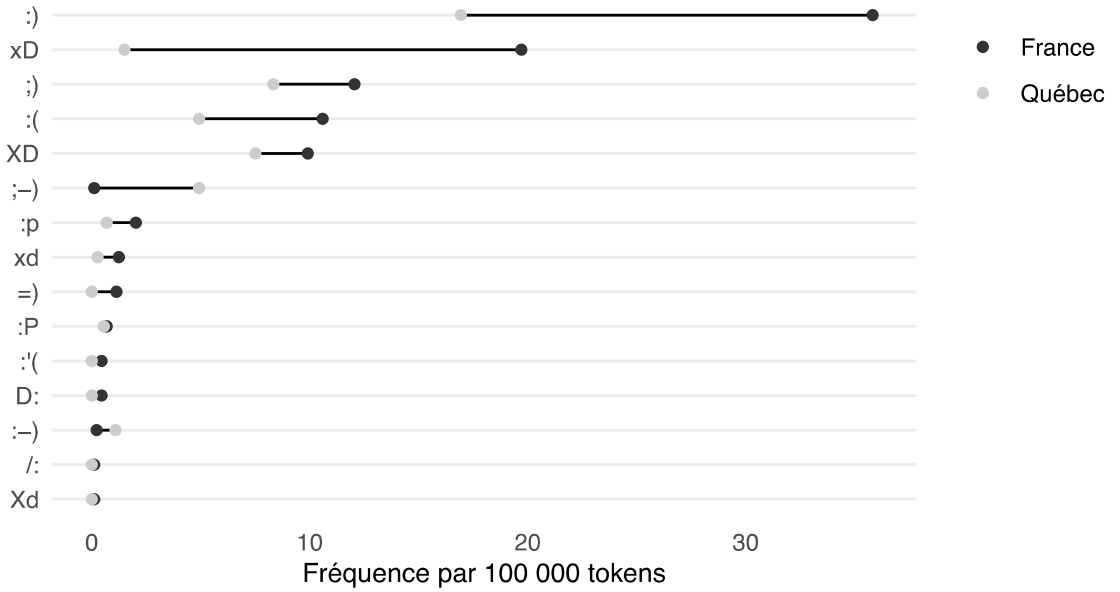

Figure 5: Fréquences relatives des émoticônes identifiées dans les corpus de tweets

3.1.3 Comparaison entre fréquence des émojis et émoticônes

Sur Twitter, les émoticônes sont considérablement moins fréquentes que les émojis: dans le corpus de tweets français, les émojis sont 27.06 fois plus fréquents que les émoticônes; dans le corpus québécois, ils sont 52.04 fois plus

fréquents que les émoticônes. Le test de Wilcoxon-Mann-Whitney montre que ces différences sont significatives (pour la France, $W = 6543214$, $p < .001$; pour le Québec, $W = 204562$, $p < .001$). Dans le corpus de r/France, les émoticônes sont plus fréquentes que les émojis, selon le test de Wilcoxon-Mann-Whitney ($W = 16710256$, $p < .001$). Il n'y a pas de différence significative entre émojis et émoticônes dans le corpus de r/Quebec ($W = 2741012$, $p = .50$).

3.2 Effet du genre sur la fréquence des émojis et émoticônes

3.2.1 Émojis

La proportion d'utilisateurs et utilisatrices uniques, c'est-à-dire de personnes qui ont utilisé au moins un émoji, est similaire pour les trois groupes de genre dans le corpus français de tweets: elle est de 68.75% pour les femmes, de 63.20% pour les hommes et de 70.33% pour les personnes non binaires. Au Québec, 61.70% des femmes et 53.55% des hommes ont utilisé au moins un émoji. Les fréquences relatives médianes par 1000 tokens des émojis sur Twitter sont présentées dans la Figure 6, avec les intervalles de confiance de 95% calculés par bootstrap avec le package R boot (Canty & Ripley 2022). Cette technique statistique a été choisie car elle permet de calculer des intervalles de confiance à partir de distributions fortement non normales (Haukoos & Lewis 2005), comme c'est le cas ici. Le diagramme en barres montre que, dans le corpus français comme dans le corpus québécois, la fréquence médiane des émojis est bien plus élevée chez les femmes que chez les hommes. Dans le corpus français, la fréquence médiane des émojis est similaire chez les hommes et chez les personnes non binaires. Les intervalles de confiance sont de grande taille. En effet, certains échantillons sont de taille réduite, et les données de fréquence sont généralement surdispersées (Hilbe 2014): ici, la majorité des internautes utilise un nombre réduit d'émojis, et une minorité en utilise énormément. La fonction boxplot() de R fait ainsi état de 231 valeurs aberrantes dans le corpus français, qui correspondent à des personnes ayant utilisé entre 125.18 et 700 émojis par 1000 tokens.

