Corpus de tweets et de SMS annotés pour l'observation de phénomènes linguistiques en français « non standard »

Louise Tarrade[1] Cédric Lopez[1]
(1) Viseo R&D, 4 avenue doyen Louis Weil, 38000 Grenoble, France

RESUME

Les tweets et les SMS ont pour spécificité d'être des textes comportant des phénomènes linguistiques qui dérogent aux règles normées de la langue. La multitude de ces phénomènes nous a conduit à développer une typologie spécifique à ce genre de texte que nous avons utilisée pour annoter un corpus composé de 1000 SMS et 1000 tweets. Un tel corpus annoté constitue un apport d'intérêt pour le TAL et les études sociolinguistiques. Dans cet article, nous présentons ce corpus annoté selon des phénomènes linguistiques d'ordre morpho-lexical et morpho-syntaxique et nous en dressons un premier bilan.

ABSTRACT

Tweets and SMS corpus annotated for the observation of linguistic phenomena in "non-standard" French.
Tweets and SMS are texts with the specificity of not respecting language norms. The multitude of these phenomena led us to develop a specific typology that we used to annotate a corpus consisting of 1000 SMS and 1000 tweets. Such a corpus is of interest specifically for NLP and sociolinguistic studies. In this paper, we present this annotated corpus according to linguistic phenomena at a morpholexical and morphosyntactical level, drawing up an overview of the general annotations.

MOTS-CLES : Tweets, SMS, Typologies, Annotations, Phénomènes linguistiques
KEYWORDS: Tweets, SMS, Typologies, Annotations, Linguistic phenomena

1 Introduction

E-mails, messages de forums, chats, SMS, tweets, *etc.* constituent une forme de communication qui a la particularité de produire des textes souvent qualifiés de « non standard », car les règles normées de la langue y sont rarement respectées ce qui les rend difficilement analysables par les outils de traitement automatique de la langue. Par exemple, l'écriture SMS se caractérise par la présence de nombreuses formes scripturales : squelettes consonantiques ("slt" (salut)), apocopes ("ordi" (ordinateur)), substitutions phonétisées ("2m1" (demain)), binettes/emoji ("^^", ":)", ☺) — la liste est longue. Or, de tels textes sont le support de nombreuses applications en TAL, telles que l'extraction d'informations médicales à partir des SMS des patients (Stenner *et al.*, 2012), l'analyse d'opinions et de sentiments dans les tweets (Vinodhini & Chandrasekaran, 2012), ou encore la synthèse vocale (Ill & Ford, 2011).

Dans le cadre d'un projet de normalisation automatique de ces textes (du français non standard vers le français standard), nous avons construit un corpus annoté de tweets et de SMS en français, dont l'annotation porte sur les phénomènes linguistiques représentés dans deux typologies, l'une d'ordre morpho-lexical et l'autre d'ordre morpho-syntaxique. Le développement de ce corpus a été motivé par un double objectif : tout d'abord, repérer quels sont les phénomènes présents dans chaque type de messages et dans quelle proportion ils apparaissent. Ainsi, il sera possible d'évaluer la possibilité de développer un unique module de normalisation pour les deux types de messages. Avoir connaissance de leur fréquence permet également d'évaluer sur quels phénomènes les efforts doivent être fournis prioritairement en termes de développement de l'outil de normalisation automatique. De plus, de telles observations peuvent être utiles dans le cadre de recherches sociolinguistiques ainsi que dans des tâches du TAL, notamment pour guider le développement d'un outil de normalisation automatique du français non standard vers le français standard.

Dans cet article, nous rappelons[1] succinctement les typologies que nous avons développées pour l'annotation des phénomènes linguistiques dans les tweets et les SMS (section 2), puis nous présentons le protocole d'annotation (section 3). Enfin, nous présentons le corpus annoté (section 4).

2 Typologies

À partir des typologies de phénomènes linguistiques dédiées aux écrits « non standard » (Roche *et al.*, 2016), (Fairon *et al.*, 2006), (Cougnon *et al.*, 2013) et (Anis, 2004), nous avons développé deux typologies (Tarrade *et al.*, 2017) dans le cadre d'un projet d'élaboration d'un outil de normalisation automatique de tweets et de SMS. La première répertorie les différents phénomènes présents dans ce type de messages au niveau morpho-lexical, tandis que la deuxième porte sur le niveau morpho-syntaxique. Ces nouvelles typologies fusionnent les précédentes en représentant des phénomènes à prendre en compte dans la tâche de normalisation tels que l'hyper-segmentation (*toute fois→toutefois*), l'écrasement (*jsuis→chuis*) ou encore les cas de code-switching, de néologismes, jargons ou mots en verlan, mais aussi la notion de pointeurs (*i.e.* mentions et hashtags). Les modifications de la graphie complète ou partielle d'un mot sont également catégorisées en fonction de leur incidence sur la prononciation de leur équivalent standard ou non. Ces typologies et les raisons de leur création sont plus précisément décrites dans (Tarrade *et al*., 2017).

