Typologies pour l'annotation de textes non standard en français

Louise Tarrade[1] Cédric Lopez[1] Rachel Panckhurst[2] Georges Antoniadis[3]
(1) Viseo R&D, 4 avenue doyen Louis Weil, 38000 Grenoble, France
(2) Praxiling UMR 5267 CNRS & Université Paul-Valéry Montpellier 3, France
(3) LIDILEM, Université Grenoble Alpes, Grenoble, France
prenom.nom@viseo.com, rachel.panckhurst@univ-montp3.fr,
georges.antoniadis@univ-grenoble-alpes.fr

RESUME

La tâche de normalisation automatique des messages issus de la communication électronique médiée requiert une étape préalable consistant à identifier les phénomènes linguistiques. Dans cet article, nous proposons deux typologies pour l'annotation de textes non standard en français, relevant respectivement des niveaux morpho-lexical et morpho-syntaxique. Ces typologies ont été développées en conciliant les typologies existantes et en les faisant évoluer en parallèle d'une annotation manuelle de tweets et de SMS.

ABSTRACT

Typologies for the annotation of non-standard French texts.
The task of automatic normalization of messages resulting from mediated electronic communication requires a preliminary step consisting of identifying linguistic phenomena. In this paper, two typologies for annotating non-standard texts in French, from both morpholexical and morphosyntactic perspectives are proposed. We have elaborated these from existing typologies and modified them according to manual annotation of tweets and SMS.

MOTS-CLÉS : typologie, SMS, tweets, normalisation
KEYWORDS: typology, SMS, tweets, normalization

1 Introduction

Ces dernières années ont été marquées par l'avènement des messages textuels courts, notamment les tweets et les SMS. Les outils de traitement automatique de la langue sont généralement conçus pour traiter du texte « standard », or ces messages courts font l'objet d'écrits présentant des variantes de graphie qui s'éloignent de la langue standardisée, ce qui implique un traitement automatique difficile. De nombreuses applications nécessitent pourtant l'analyse de tels textes, par exemple 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). Afin de pallier cette difficulté il est nécessaire de transformer les écrits « non standard » en écrits « standard », *i.e.* de normaliser, ou transcoder (Beaufort & *al.* 2010). Un premier pas vers le développement d'un outil de normalisation consiste à dresser une typologie fine des phénomènes linguistiques apparaissant dans les écrits non standard. Une telle typologie permettra d'annoter des textes et ainsi d'identifier les phénomènes les plus fréquents dans un corpus

donné. Nous pourrons alors ensuite focaliser le travail sur le traitement des phénomènes, en fonction de leur fréquence, avec des modes de traitement adaptés.

Dans cet article, nous discutons des typologies déjà existantes (section 2), puis nous nous focalisons sur la construction d'une typologie représentant les phénomènes linguistiques présents dans les tweets et les SMS (section 3).

2 Travaux antérieurs

Le développement de typologies dans le contexte du discours numérique médié (désormais DNM) (Panckhurst, 2017) a logiquement suscité l'attention des chercheurs cette dernière décennie. Une des premières typologies répertoriant les variations graphiques et les aspects morpho-lexicaux de la DNM a été établie par (Anis, 2004). La typologie de (Fairon *et al.*, 2006) consiste en une classification générale de l'écriture SMS, incluant donc des phénomènes liés à l'orthographe phonétique, la morphosyntaxe, à la syntaxe et au discours. Nous avons besoin d'augmenter le degré de granularité considéré dans ces typologies afin de distinguer, par exemple, le cas des abréviations sémantisées (t→te/tu) (Roche *et al.* 2016), des squelettes consonantiques (dsl→désolé), tous deux classés parmi les abréviations. Plus récemment, (Cougnon *et al.*, 2013) proposent une typologie des stratégies réductionnelles de l'écriture SMS, afin de « *présenter les variations graphiques présentes dans l'écrit SMS* » (Cougnon *et al.*, 2013). Dans cette typologie, les phénomènes n'appartenant pas à des stratégies réductionnelles (les variations orthographiques) sont considérés, et l'on y inclut également des « erreurs » telles qu'une mauvaise utilisation du temps, l'inversion indicatif/impératif, l'accord du participe passé, *etc*., que nous réservons à une typologie d'ordre morpho-syntaxique. Enfin, la typologie publiée dans (Panckhurst, 2009), modifiée très récemment dans (Roche *et al.*, 2016) porte exclusivement sur les phénomènes de néographie de l'écriture SMS, en tenant compte des typologies précédentes. Son originalité repose entre autres sur le fait qu'elle distingue des phénomènes tels que la substitution, la réduction, la suppression ou l'ajout.