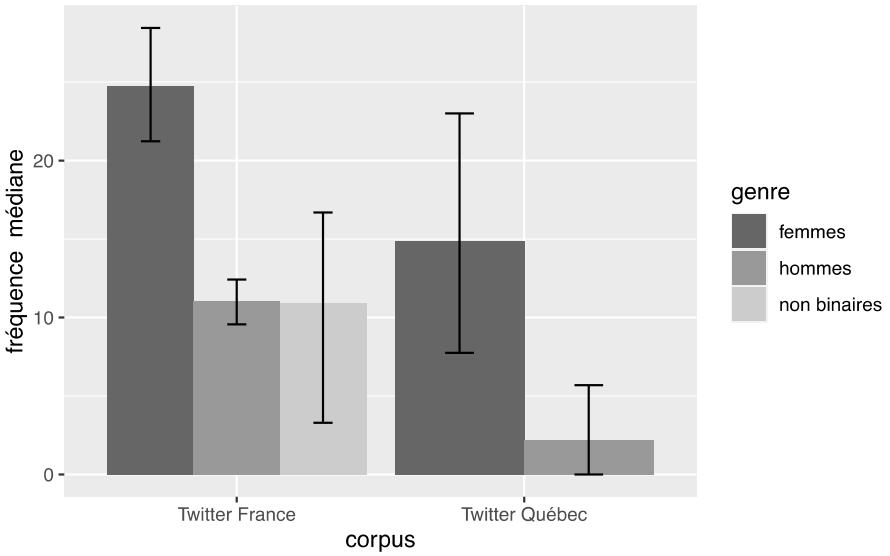

Figure 6: Fréquence médiane des émojis par 1000 tokens dans les corpus de tweets

Cette surdispersion est également présente dans les corpus de Reddit. On y note une proportion plus faible d'utilisateurs uniques d'émojis que sur Twitter: elle est de 42.44% pour les femmes et de 35.50% pour les hommes dans le corpus français, et de 56.16% pour les femmes et de 45.49% pour les hommes dans le corpus québécois. Dans les deux corpus, il y a donc davantage d'utilisatrices uniques que d'utilisateurs uniques. La fréquence médiane par 1000 tokens est de 0, puisque, dans chaque groupe, moins de la moitié des internautes ont utilisé un émoji. Les fréquences moyennes, qui sont ici peu fiables au vu de la dispersion importante des données, suggèrent que, comme sur Twitter, les femmes utilisent davantage d'émojis que les hommes. Dans le corpus de r/France, les femmes ont produit en moyenne 1.46 émoji par 1000 tokens (*SD* = 6.51) et les hommes 0.57 émoji (*SD* = 2.37). Dans le corpus québécois, la fréquence moyenne est de 1.83 pour les femmes (*SD* = 4.80) et de 1.08 pour les hommes (*SD* = 4.18).

Pour mesurer les effets du genre et de la variété de français sur la fréquence des émojis et émoticônes sur Twitter et Reddit, nous avons utilisé la régression binomiale négative, modèle capable de gérer la surdispersion propre aux données de corpus (Hilbe 2014). Les modèles ont été créés avec le package R MASS (Venables & Ripley 2002). Ils ont pour variable dépendante la fréquence brute des émojis. Nous avons ajouté un offset correspondant au nombre de tokens produits par chaque personne, ce qui permet de prendre en compte le fait que les sous-corpus ont des tailles différentes (Zuur et al. 2015). Les variables indépendantes intégrées dans la construction des modèles sont le genre, le pays, et l'interaction entre genre et pays. L'intégration de l'interaction part du principe qu'il est par exemple possible que les femmes de r/Quebec aient des pratiques différentes des femmes de r/France. Nous avons ensuite utilisé la fonction step() de R pour déterminer les modèles optimaux. Les

modèles sont présentés avec coefficients et intervalles de confiance (Figure 7); les coefficients ont été exponentialisés pour faciliter l'interprétation. Le niveau de référence du modèle 1 est tous les hommes; le niveau de référence du modèle 2 est les hommes de r/France.

Dans le modèle 1, le processus de sélection n'a retenu ni l'interaction entre genre et pays, ni l'effet du pays, ce qui signifie que selon le modèle, il n'y a pas de différence d'usage des émojis entre France et Québec. L'effet du genre est significatif: les femmes utilisent plus d'émojis que les hommes; le coefficient indique une taille d'effet de 1.495, qui signifie que quand les hommes utilisent 1 émoji, les femmes en produisent 1.495. Le modèle 2 a retenu l'interaction du genre et du pays. Il montre que, dans le corpus français, la fréquence des émojis est significativement plus importante chez les femmes que chez les hommes, avec une taille d'effet de 2.38. Les hommes de r/Quebec produisent deux fois plus d'émojis que les hommes français. Il n'y a en revanche pas de différence significative entre les hommes de r/France et les femmes de r/Quebec. En changeant le niveau de référence du modèle, on peut comparer les hommes et les femmes du corpus québécois: comme sur r/France, ces dernières produisent significativement plus d'émojis que les hommes, même si la taille d'effet est plus faible (1.79). Il n'y a pas de différence entre les hommes de r/Quebec et les femmes de r/France. Enfin, les femmes du corpus français utilisent significativement moins d'émojis que les femmes du corpus québécois; selon le modèle, quand une femme de r/Quebec produit un émoji, une femme de r/France en produit 0.65.