La première typologie, représentant des phénomènes du niveau morpho-lexical, est divisée en trois catégories : 1) la substitution recense les cas de « remplacement de la graphie ou une partie de la graphie par une autre » (Panckhurst, 2009), cette dernière pouvant préserver ou non la prononciation du mot standard (*c→c'est*, *ca→ça*, *ossi→aussi*), mais également d'autres cas de substitution comme les contractions (*p'tit→petit*) ou les mots en verlan (*wam→moi*), par exemple. 2) la réduction répertorie des phénomènes qui correspondent à « un enlèvement de certains caractères et résultent nécessairement en un nombre inférieur de caractères » (Panckhurst, 2009) et qui ne correspondent pas un cas de substitution (qui peut aussi résulter en un nombre inférieur de caractères), tels que l'agglutination (*jattends*), les abréviations sémantisées (*m→me*), les squelettes consonantiques (*dsl→désolé*), pour ne citer qu'eux. 3) les ajouts correspondent à l'augmentation du nombre de caractères (ajouts phonétisés (*namour→amour*), allongements (*suuuuper*)) ou à l'ajout d'éléments tels que des smileys, des emoji, des mentions ou des hashtags, et quelques autres symboles. Une vue des typologies utilisées est consultable en Figure 2.

[1] Les typologies et les choix effectués au moment de leur conception sont détaillés dans (Tarrade *et al.*, 2017)

La seconde typologie, de niveau morpho-syntaxique, n'a pas pour ambition de décrire l'ensemble des phénomènes morpho-syntaxiques présents dans les écrits non standard, mais d'apporter une couche d'information supplémentaire aux annotations morpho-lexicales. Par exemple, elle permet de préciser si un hashtag (#) joue un rôle syntaxique ou non dans la phrase, ou encore si une modification de la graphie partielle d'un mot est due à une inversion entre le participe passé et l'infinitif ou à une erreur d'accord. Cette typologie répertorie également des phénomènes qui nous semblent important à signaler lors de l'annotation, tels que les ellipses (lexicales ou grammaticales), l'absence de ponctuation, ou les troncations de texte (en particulier pour les tweets). Toutefois, il serait peut-être intéressant de l'élargir pour considérer également les formes syntaxiques non standard (« ramenez moi les » par exemple).

3 Protocole d'annotation

Les annotations ont été réalisées à l'aide de l'outil d'annotation Brat[2] (Stenetorp *et al.,* 2012). Cet outil a notamment été choisi car :

– il permet d'attribuer plusieurs annotations à un élément textuel donné (ce qui est indispensable pour indiquer différents phénomènes observables sur un mot ou un syntagme). Par exemple, « ke il » contiendra deux annotations : l'une indiquant qu'il s'agit d'un cas de voyelle non élidée, et l'autre indiquant qu'il s'agit également d'un cas de remplacement d'une partie de la graphie par une autre (dans l'exemple, « k » à la place de « qu »).
– il permet d'annoter les *offsets*, ce qui est nécessaire pour signaler certains phénomènes comme l'ellipse ou l'absence de ponctuation, entre autres.
– il donne la possibilité d'ajouter des « notes » rattachables à un mot ou syntagme : nous les utilisons pour indiquer la forme normalisée (*i.e.* en français standard).

Le corpus présenté dans la section suivante a été annoté entièrement manuellement, à partir d'une version stabilisée de nos typologies. Au cours de l'annotation, de légers remaniements ont été effectués dans la typologie d'ordre morpho-lexical : les phénomènes de contraction et de compactage ont été ajoutés et une distinction a été effectuée entre les substitutions de signes diacritiques entraînant une modification de la prononciation et celles la conservant. Afin de prendre en compte ces évolutions mineures, une deuxième phase d'annotation a été effectuée.

Indépendamment des typologies, au cours de l'annotation, certains choix étaient inévitables, tels que celui de considérer comme standard le non-emploi de la double négation, de ne proposer de normalisation qu'aux mentions et hashtags jouant un rôle syntaxique dans la phrase ou de considérer comme standard les termes présents dans le dictionnaire. La tâche ayant une part de subjectivité, un des points importants de l'annotation était de respecter une cohérence dans les choix effectués.

