TREMoLo-Tweets corpus : guide d'annotation pour un corpus annoté en registres de langue pour le français

Jade Mekki, Delphine Battistelli, Gwénolé Lecorvé, Nicolas Béchet

Guide d'annotation pour un corpus annoté en registres de langue français

Jade Mekki[1,2] Delphine Battistelli[2] Gwénolé Lecorvé[1,3] Nicolas Béchet[4]

[1]Univ Rennes, CNRS, IRISA / Lannion-Vannes, France
[2]Universtité Paris Nanterre, CNRS, MoDyCo / Nanterre, France
[3]Orange Labs / Lannion, France
[4]Université de Bretagne Sud, CNRS, IRISA / Vannes, France
firstName.lastName @ {irisa.fr, orange.com, parisnanterre.fr}

Résumé

This work is part of the TREMoLo project [1] dedicated to language registers (casual, neutral, and formal). Here, we present an annotation guide grounded on a linguistic analysis of language registers and *Computer-Mediated Communications* (*CMCs*). It gives instructions for annotating French tweets according to the tripartition casual, neutral and formal. First, it specifies and defines the elements specific to CMCs. Then, it presents the protocol for annotating tweets in language registers. All annotation choices has to be justified by at least one linguistic descriptor. The complete list of linguistic descriptors is presented with examples at the end of the annotation guide.

1. http://tremolo.irisa.fr

Table des matières

Table des figures

Liste des tableaux

1 Introduction

Le registre de langue dans lequel se situe un texte (à l'oral comme à l'écrit) apparaît comme un trait saillant. Il renvoie au contexte d'énonciation dans lequel il est —ou a été —produit (et qui comprend notamment la relation du locuteur avec ses interlocuteurs). Parmi les manifestations possibles de ce phénomène sociolinguistique, le partitionnement en registres tels que familier, courant et soutenu est probablement le plus répandu. Nous présentons ici un guide d'annotation de textes en français selon la tripartition familier, courant et soutenu. Une des contributions sur le plan linguistique est d'y inclure certains éléments spécifiques aux discours numériques (en particulier les tweets) comme les hashtags et les émoticônes. Plus largement il prend part au projet ANR TREMoLo [2] dont « les objectifs sont de progresser dans l'étude des registres de langue et de développer des méthodes automatiques de transformation de textes d'un registre vers un autre. » [3].

Le registre de langue dans lequel se situe un texte (à l'oral comme à l'écrit) apparaît comme un trait saillant. Il renvoie au contexte d'énonciation dans lequel il est —ou a été —produit (et qui comprend notamment la relation du locuteur avec ses interlocuteurs). Parmi les manifestations possibles de ce phénomène sociolinguistique, le partitionnement en registres tels que familier, courant et soutenu est probablement le plus répandu. Si des corpus comme GYAFC (Rao et Tetreault 2018) — où ce type de variations est appelé « niveau de formalité » — ont récemment popularisé le domaine, celui-ci est encore globalement peu étudié en traitement automatique des langues (TAL), et particulièrement en dehors de l'anglais. Par ailleurs, bien que de nouveaux types de textes aient émergé depuis les deux dernières décennies — tels que les tweets, et plus généralement ceux que l'on range sous le terme *CMO* —, les travaux sur les registres de langue traitent surtout des types plus classiques de textes dont les caractéristiques sont plus ou moins connues de la littérature linguistique (on associera ainsi généralement par exemple les insultes au registre familier et la diversité de connecteurs logiques à du registre soutenu). Dès lors, les analyses de corpus CMO en termes de registres de langue constituent un défi tant pour la linguistique descriptive que pour les différentes applications en TAL. Pour répondre à ces enjeux, ce guide d'annotation propose un protocole d'annotation de CMO en proportions de registres de langue.

La constitution d'un corpus de textes écrits représentatif de l'usage réel des registres de langue présente deux difficultés majeures : tout d'abord le lien biunivoque fort entre certains registres et certains types de textes (par exemple le soutenu associé à des romans de la littérature classique, le familier aux forums de discussion, et le courant à des dépêches journalistiques) ; ensuite l'association quasi immédiate de la modalité orale avec le registre familier d'une part, et de la modalité écrite avec les registres courant ou soutenu d'autre part (Gadet 2000 ; Rebourcet 2008). Pour répondre à ces biais, nous avons choisi de construire notre corpus à partir d'un seul type de textes issu des CMO définis comme

2. https://anr.fr/Projet-ANR-16-CE23-0019
3. https://tremolo.irisa.fr/fr/

« *toute communication humaine qui se produit à travers l'utilisation de deux ou plusieurs dispositifs électroniques* » (McQuail 2010). Un des intérêts des CMO sur le plan linguistique réside dans le fait qu'ils contribuent à créer un « *parlécrit* » (Jacques 1999) par le caractère instantané des échanges qu'ils matérialisent ; l'intérêt des tweets en particulier parmi les CMO est leur limite à 280 caractères, imposée par Twitter, ce qui homogénéise la taille des textes produits et analysés.