Coefficients	**Modèle 1: Twitter**	**Modèle 2: Reddit**
Intercept	0.030 (1.291, 1.732)**	0.0005 (0.0005, 0.001)**
Genre: femmes	1.495 (0.027, 0.032)**	2.382 (1.918, 2.983)**
Pays: Québec	-	2.015 (1.762, 2.308)**
Interaction genre (femmes) et pays (Québec)	-	0.754 (0.524, 1.091)

Figure 7: Modèles de régression binomiale négative. Légende: ** $p < .001$; * $p < .01$

3.2.2 Émoticônes

Dans le corpus de r/France, les émoticônes ont été utilisées au moins une fois par 56.46% des femmes et 52.35% des hommes. Les proportions sont similaires dans le corpus de Reddit québécois, avec 55.09% d'utilisatrices uniques et 57.59% d'utilisateurs uniques. Les femmes ont utilisé davantage d'émoticônes que les hommes dans les deux corpus, comme le montre la Figure 8, avec des fréquences médianes de 0.17 pour les femmes des deux corpus, et de 0.11 et de 0.05 pour les hommes du corpus français et québécois. Les

intervalles de confiance indiquent toutefois d'importantes variations individuelles, surtout chez les femmes, qui sont beaucoup moins nombreuses que les hommes dans les corpus.

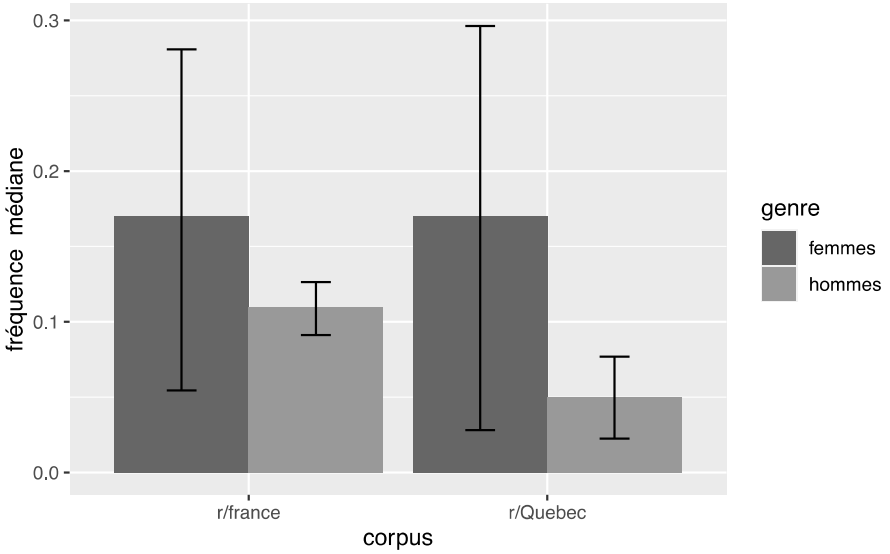

Figure 8: Fréquences relatives médianes pour 1000 tokens des émoticônes dans les corpus de Reddit

Même si la fréquence relative des émoticônes est similaire dans les corpus de tweets et dans les corpus de Reddit, il y a beaucoup moins d'utilisateurs et d'utilisatrices uniques d'émoticônes sur Twitter: seulement 10.02% de femmes, 10.94% d'hommes et 18.68% de personnes non binaires dans les tweets français, et 10.11% de femmes et 14.79% d'hommes dans les tweets québécois. Les fréquences médianes sont donc de 0 pour tous les groupes. Les fréquences moyennes suggèrent que les hommes ont employé davantage d'émoticônes que les femmes, en France comme au Québec: elles sont de 0.95 (*SD* = 5.52) pour les femmes en France, de 1.48 pour les hommes (*SD* = 11.34) et de 3.22 pour les personnes non binaires (*SD* = 21.14). Dans le corpus québécois, les émoticônes ont une fréquence moyenne relative de 1.73 (*SD* = 10.78) chez les femmes et de 2.73 chez les hommes (*SD* = 17.21). Encore une fois, la variation individuelle est forte, comme le montrent les écarts types. Le modèle de régression binomiale négative indique que les effets du genre et du pays sont tous les deux significatifs sur Reddit (Figure 9). Sur r/France comme sur r/Quebec, les femmes produisent davantage d'émoticônes que les hommes: le modèle indique qu'elles en produisent 1.51 quand les hommes en produisent 1. Les internautes de r/France ont utilisé davantage d'émoticônes que les internautes de r/Quebec.

	Émoticônes
Intercept	0.001 (0.001, 0.001)**
Genre: femmes	1.514 (1.345, 1.708)**
Pays: r/Quebec	0.827 (0.759, 0.902)**
Interaction genre (femmes) et pays (Québec)	-

Figure 9: Modèle de régression binomiale négative. Légende: ** $p < .001$; * $p < .01$

Comme près de neuf internautes sur dix n'ont pas utilisé d'émoticônes sur Twitter, nous avons créé un modèle de régression "zero-inflated", capable de gérer un important nombre de zéros (Hilbe 2014). Le modèle prédit l'effet du genre et du pays sur la fréquence des émoticônes parmi les personnes qui ont utilisé des émoticônes, et la probabilité que les internautes produisent une émoticône. Le modèle (Figure 10) montre que les hommes et les personnes non binaires produisent significativement moins d'émoticônes que les femmes, et qu'il y a moins d'émoticônes dans le corpus québécois que dans le corpus français. Il indique aussi que les hommes sont moins susceptibles d'utiliser au moins une émoticône que les femmes, en France comme au Québec, et que les internautes québécois sont moins susceptibles d'en produire que les internautes français.