Les normalisations en français standard ont été ajoutées simultanément aux annotations. Pour chaque lexie non-standard annotée, son équivalent normalisé est indiqué sous forme d'une note textuelle (voir point précédent). Par forme non-standard nous entendons toute forme non présente dans le dictionnaire Larousse en ligne[3]. Les lexies n'ayant pas d'entrée dans le dictionnaire ont été soumises à une normalisation tenant compte du contexte d'un point de vue à la fois syntaxique et sémantique.

[2] http://brat.nlplab.org/
[3] http://www.larousse.fr/dictionnaires/francais-monolingue

4 Présentation du corpus annoté

Mille tweets et mille SMS ont été annotés. L'ensemble de ces messages constitue le corpus présenté dans cet article, dont les métadonnées sont synthétisées dans le Tableau 1.

Les tweets ont été collectés à l'aide de l'API Twitter dans le cadre de la compétition CAP (Lopez *et al*, 2017). Ces tweets « tout venant » ont été recueillis sans filtre spécifique. Le corpus, annoté manuellement, contient 4450 annotations, dont 1700 qui concernent le niveau morpho-syntaxique, 2741 le niveau morpho-lexical et 9 concernent des cas d'indécision. À noter que certains tweets présentent la particularité d'être tronqués car ils sont postés depuis une autre application (nous ne pouvons contrôler leur provenance) et sont limités à 140 caractères. Nous avons décidé de ne pas annoter ces phrases tronquées.

Les SMS proviennent du corpus *88milSMS* (Panckhurst *et al*., 2014). Les 1000 premiers SMS ont été retenus pour l'annotation. Le nombre d'annotations effectuées sur les SMS s'élève à 5296, dont 4036 au niveau morpho-lexical, 1259 au niveau morpho-syntaxique et 1 cas d'indécision. Les annotations sont donc plus nombreuses dans ce type de message, notamment au niveau morpho-lexical.

Corpus utilisés	SMS : sous corpus de 88milSMS	Tweets collectés avec l'API twitter
Auteurs des annotations	L. Tarrade	L. Tarrade
Auteurs du corpus	R. Panckhurst, C. Détrie, C. Lopez, C. Moïse, M. Roche, B. Verine	C. Lopez
Genre du corpus	SMS	Tweets
Type d'annotation	Phénomènes linguistiques (morpho-lexicaux, morpho-syntaxiques)	
Taille du corpus	1000	1000
Licence	CC BY	
Version actuelle	v. 1	
Site Web, gestion du développement et des corrections	À venir, contacter les auteurs.	
Publication de référence	TARRADE L. & LOPEZ C. (2017). Corpus de tweets et de SMS annotés pour l'observation de phénomènes linguistiques en français « non standard ». *TALN 2017, cet article.*	

TABLE 1 : Métadonnées du corpus

Notons que pour un même nombre de messages annotés, les SMS contiennent plus d'annotations que les tweets, avec 4450 annotations pour les tweets et 5296 pour les SMS (Table 2).

Au niveau morpho-lexical, les SMS comportent proportionnellement plus de réductions que les tweets (29% contre 17%) (Figure 1). Si dans les tweets nous comptons plus de cas d'ajout que de réduction (41% contre 17%), c'est l'inverse dans les SMS qui ne comptent que 16% de phénomènes d'ajout pour 29% de phénomènes de réduction. Cela s'explique notamment par le fait que les tweets

contiennent une grande quantité d'hashtags (#) et de mentions (@), absents dans les SMS. Les phénomènes de substitution se révèlent les plus fréquents dans ces deux types de messages.

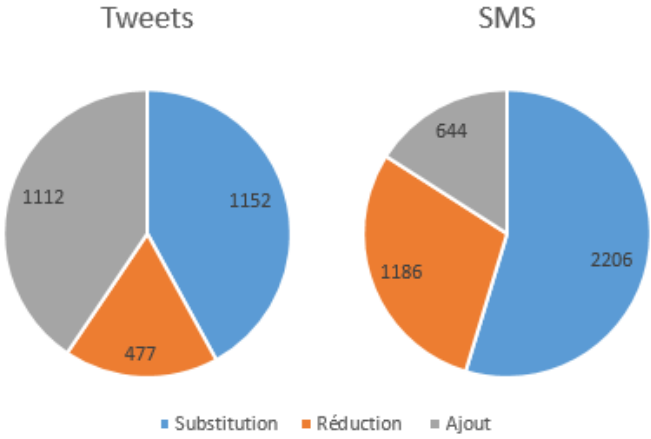

FIGURE 1 : Répartition des phénomènes morpho-lexicaux en nombre d'annotations

A propos des phénomènes de substitution, leur fréquence est assez similaire dans les deux types de messages. Cependant, les SMS ont une proportion extrêmement élevée (62%) de modifications d'une partie de la graphie d'un mot, contre 32% seulement pour les tweets. Cela peut en partie s'expliquer par le fait que les SMS contiennent un grand nombre de suppressions de signes diacritiques, contrairement aux tweets.