Ces différentes typologies ont principalement été développées dans un but descriptif et se sont focalisées sur les SMS. Dans la pratique, l'annotation manuelle fondée sur les typologies antérieures implique des choix subjectifs et/ou multiples. À partir des typologies citées précédemment, notre objectif est de 1) dresser une typologie dont les classes ont une intersection minimale afin de faciliter la tâche d'annotation manuelle, 2) étendre la typologie à des phénomènes apparaissant dans twitter, pour lequel il semble qu'aucune typologie n'ait été établie.

3 Conception et présentation de la typologie

Notre typologie concilie les typologies de (Roche *et al.*, 2016), (Fairon *et al.*, 2006), (Cougnon *et al.*, 2013) et (Anis, 2004) avec l'objectif d'une annotation dans le cadre d'une tâche de normalisation automatique. Nous exposons dans cette section notre méthodologie qui a conduit à la construction de notre typologie.

3.1 Vers une typologie modifiée pour le TAL

(Roche *et al.*, 2016) considèrent deux phénomènes possibles de suppression graphique : *typographie et ponctuation* (l'absence de ponctuation finale par exemple) et *signes diacritiques* (la suppression

des signes diacritiques) ; or, nous considérons la catégorie *signes diacritiques* comme étant plus appropriée à la définition de la substitution (remplacement de la graphie ou une partie de la graphie par une autre), et l'absence de ponctuation comme relevant plutôt du niveau syntaxique. De fait, notre typologie conserve trois catégories principales : substitution, réduction et ajout, que nous présenterons plus précisément dans la section suivante.

Dans un souci de démarcation plus nette des frontières entre les phénomènes, nous avons choisi de ne pas conserver la distinction effectuée par (Roche *et al.*, 2016) entre les variations graphiques et phonétisées. En effet, il nous semble important de repérer les modifications de graphies qui altèrent la prononciation du mot standard ou non, dans le cadre d'un éventuel recours à l'aspect phonétique du texte lors de la tâche de normalisation ; or, dans la typologie de (Roche *et al.,* 2016), la distinction phonétisé/graphique ne semble pas avoir cette vocation. Par exemple, le cas « mwa » (au lieu de « moi ») correspond à une substitution graphique avec variation dans la typologie de (Roche *et al.*, 2016) ; dans notre typologie, nous le considérons comme un cas de substitution de la graphie partielle d'un mot avec correspondance phonologique. Nous avons également réorganisé certains phénomènes : c'est le cas, par exemple, des abréviations de consonnes doubles ou de chute du « e » instable, de fin muette d'un mot ou de son début muet, répertoriés comme des cas de réductions par (Roche *et al.*, 2016), mais que nous préférons assimiler à des cas de substitutions graphiques. De la même façon, la réduction phonétisée entière (*c→ces*) nous paraît être un cas de substitution d'un mot par une lettre unique avec correspondance phonologique, étant l'initiale de celui-ci.

La tâche de normalisation automatique à laquelle nous aspirons nous impose d'avoir une typologie la plus couvrante possible, avec un niveau de description des phénomènes assez fin ; c'est pourquoi ont été intégrés à la typologie des phénomènes tels que l'écrasement décrit par (Anis, 2004) et (Fairon, 2006), mais aussi les possibilités de réduction, d'ajout ou de substitution autres que les phénomènes décrits dans notre typologie, qui peuvent s'assimiler à la majorité des notions d'omission, d'adjonction et de confusion décrites par (Cougnon *et al.*, 2013) dans la catégorie des erreurs orthographiques (*enregirstré→enregistré* ou *descentre→descendre* par exemple). Le phénomène d'hyper-segmentation (*toute fois→toutefois*), peu souvent inclus dans les différentes typologies, nous a paru important à ajouter. Ensuite, tout comme (Anis, 2004) qui proposait dans sa typologie les anglicismes, nous avons décidé de les inclure sous le prisme plus global du *code-switching*, dans sa plus large acception. Au même titre, nous avons ajouté les cas de mots en verlan. Les néologismes et jargons sont également présents dans notre typologie.