	Émoticônes
Modèle "count"	
Intercept	0.002 (-6.095, -5. 841)**
Hommes	0.722 (-0.474, -0.177)**
Non binaires	0.560 (-1.002, -0.158)**
Québec	0.512 (-0.810, -0.527)**
Modèle "zero"	
Intercept	0.004 (-5.933, -5.191)**
Hommes	0.505 (-1.150, -0.216)**
Non binaires	0.292 (-2.342, -0.120)
Québec	0.288 (-1.830, -0.658)**

Figure 10: Modèle de régression "zero inflated". Légende: ** $p < .001$; * $p < .01$

3.3 Effet de l'âge sur l'utilisation des émojis

Nous avons uniquement analysé l'effet de l'âge sur la fréquence des émojis dans le corpus de tweets français, qui est le seul à comporter une annotation d'âge. Pour cette analyse, nous n'avons pas pris en compte les personnes non binaires, car elles n'étaient pas assez nombreuses dans le corpus. Nous avons créé trois groupes d'âge: les adolescent·es (11 à 18 ans, $N = 328$), les jeunes

adultes (23 à 27 ans, *N* = 860) et les adultes (31 ans et plus, *N* = 299). Les fréquences relatives médianes des émojis par groupes d'âge et de genre sont présentées dans un graphique d'interaction (Figure 11).

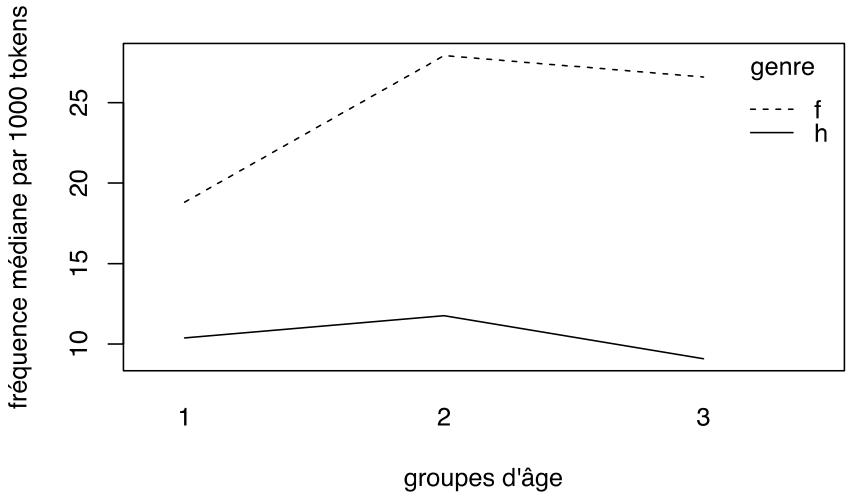

Figure 11: Fréquence médiane des émojis par 1000 tokens dans les corpus de tweets

Il semble y avoir un effet de l'âge chez les femmes: les adolescentes ont produit moins d'émojis que les deux autres groupes d'âge. Lors de la création du modèle de régression négative binomiale, le processus de sélection des variables a toutefois uniquement retenu l'effet du genre. (Figure 12). Nous pouvons en conclure que l'âge n'est pas corrélé avec l'utilisation des émojis dans le corpus de tweets français.

	Émojis
Intercept	0.001 (0.001, 0.001)**
Genre: femmes	1.547 (1.030, 2.328)*

Figure 12: Modèle de régression. Légende: ** $p < .001$; * $p < .01$

4. Conclusions

Nos analyses ont révélé plusieurs différences significatives dans l'utilisation des émoticônes et émojis sur Twitter et Reddit. Il y a tout d'abord une variation liée au type de plateforme: les émojis sont considérablement plus fréquents sur Twitter que sur Reddit. Il semble que Twitter, peut-être par sa limite de caractères, est plus propice à l'utilisation d'émojis et probablement d'autres phénomènes typiques du français d'internet, comme les acronymes, abréviations et autres graphies non standard. Nos résultats suggèrent également que, sur Twitter, les émojis ont supplanté les émoticônes, qui sont utilisées par environ 10% des utilisatrices et des utilisateurs seulement. Le "déclin" des émoticônes au profit des émojis sur Twitter a été noté dans une

étude longitudinale (Pavalanathan & Eisenstein 2016); de plus, les émoticônes sont 12 fois moins fréquentes dans notre corpus français que dans un corpus de tweets italiens datant de 2012 et 2013 (Spina 2019). Il est possible que les émoticônes ne fassent plus partie du répertoire d'une majorité d'utilisateurs de Twitter parce que ceux-ci écrivent généralement leurs tweets depuis un smartphone ou une tablette, qui proposent des claviers d'émojis. En effet, 81.96% des tweets du corpus français et 80.67% du corpus québécois ont été mis en ligne depuis une application mobile, selon les données recueillies grâce à l'API de Twitter. En revanche, les émojis ne sont pas plus fréquents que les émoticônes sur r/Quebec, et sont moins fréquents qu'elles sur r/France. Cela peut être dû à plusieurs facteurs: le fait qu'il n'y ait pas de limite de caractères pour les commentaires sur Reddit, le fait que les internautes ne semblent pas majoritairement accéder au site depuis un smartphone, ou encore le style d'écriture "geek" typique de Reddit, qui désapprouve l'usage d'émojis (Flesch 2020).