Dans notre corpus, les phénomènes de réduction sont bien plus nombreux dans les SMS (1186) que dans les tweets (477). Les SMS contiennent en effet un grand nombre d'agglutinations (*jattends* => *j'attends*), d'abrégements morpho-lexicaux (ce qui s'explique par le fait que l'apocope est très fréquente dans les SMS, contrairement aux tweets qui contiennent néanmoins proportionnellement plus de sigles ou d'acronymes que les SMS), d'abréviations et de squelettes consonantiques. Notons que les proportions des phénomènes dans les tweets et les SMS sont relativement similaires, si ce n'est une inversion des tendances entre les agglutinations (31% dans les SMS contre 28% dans les tweets) et les abrégements morpho-lexicaux (19% dans les SMS contre 37% dans les tweets) ainsi qu'une proportion bien plus élevée des squelettes consonantiques et des abréviations dans les SMS que dans les tweets.

Concernant les phénomènes d'ajout, le corpus annoté met en évidence la présence exclusive des pointeurs (hashtags et mentions) et des emoji dans les tweets, mais également la proportion écrasante des smileys dans les SMS par rapport aux tweets (59% contre 3%). Les phénomènes d'allongement (*suuuuuper* => *super*) sont également largement majoritaires dans les SMS (20% contre 5%). Quant aux autres phénomènes, ils sont répartis de façon homogène entre les deux types de messages, même si les ajouts phonétisés et les onomatopées/interjections sont plus fréquents dans les SMS que dans les tweets.

Niveaux	Cat. Phénomènes	Phénomènes	Tweets	SMS	Corpus total
Niveau morpho-lexical			**2741**	**4036**	6777
	Substitution		1152	2206	3358
		Graphie complète	66	134	200
		Graphie partielle	366	1375	1741
		Typographie	538	514	1052
		Rébus	30	3	33
		Écrasement	0	4	4
		Code-switching	53	111	164
		Néologisme/jargon	51	20	71
		Verlan	8	1	9
		Contraction	0	5	5
		Autre	40	39	79
	Réduction		477	1186	1663
		Abrégement morpho-lexical	177	220	397
		Abréviation sémantisée	16	84	100
		Squelette consonantique	41	184	225
		Agglutination	133	366	499
		Compactage	42	96	138
		Abréviation	64	212	276
		Autre	4	24	28
	Ajout		1112	644	1756
		Ajout phonétisé	2	20	22
		Allongement	56	130	186
		Smiley	35	381	416
		Emoji	238	0	238
		Onomatopée/interjection	69	87	156
		Hyper-segmentation	12	8	20
		Pointeur	666	0	666
		Symbole	17	8	25
		Autre	17	10	27
Niveau morpho-syntaxique			**1700**	**1259**	2959
		Sans rôle syntaxique	433	0	433
		Typographie et ponctuation	691	880	1571
		Conversion	9	3	12
		Inversion participe passé/infinitif	25	22	47
		Inversion mot grammatical	33	37	70
		Accord	112	137	249
		Ellipse	129	138	267
		Répétition	11	39	50
		Troncation texte	255	0	255
		Autre	2	3	5
Indécision			**9**	**1**	10
Total annotations			**4450**	**5296**	**9746**

TABLE 2 : vue d'ensemble du corpus en termes de nombre d'annotations pour chaque catégorie

Enfin, au niveau morpho-syntaxique, nous constatons sans grand étonnement que les problèmes de typographie et ponctuation sont majoritaires dans les deux types de messages, et que la catégorie

s*ans rôle syntaxique* n'est absolument pas représentée dans le corpus de SMS. Effectivement, les troncations de textes sont exclusives aux tweets (tronqués au-delà de 140 caractères), et seules les mentions (@), les hashtags (#) et les retweets (*RT*) peuvent être concernés par l'étiquette « sans rôle syntaxique » ; or, ces éléments sont absents des SMS annotés. Le nombre de « fautes » d'accord, d'ellipses et de répétitions est légèrement supérieur dans les SMS (respectivement, 11%, 11% et 3% dans les SMS contre 7%, 8% et 1% dans les tweets). Pour les autres phénomènes, ils se retrouvent à peu de choses près à proportions égales dans les deux types de messages.