Des éléments apparaissant quasi exclusivement dans les tweets ont également été ajoutés, notamment les pointeurs, *i.e.* les mentions et les hashtags, qui peuvent par ailleurs jouer un rôle syntaxique. Cela nous a conduit à considérer le niveau morpho-syntaxique (*i.e.* incluant le niveau syntaxique). Nous considérons donc une seconde typologie, n'ayant pas pour but de décrire exhaustivement les phénomènes morpho-syntaxiques présents dans ce type de texte, mais d'apporter des précisions concernant le niveau morpho-lexical, sur lequel porte principalement notre typologie. Ceci permettra d'affiner les premières annotations et d'apporter un niveau supplémentaire d'information qui constituera une précieuse aide dans l'élaboration de l'outil de normalisation automatique.

Disposant d'un corpus de tweets obtenus grâce à l'API twitter et du corpus de SMS *88milSMS* (http://88milsms.huma-num.fr/ et https://hdl.handle.net/11403/comere/cmr-88milsms), nous avons annoté 1 000 tweets et 1 000 SMS afin de tester notre typologie et de la faire évoluer le cas échéant. À partir de ces annotations, portant sur les phénomènes décrits dans cette typologie, nous avons réadapté cette dernière en ajoutant les phénomènes de contraction et de compactage, ou en faisant la distinction entre les suppressions de signes diacritiques entraînant une approximation phonologique

ou non. Effectivement, l'annotation des SMS et des tweets a révélé que la suppression des signes diacritiques pouvait entrainer une modification de la prononciation du mot dans 43% des cas. Ces évolutions mineures ont été considérées dans une deuxième étape d'annotation manuelle qui nous a déjà permis d'observer, par exemple, une légère différence de proportion dans les phénomènes présents au niveau morpho-lexical dans les tweets et les SMS (Figure 1) (Tarrade & Lopez, 2017).

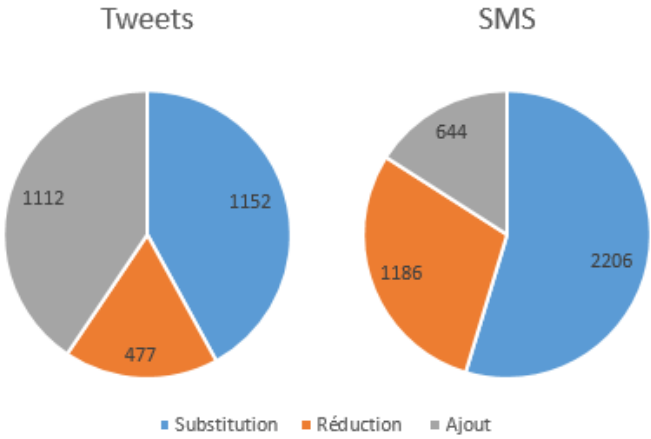

FIGURE 1: Répartition des phénomènes morpho-lexicaux en nombre d'annotations (Tarrade & Lopez, 2017)

Nous proposons ainsi deux typologies développées et mises à l'épreuve dans une tâche d'annotation manuelle pour deux types de textes différents (tweets, SMS), couvrant les niveaux morpho-lexicaux et morpho-syntaxiques.

3.2 Typologies

Nous décrivons dans cette section les catégories de nos typologies relevant respectivement du niveau morpho-lexical (Figures 2, 3 et 4) et morpho-syntaxique (Figure 5). Les exemples proviennent soit de (Roche *et al.*, 2016) et de (Fairon, 2006) pour les catégories communes, soit des corpus de tweets et de SMS que nous avons annotés.

3.2.1 Substitution (typologie morpho-lexicale)

Concernant la définition de la substitution, nous élargissons celle de (Panckhurst, 2009) « La substitution correspond à un remplacement de la graphie ou une partie de la graphie par une autre », en tenant compte du fait que les graphies de substitution peuvent ou non préserver la prononciation du mot, cette distinction étant indispensable dans l'éventualité où l'on souhaiterait traiter des cas de substitution à l'aide d'une étape de phonétisation. Les phénomènes de substitution peuvent donc concerner la graphie complète d'un mot ou groupe de mots, ou une partie de leur graphie. Il peut y avoir soit une correspondance phonologique entre la forme standard et la forme concernée par le phénomène, soit une approximation phonologique.