En ce qui concerne les types d'émojis et d'émoticônes, on note moins de variations dans les émoticônes que dans les émojis, ce qui est sans doute lié au fait que nous avons utilisé une liste d'émoticônes plus réduite que la liste d'émojis d'Unicode; nous n'avons pas identifié toutes les émoticônes utilisées, dont les kaomojis, qui représentent des visages de façon verticale (>.<, par exemple). Nos résultats montrent que les émoticônes "sans nez" sont beaucoup plus fréquentes que les émoticônes avez nez, ce qui reflète une tendance à l'abréviation notée par d'autres études (Oleszkiewicz et al. 2017; Schnoebelen T. 2012; Thompson & Filik 2016). Par ailleurs, malgré le vaste répertoire d'émojis à leur disposition, les utilisatrices et utilisateurs privilégient deux catégories: les "smileys et émotions" et les "visages et personnes".

Nos résultats révèlent une variation géographique. Sur Reddit, les internautes du corpus québécois produisent plus d'émojis que les internautes de r/France. Sur Reddit comme sur Twitter, ils utilisent moins d'émoticônes que ceux de France. Cela pourrait indiquer une différence entre le français d'internet québécois et le français d'internet de France; des différences d'usage au sein des espaces francophones ont en effet été mises en évidence par Cougnon (2010) dans des corpus de SMS couvrant la France, la Belgique, la Suisse et le Québec. Toutefois, les corpus de Twitter étant considérablement plus petits que les corpus de Reddit, il faut interpréter ce résultat avec précaution. Sur Reddit, il pourrait également refléter une variation interne au site, liée à des différences de pratiques linguistiques entre deux communautés. Les subreddits peuvent en effet constituer de véritables communautés de pratique, qui ont chacune un "répertoire de normes sociales et linguistiques" (Leuckert & Leuckert 2020: 36).

Dans tous les corpus, nos résultats montrent qu'il existe d'importantes variations individuelles dans la fréquence des émojis et émoticônes, avec une faible proportion de personnes qui les utilisent très fréquemment. Il y a une forte variation intra-genre, ainsi que des similitudes entre femmes et hommes: par

exemple, sur Twitter, la proportion de femmes et d'hommes qui utilisent des émoticônes est similaire. Les modèles de régression, capables de gérer ces variations, indiquent toutefois qu'émoticônes et émojis sont plus fréquemment produits par les femmes que par les hommes, comme l'ont montré de nombreuses autres études (Baron 2004; Chen et al. 2017; Coats 2017; Flesch 2020; Fullwood et al. 2013; Holtgraves 2011; Spina et al. 2013). Il est possible que les émojis et émoticônes soient pour les femmes une façon de construire leur identité de genre en ligne, en utilisant des marqueurs de féminité. Puisque ces procédés ont entre autres pour rôle de signaler l'empathie et de renforcer les liens sociaux, il est aussi possible que les femmes effectuent davantage de travail relationnel dans les interactions (Fishman 1978). En effet, comme le montre la faible proportion de femmes dans nos corpus, internet n'a pas effacé les inégalités, qui peuvent se refléter dans les pratiques de communication. Les hommes investissent davantage les espaces en ligne que les femmes, surtout sur les plateformes où sont discutés des sujets qui relèvent de la sphère publique comme la politique et l'actualité (Bode 2017; Meyer & Carey 2015; Peacock & Duyn 2021), et dont font partie Twitter, r/France et r/Quebec. Dans ces contextes très masculins, les femmes pourraient faire une "maintenance conversationnelle" (Maltz & Borker 1982: 202) plus importante que les hommes, notamment en utilisant davantage d'émoticônes et d'émojis.

Enfin, notre analyse du corpus français de tweets suggère que l'âge n'a pas d'effet significatif sur la fréquence des émojis, alors que les adolescent·e.s produisent généralement davantage de procédés typiques de la langue d'internet (Tagliamonte 2016). Ce résultat peut être interprété de deux façons différentes: soit il n'y a jamais eu d'effet de l'âge sur la fréquence des émojis dans le français d'internet, soit les habitudes des internautes sont en train de changer. Il est possible que les adolescent·e.s de 2022 aient délaissé les émojis, peut-être au profit d'autres formes du français d'internet.

BIBLIOGRAPHIE

Alloing, C. & Pierre, J. (2021): L'usage des émoji sur Twitter: Une grammaire affective entre publics et organisations? Communication & Langages, 2, 269-298. https://doi.org/10.3917/comla1.208.0269

Baron, N. S. (2004): See you online: Gender issues in college student use of Instant Messaging. Journal of Language and Social Psychology, 23(4), 397-423. https://doi.org/10.1177/0261927X04269585

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S. & Matsuo, A. (2018): quanteda: An R package for the quantitative analysis of textual data. Journal of Open Source Software, 3(30), 774. https://doi.org/10.21105/joss.00774

Bode, L. (2017): Closing the gap: Gender parity in political engagement on social media. Information, Communication & Society, 20(4), 587-603. https://doi.org/10.1080/1369118X.2016.1202302

Canty, A. & Ripley, B. D. (2022): boot: Bootstrap R (S-plus) functions [R].