5 Conclusion

Dans cet article, un corpus annoté constitué de 1000 tweets et 1000 SMS a été présenté. L'annotation porte d'une part sur les phénomènes linguistiques responsables de la transformation d'une lexie en son équivalent « non standard » et d'autre part sur le résultat de la normalisation en français standard.

Si la plupart des phénomènes sont communs aux deux types de messages, ils ne le sont cependant pas dans les mêmes proportions. L'observation de ce corpus annoté montre que les SMS sont moins conformes que les tweets au français standard et met en évidence les taux de représentation de chaque phénomène, qui nous permettront d'orienter le développement de l'outil de normalisation.

Notre prochaine étape consistera à utiliser ce corpus annoté pour développer un outil de normalisation automatique de tweets et de SMS.

Références

ANIS J., DE FORNEL, M., & FRAENKEL B. (organisateurs, 2004). La communication électronique : Approches linguistiques et anthropologiques. Colloque international, EHESS, Paris, 5-6 février 2004.

COUGNON L.-A., ROEKHAUT S., & BEAUFORT R. (2013). Typologies de variation graphique dans l'écrit SMS. S. Baddeley, F. Jejcic et C. Martinez (éd.)*, L'orthographe en quatre temps*, 20, 129-148.

FAIRON C., KLEIN J.-R., & PAUMIER S. (2006). *Le langage SMS : étude d'un corpus informatisé à partir de l'enquête « Faites don de vos SMS à la science »* (p. 31-47). Louvain-la-Neuve : UCL, Presses Univ. de Louvain.

ILL, E. T. G. & FORD, C. S. (2011). *U.S. Patent Application No. 12/983,946*.

KAUFMANN M. & KALITA J. (2010). Syntactic normalization of twitter messages. In *International Conference on Natural Language Processing*.

LOPEZ C., PARTALAS I., BALIKAS G., DERBAS N., MARTIN A., REUTENAUER C., SEGOND F., AMINI M.-R. (2017) French Named Entity Recognition in Twitter Challenge, In: Actes de CAP'17, à paraître.PANCKHURST R. (2009). Short Message Service (SMS) : typologie et problématiques futures., in : *Polyphonies, pour Michelle Lanvin*. Sous la dir. de Teddy Arnavielle. Université Paul-Valéry Montpellier 3, 33-52.

PANCKHURST R., DETRIE C., LOPEZ C., MOÏSE C., ROCHE M., VERINE B. (2014) « 88milSMS. A corpus of authentic text messages in French », produit par l'Université Paul-Valéry Montpellier III et le CNRS, en collaboration avec l'Université catholique de Louvain, financé grâce au soutien de la MSH-M et du Ministère de la Culture (Délégation générale à la langue française et aux langues de France) et avec la participation de Praxiling, Lirmm, Lidilem, Tetis, Viseo. ISLRN : 024-713-187-947-8 http://88milsms.huma-num.fr/

ROCHE M., VERINE B., LOPEZ C. & PANCKHURST R. (2016). « La néographie dans un grand corpus de SMS français : 88milSMS ». In : *La neología en las lenguas románicas Recursos, estrategias y nuevas orientaciones*, Actes du colloque *Cineo 2015*, 22-24 octobre, Salamanque. Sous la dir. de Joaquín García Palacios, Goedele De Sterck, Daniel Linder, Nava Maroto, Miguel Sánchez Ibáñez et Jesús Torres del Rey. Studien zur romanischen Sprachwissenschaft und interkulturellen Kommunikation. Frankfurt, Peter Lang.: DOI: http://dx.doi.org/10.3726/978-3-631-69859-4, 279-302.

STENETORP, P., PYYSALO, S., TOPIĆ, G., OHTA, T., ANANIADOU, S., & TSUJII, J. I. (2012). BRAT: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 102-107.

STENNER S. P., JOHNSON K. B., & DENNY J. C. (2012). PASTE: patient-centered SMS text tagging in a medication management system. *Journal of the American Medical Informatics Association*, *19*(3), 368-374.

TARRADE L., LOPEZ C., PANCKHURST R., ANTONIADIS G. (2017). Typologies pour l'annotation de textes non standard en français. *TALN 2017*.

VINODHINI, G., & CHANDRASEKARAN, R. M. (2012). Sentiment analysis and opinion mining: a survey. *International Journal*, *2*(6), 282-292.