Substitution	graphie complète d'un mot ou groupe de mot	correspondance phonologique	lettre		c->s'est/ses/sais/sait, g->j'ai, i->y, k->qu'à/cas, l->elle, m->aime, n->haine, o->eau/au, q->cul, r->aire/air, s->est-ce, u->eu
				initiale	b->bé (aussi interjection), c->c'est/ces, d->des/dé/dès, e->eu, h->hache, j->j'y, o->oh, p->paix/pet, t->t'es/tes, v->vais
			chiffre		7->cet
			symbole		+ ->plus
		approximation phonologique			k->que, c->se/ça, z->je
	graphie partielle d'un mot	correspondance phonologique	lettre		bo->beau, Dcédé->décédé, ossi->aussi
			chiffre		2main->demain
			symbole		+sieurs->plusieurs
			graphie		bizes->bises, bisoux->bisous, mwa-moi
				consonne double	come->comme
				chute du "e" instable	douch->douche
				lettre muette enlevée	fou->fous, pa->pas, peu->peut, vou->vous
				lettre muette ajoutée	peux->peu, as->a
				initiale muette	ôtel->hôtel
			signe diacritique		voila->voilà, tantôt->tantot
		approximation phonologique			nan->non, kikou->coucou(avec graphie du 'c' modifiée), pô->pas, ui->oui
			signe diacritique		ca->ça, appelé->appele
	typographie				est ce que->est-ce que
		casse d'un caractère			suzanne -> Suzanne
		élision			m en->m'en
	rébus	caractère spécial			de grandes @ (oreilles)
		emoji			Merci pour vos RTs & 👏
	écrasement				jsuis->chuis (précédemment : agglutination+compactage)
	code-switching				yes -> oui, screen->écran
	néologisme/jargon				catche (compris), walou (rien)
	verlan				wam->moi
	contraction				p'tit->petit
	autre				chzton->chaton, les->la

FIGURE 2: Typologie morpho-lexicale : les cas de substitution

Toujours dans un souci d'adaptation à notre future tâche de normalisation automatique, il nous a paru nécessaire que la granularité de notre typologie soit assez fine. C'est pour cette raison que, comme dans la typologie proposée par (Fairon, 2006), une catégorie est dédiée aux différents types de modifications d'une graphie (ne correspondant pas à une autre possibilité du même niveau) ayant conservée une correspondance phonologique avec le mot d'origine. Sont alors distingués des cas particuliers tels que la substitution d'une consonne doublée par une consonne simple, la chute du « e » instable, de l'initiale muette ou de la lettre muette (qui peut également être ajoutée). Ces éléments sont classés pour la plupart comme des phénomènes de réduction dans les typologies de (Anis, 2004) et de (Roche *et al.*, 2016). Nous préférons les considérer comme des substitutions, car nous pensons cette méthode de classement plus adaptée à la tâche de normalisation, dans l'éventualité où la forme phonétique des mots serait utilisée dans celle-ci, par exemple.

Le phénomène de code-switching est considéré ici dans une large acception, comme c'est également le cas des néologismes ou jargons. Ces phénomènes, tout comme l'utilisation du verlan, sont considérés en tant que phénomènes de substitution. Effectivement, même s'ils n'obéissent pas à la définition stricte donnée précédemment de la substitution, ils correspondent toutefois à une substitution au sens large, car ils sont fréquemment utilisés à la place d'autres termes équivalents en français standard.

3.2.2 *Réduction (typologie morpho-lexicale)*

Réduction	Abrégement morpho-lexical	troncation	apocope		ordi -> ordinateur
			aphérèse		tain -> putain
		sigle/acronyme			"tlm"-> tout le monde, lol->laughing out loud
	abréviation sémantisée				b->ben (aussi interjection)/bien/bon, c->ce, d->de/du/dans, e->eh/et/est/être, g->gueule, h->heure, i->il/île, j->je/jeux, l->le/lesbienne/la, m->me/même/mais/mes/mètre, n->ne, p->page/peux/peut, q->que/qui, 't->tu/te/ton, u->universitaire, v->vœux/veut/veux/va/vas, w->week[-end]
	squelette consonantique				dsl->désolé, pr->pour
	agglutination				jattends -> j'attends
		avec élision			j'vois -> je vois
	compactage				jsuis->jesuis (précédemment : agglutination)
	abréviation				espr->espère
		consonne			dc->donc, cdlt->cordialement, RT->retweet
	autre				chaon->chaton

FIGURE 3: Typologie morpho-lexicale : les cas de réduction

La réduction « correspond à un enlèvement de certains caractères et résulte nécessairement en un nombre inférieur de caractères » (Panckhurst, 2009). Par ailleurs, d'un point de vue de la tâche de normalisation, il ne nous a pas paru pertinent de conserver une distinction entre les sigles et les acronymes. Cependant, préciser les catégories d'agglutination et d'abréviation nous semble approprié ; nous avons ajouté les cas d'agglutination avec élision de la première voyelle, ou encore les cas d'abréviations ne formant pas de squelettes consonantiques mais étant tout de même composées uniquement de consonnes du mot d'origine.