Cassell, J., Huffaker, D., Tversky, D. & Ferriman, K. (2006): The language of online leadership: Gender and youth engagement on the Internet. Developmental Psychology, 42(3), 436-449. https://doi.org/10.1037/0012-1649.42.3.436

Chen, Z., Lu, X., Shen, S., Ai, W., Liu, X. & Mei, Q. (2017): Through a gender lens: An empirical study of emoji usage over large-scale Android users. arXiv:1705.05546 [cs]. http://arxiv.org/abs/1705.05546

Coats, S. (2017): Gender and lexical type frequencies in Finland Twitter English. Studies in Variation, Contacts and Change in English, 19. http://www.helsinki.fi/varieng/series/volumes/19/coats/

Cougnon, L.-A. (2010): Orthographe et langue dans les SMS. Conclusions à partir de quatre corpus francophones. Éla. Études de linguistique appliquée, 160(4), 397-410. https://doi.org/10.3917/ela.160.0397

Countries with most Twitter users 2022. (2023): Statista. https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/

Del-Teso-Craviotto, M. (2008): Gender and sexual identity authentication in language use: The case of chat rooms. Discourse Studies, 10(2), 251-270. https://doi.org/10.1177/1461445607087011

Derks, D., Bos, A. E. R. & von Grumbkow, J. (2008): Emoticons in computer-mediated communication: Social motives and social context. CyberPsychology & Behavior, 11(1), 99-101. https://doi.org/10.1089/cpb.2007.9926

Dresner, E. & Herring, S. C. (2010): Functions of the nonverbal in CMC: Emoticons and illocutionary force. Communication Theory, 20(3), 249-268. https://doi.org/10.1111/j.1468-2885.2010.01362.x

Emoji Counts, v15.0. (s. d.): Unicode. https://unicode.org/emoji/charts/emoji-counts.html

Fishman, P. M. (1978): Interaction: The work women do. Social Problems, 25(4), 397-406. https://doi.org/10.2307/800492

Flesch, M. (2020): Lol thats how reddit talks ;). Le site américain Reddit comme espace de variation de l'anglais. Étude de corpus intersectionnelle et quantitative d'usages non standard, au prisme du genre, de l'âge et de l'ethnicité (Université de Lorraine): https://www.theses.fr/2020LORR0192

Fullwood, C., Orchard, L. J. & Floyd, S. A. (2013): Emoticon convergence in Internet chat rooms. Social Semiotics, 23(5), 648-662. https://doi.org/10.1080/10350330.2012.739000

Fullwood, C., Quinn, S., Chen-Wilson, J., Chadwick, D. & Reynolds, K. (2015): Put on a Smiley Face: Textspeak and Personality Perceptions. Cyberpsychology, Behavior, and Social Networking, 18(3), 147-151. https://doi.org/10.1089/cyber.2014.0463

Garrison, A., Remley, D., Thomas, P., & Wierszewski, E. (2011): Conventional faces: Emoticons in instant messaging discourse. Computers and Composition, 28(2), 112-125. https://doi.org/10.1016/j.compcom.2011.04.001

Guntuku, S. C., Li, M., Tay, L. & Ungar, L. H. (2019): Studying cultural differences in emoji usage across the east and the west. Proceedings of the international AAAI conference on web and social media, 13, 226-235. https://doi.org/10.1609/icwsm.v13i01.3224

Halté, P. (2017): Positionnement syntaxique des interjections et des émoticônes: Modalisation, portée, visée. Cahiers de Praxématique, 69. https://doi.org/10.4000/praxematique.4680

Haukoos, J. S. & Lewis, R. J. (2005): Advanced statistics: Bootstrapping confidence intervals for statistics with "difficult" distributions. Academic Emergency Medicine, 12(4), 360-365. https://doi.org/10.1197/j.aem.2004.11.018

Hilbe, J. M. (2014): Modeling count data. Cambridge University Press.

Holtgraves, T. (2011): Text messaging, personality, and the social context. Journal of Research in Personality, 45(1), 92-99. https://doi.org/10.1016/j.jrp.2010.11.015

Hu, L. & Kearney, M. W. (2021): Gendered tweets: Computational text analysis of gender differences in political discussion on Twitter. Journal of Language and Social Psychology, 40(4), 482-503. https://doi.org/10.1177/0261927X20969752

Hvitfeldt, E. (2022): Emoji [R]. https://emilhvitfeldt.github.io/emoji/

Kavanagh, B. (2016): Emoticons as a medium for channeling politeness within American and Japanese online blogging communities. Language & Communication, 48(Supplement C), 53-65. https://doi.org/10.1016/j.langcom.2016.03.003

Kearney, M. W. (2019): rtweet: Collecting and analyzing Twitter data. Journal of Open Source Software, 4(42), 1829. https://doi.org/10.21105/joss.01829

Koch, T. K., Romero, P. & Stachl, C. (2022): Age and gender in language, emoji, and emoticon usage in instant messages. Computers in Human Behavior, 126, 106990. https://doi.org/10.1016/j.chb.2021.106990

Leuckert, S. & Leuckert, M. (2020): Towards a digital sociolinguistics. Communities of practice on Reddit. In S. Rüdiger & D. Dayter (éds.), Corpus approaches to social media. Amsterdam (John Benjamins Publishing Company), 15-40

List of emoticons. (2020): In Wikipedia. https://en.wikipedia.org/w/index.php?title=List_of_emoticons&oldid=949712309

Magué, J.-P., Rossi-Gensane, N. & Halté, P. (2020): De la segmentation dans les tweets: Signes de ponctuation, connecteurs, émoticônes et émojis. Corpus, 20. https://doi.org/10.4000/corpus.4619

Maltz, D. N. & Borker, R. A. (1982): A cultural approach to male-female miscommunication. In A cultural approach to interpersonal communication: Essential readings. Wiley Malden.

Meyer, H. K., & Carey, M. C. (2015): Men more likely to post online newspaper comments. Newspaper Research Journal, 36(4), 469-481. https://doi.org/10.1177/0739532915618417

Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P. & Rosenquist, J. N. (2011): Understanding the demographics of Twitter users. ICWSM. https://doi.org/10.1609/icwsm.v5i1.14168

Na'aman, N., Provenza, H. & Montoya, O. (2017): Varying linguistic purposes of emoji in (Twitter) context. Proceedings of ACL 2017, Student research workshop, 136-141.

Oleszkiewicz, A., Karwowski, M., Pisanski, K., Sorokowski, P., Sobrado, B. & Sorokowska, A. (2017): Who uses emoticons? Data from 86702 Facebook users. Personality and Individual Differences, 119(Supplement C), 289-295. https://doi.org/10.1016/j.paid.2017.07.034

Original Board Thread in which:-) was proposed. (s. d.): https://www.cs.cmu.edu/~sef/Orig-Smiley.htm

Pavalanathan, U. & Eisenstein, J. (2016): More emojis, less:) The competition for paralinguistic function in microblog writing. First Monday, 21(11): https://doi.org/10.5210/fm.v21i11.6879

Peacock, C. & Duyn, E. V. (2021): Monitoring and correcting: Why women read and men comment online. Information, Communication & Society, 1-16. https://doi.org/10.1080/1369118X.2021.1993957

Pérez-Sabater, C. (2013): The linguistics of social networking: A study of writing conventions on Facebook. Linguistik Online, 56(6): https://doi.org/10.13092/lo.56.257

Podolak, M. (2021, avril 9): How to scrape large amounts of Reddit data. The Startup. https://medium.com/swlh/how-to-scrape-large-amounts-of-reddit-data-using-pushshift-1d33bde9286

Provine, R. R., Spencer, R. J. & Mandell, D. L. (2007): Emotional expression online: Emoticons punctuate website text messages. Journal of Language and Social Psychology, 26(3), 299-307. https://doi.org/10.1177/0261927X06303481

R Core Team. (2021): R: A language and environment for statistical computing. https://www.R-project.org/

Reddit - Press. (2021): Reddit by the numbers. Redditinc.com. https://www.redditinc.com/press

Réseaux sociaux les plus utilisés 2023. (2023): Statista. https://fr.statista.com/statistiques/570930/reseaux-sociaux-mondiaux-classes-par-nombre-d-utilisateurs/

Rezabeck, L. L. & Cochenour, J. J. (1995): Emoticons: Visual Cues for Computer-Mediated Communication. http://eric.ed.gov/?id=ED380096

Sampietro, A., Felder, S. & Siebenhaar, B. (2022): Do you kiss when you text? Cross-cultural differences in the use of the kissing emojis in three WhatsApp corpora. Intercultural Pragmatics, 19(2), 183-208.

Schnoebelen, J. (2012): Emotions are relational: Positioning and the use of affective linguistic resources (Stanford University): https://stacks.stanford.edu/file/druid:fm335ct1355/Dissertation_Schnoebelen_final_8-29-12-augmented.pdf

Schnoebelen, T. (2012): Do you smile with your nose? Stylistic variation in twitter emoticons. University of Pennsylvania Working Papers in Linguistics, 18(2), 118-125.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E. P. & Ungar, L. H. (2013): Personality, gender, and age in the language of social media: The open-vocabulary approach. PLoS ONE, 8(9), e73791. https://doi.org/10.1371/journal.pone.0073791

Settanni, M. & Marengo, D. (2015): Sharing feelings online: Studying emotional well-being via automated text analysis of Facebook posts. Frontiers in Psychology, 6. https://doi.org/10.3389/fpsyg.2015.01045

Skovholt, K., Grønning, A. & Kankaanranta, A. (2014): The communicative functions of emoticons in workplace e-Mails::-): Journal of Computer-Mediated Communication, 19(4), 780-797. https://doi.org/10.1111/jcc4.12063

Spina, S. (2019): Role of emoticons as structural markers in Twitter interactions. Discourse Processes, 56(4), 345-362. https://doi.org/10.1080/0163853X.2018.1510654

Spina, S., Cancila, J. & others. (2013): Gender issues in the interactions of Italian politicians on Twitter: Identity, representation and flows of conversation. International Journal of Cross-Cultural Studies and Environmental Communication, 2(2), 147-157.