3.2.3 *Ajout (typologie morpho-lexicale)*

Ajout	phonétisé	avec variation			oki->ok, ouaip->ouais
		liaison			zèt->êtes, namour->amour
	allongement	ponctuation			*!!!!!!!!!!!!!!!
		lettre			suuuuuper->super
	smiley				:-)
	emoji				🙆🌠🙆 Bonne nuit à tous mes Amis 💎
	onomatopée/interjection				snif, bof
	hyper-segmentation	mot	espace		toute fois->toutefois
			ponctuation		ti-pe->type (avant : i->y (graphie))
		voyelle non élidée			que il->qu'il
	pointeur	hashtag			#Ronaldinho ne s'entraînait pas [...]
		mention			Pr @NicolasSarkozy la création de [...]
	symbole				la *star*
	autre				chzat->chat

FIGURE 4: Typologie morpho-lexicale : les cas d'ajout

L'ajout peut être défini comme l'augmentation du nombre de caractères ou d'éléments tels que des smileys, des emoji, des mentions ou des hashtags, et plus généralement des symboles quelconques. De façon plus fine que les typologies précédentes, nous spécifions si l'ajout concerne, par exemple, une voyelle non-élidée, une séparation du mot par une espace ou par une ponctuation (par exemple dans « S.U.P.E.R »).

3.2.4 Typologie morpho-syntaxique

La typologie morpho-syntaxique (Figure 5) a pour but principal d'apporter une couche d'information nécessaire lors de l'annotation. Deux phénomènes sont relatifs aux tweets : c'est le cas de la troncation de texte, fréquente dans les tweets limités à 140 caractères, et de la catégorie *sans rôle syntaxique*. Cette catégorie nous paraît indispensable ici, puisque comme le soulignent (Kaufmann & Kalita, 2010) les *hashtags* et les mentions ont pour particularité de ne pas toujours jouer un rôle syntaxique dans la phrase.

Niveau morpho-syntaxique					
	sans rôle syntaxique				Moi quand la bourse va arriver #CROUS, RT @Guigabriel92 : [...]
	typographie et ponctuation				guillemets, ponctuations finales, etc.
	conversion				sms-moi qud tu arriv
	Inversion participe passé/infinitif				Ben sa va arrivé->Ben ça va **arriver**
	Inversion des mots grammaticaux				pas comme sa-> pas comme **ça**
	accord	genre			quel annee -> **quelle** année
		nombre			je t'enverrai les photo -> je t'enverrai les **photos**
		personne			moi j'attend 18:15->moi j'**attends** 18:15
	ellipse	mot grammatical			à 17h y a mec dans le bus -> à 17h **il** y a **un** mec dans le bus
		mot lexical			il avec le chat
	Répétition				tu tu vois
	troncation du texte				RT @PlanBatiment : Le rêve pour Philippe Pelletier, la rénovation des bâtiments scolaires. Un @PlanBatiment scolaire ! Envie de porter cette ...
	Autre				

FIGURE 5: Typologie morpho-syntaxique

4 Conclusion

Dans ce travail, nous avons défini deux typologies pour l'annotation du français non standard, l'une morpho-lexicale et l'autre morpho-syntaxique, testées sur 1 000 tweets et 1 000 SMS vis-à-vis de la tâche de normalisation automatique. Les résultats des annotations ainsi que les différentes observations qui en découlent sont présentés de manière plus détaillée dans (Tarrade & Lopez, 2017). La prochaine étape, avant de commencer le développement de l'outil de normalisation, est d'identifier les phénomènes les plus représentés dans les différents corpus. Nous pourrons alors développer un outil de normalisation automatique de textes en français, prenant en compte le type de message et établissant des priorités au niveau des phénomènes. Dans une autre perspective, il serait également intéressant d'observer dans quelle mesure les phénomènes décrits dans ces typologies sont susceptibles de s'appliquer à d'autres langues.

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.

BEAUFORT R., ROEKHAUT S., COUGNON L. A., & FAIRON C. (2010, July). A hybrid rule/model-based finite-state framework for normalizing SMS messages. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 770-779). Association for Computational Linguistics.

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*.

PANCKHURST (2017), « Entre linguistique et informatique. Des outils de traitement automatique du langage naturel écrit (TALNE) à l'analyse du discours numérique médié (DNM) », habilitation à diriger des recherches, Université Paris-Est.

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, p. 33-52.

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, p.279-302.

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. (2017). Corpus de tweets et de SMS annotés pour l'observation de phénomènes linguistiques en français « non standard ». *TALN 2017.*

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