SUTOM. (s. d.): Sutom. Consulté le 11 février 2023, à l'adresse https://sutom.nocle.fr/

Tagg, C. (2012): Discourse of text messaging: Analysis of SMS communication. Londres (Continuum):

Tagliamonte, S. A. (2016): So sick or so cool? The language of youth on the internet. Language in Society, 45(01), 1-32. https://doi.org/10.1017/S0047404515000780

Thompson, D. & Filik, R. (2016): Sarcasm in written communication: Emoticons are efficient markers of intention. Journal of Computer-Mediated Communication, 21(2), 105-120. https://doi.org/10.1111/jcc4.12156

Tsou, A. (2016): How does the front page of the Internet behave? Readability, emoticon use, and links on Reddit. First Monday, 21(11): https://doi.org/10.5210/fm.v21i11.7013

Vandergriff, I. (2013): Emotive communication online: A contextual analysis of computer-mediated communication (CMC) cues. Journal of Pragmatics, 51, 1-12. https://doi.org/10.1016/j.pragma.2013.02.008

Venables, W. N. & Ripley, B. D. (2002): Modern Applied Statistics with S (Fourth edition): Springer.

VifEspoirPirez. (2022, janvier 17): Forum Libre [Reddit Comment]. r/france. www.reddit.com/r/france/comments/s5wivg/forum_libre_20220117/ht0k9eg/

Witmer, D. F. & Katzman, S. L. (1997): On-line smiles: Does gender make a difference in the use of graphic accents? Journal of Computer-Mediated Communication, 2(4): http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.1997.tb00192.x/full

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

Zuur, A. F., Hilbe, J. M. & Ieno, E. N. (2015): A beginner's guide to GLM and GLMM with R: A frequentist and Bayesian perspective for ecologists. Newburgh (Highland Statistics):

Annexes

Annexe 1. Liste des requêtes utilisées pour identifier l'identité de genre des internautes dans les corpus de Reddit

Pour les hommes:
"je suis un homme", "je suis un mec", "je suis un garçon", "je suis un gars", "je suis papa", "je suis certain", "je suis anxieux", "je suis confiant", "je suis content", "je suis conscient", "je suis inquiet", "je suis curieux", "je suis étudiant", "je suis français", "je suis heureux", "je suis inscrit", "je suis jaloux", "je suis prêt", "je suis surpris", "je suis vieux", "je suis satisfait", "je suis ouvert", "je suis dubitatif", "je suis preneur", "je suis innocent", "je suis végétarien", "je suis amateur", "je suis le seul", "je suis idiot", "je suis con", "je suis reconnaissant", "je suis très heureux", "je suis très surpris", "je suis très conscient", "je suis très content", "je suis très satisfait", "je suis grand", "je suis très grand"

Pour les femmes:
"je suis une femme", "je suis une nana", "je suis une meuf", "je suis une fille", "je suis maman", "je suis certaine", "je suis anxieuse", "je suis confiante", "je suis contente", "je suis consciente", "je suis inquiète", "je suis curieuse", "je suis étudiante", "je suis française", "je suis heureuse", "je suis inscrite", "je suis jalouse", "je suis prête", "je suis surprise", "je suis vieille", "je suis satisfaite", "je suis ouverte", "je suis dubitative", "je suis preneuse", "je suis innocente", "je suis végétarienne", "je suis amatrice", "je suis la seule", "je suis idiote", "je suis conne", "je suis reconnaissante", "je suis très heureuse", "je suis très surprise", "je suis très consciente", "je suis très contente", "je suis très satisfaite", "je suis grande", "je suis très grande"

Annexe 2. Expressions régulières utilisées pour identifier l'âge et les pronoms des internautes (syntaxe R)

Âge:
"\\d\\d\\sy|\\d\\d\\|\\d\\d\\||\\d\\dy|\\d\\dyo|\\d\\d\\syo|\\d\\d\\sy/o|\\d\\dy/o|\\d\\dy.o|\\d\\d\\sy.o|\\d\\d years old|\\d\\d\\sans|\\d\\da|\\d\\d\\sa|\\d\\d-|\\d\\d\\|"

Pronoms non binaires: "he/she|he / she|he \\| she|he\\|she|she/he\\b|she /he\\b|she \\| he\\b|she\\|he\\b|\\biel\\b|elle/iel|elle / iel|elle\\|iel|elle \\| iel|il/iel|il / iel|il\\|iel|il \\| iel|he/they|non-binaire|she/they|elle/iel|they/them|they / them|they \\| them|they\\|them|non binary|iel/il|iel / il|iel\\|il|iel \\| il|iel/elle|iel / elle|iel\\|elle|iel \\| elle|\\bielle\\b"

Pronoms de femmes: "she/her|she / her|she \\| her|she\\|her|elle/elle|elle / elle\\b|elle\\|elle\\b|elle \\| elle\\b"

Pronoms d'hommes: "he/him|he / him|he \\| him|he\\|him|il/lui |il / lui\\b|il \\| lui\\b|il\\|lui\\b"