2 Corpus

2.1 Présentation générale de Twitter et des tweets

Twitter est un réseau social en ligne créé en 2006 par Jack Dorsey, Evan Williams, Biz Stone et Noah Glass. Il est géré par la société Twitter Inc. Il comptabilise au dernier trimestre 2019 « 145 millions d'utilisateurs actifs quotidiens et 330 millions d'utilisateurs actifs mensuels » [4] ce qui fait de lui une plateforme emblématique du « micro-blogging » : « blogue constitué de minimessages diffusés en temps réel, qui contiennent souvent des mots-dièse et dont l'enchaînement forme des fils de discussion. » [5]. Ces « minimessages » ont une taille limitée à 280 caractères dans lesquels sont inclus les espaces. (Domenget 2013) précise que :

> « [...] Twitter est un dispositif asymétrique (venant d'une non-réciprocité possible dans les abonnements), dont la logique d'usage principale consiste à partager des contenus autour de centres d'intérêt ; échanges se réalisant entre pairs. »

Twitter est donc un réseau social en ligne qui permet d'échanger en temps réel des « minimessages » dont la taille est limitée sans nécessairement impliqué une relation de réciprocité entre les utilisateurs : par exemple un utilisateur peut être abonné à un autre utilisateur sans que ce dernier ne le soit en retour, ou bien un utilisateur peut répondre à un tweet sans avoir de réponse... De plus, comme le met en exergue (Domenget 2013) twitter permet de créer des communautés d'utilisateurs qui se regroupent selon des intérêts, des valeurs ou bien des opinions partagées.

Le message textuel produit sur twitter par les utilisateurs sont appelés des « tweets ». Un tweet est « un énoncé plurisémiotique complexe, limité à, 280 signes, fortement contextualisé et non modifiable, produit nativement en ligne sur la plateforme de microblogging Twitter. Le tweet apparaît dans le fil du twitter (ou twittos) et dans la timeline (TL) de ses abonnés. Depuis la naissance de la plateforme en 2006, ses formes ont considérablement évolué, passant d'un format simple (un énoncé inscrit dans une fenêtre) à des formats et des combinaisons variées (tweet avec des photos (figure 1), vidéos (figure 2) ou gif, avec partage, autoretweet, thread, ect.) » (Paveau 2017).

4. https://www.agencedesmediassociaux.com/twitter-chiffres-2020/
5. http://gdt.oqlf.gouv.qc.ca/ficheOqlf.aspx?Id_Fiche=26540439

FIGURE 1 – Exemple de tweet qui intègre une image

(PAVEAU 2017) propose de décrire un tweet dans sa forme conventionnelle comme composé de :

1. photo de profil de l'abonnée ;
2. nom de l'abonné ;
3. pseudo de l'abonné,
4. date du tweet, relative ou absolue ;
5. texte du tweet inscrit dans la fenêtre dédiée (280 signes espaces compris) ;
6. liste des opérations possibles signalées par des icônes sous le texte (anciennement assorties de mots-consignes) : répondre, retweeter, aimer, activité des tweets ;
7. bouton-chevron signalant un menu déroulant avec différentes fonctions telles que copier le lien du tweet, intégrer le tweet, bloquer...

Dans le cadre de notre travail nous utiliserons une « forme logocentrée » (exemples (1) et (2)), c'est à dire « une présentation du tweet qui ne retient que les éléments de contenu langagier, au détriment de l'ensemble des éléments discursifs et technodiscursifs mentionnés plus haut » (PAVEAU 2017). De fait notre corpus tend à représenter la variation des registres de langue circonscrite au contenu langagier.

(1) Mon corps secrète encore de l'endorphine avec cette victoire du King @X #ItalianGP #F1 #MonzaGP

(2) Mdr je fait les hype night chill et je me fait tuer par @X et il danse mdr jsp si cets le vrai cets quoi sont epic svp

2.2 Présentation des éléments linguistiques spécifiques aux tweets

Bien que nous renoncions à sa « forme écologique » certains éléments linguistiques restent spécifiques des tweets ou bien de l'écriture numérique car

FIGURE 2 – Exemple de tweet qui intègre une vidéo

ils résultent d'une production uniquement possible avec la plateforme Twitter. (PAVEAU 2013) propose une typologie des formes langagières qu'exploite Twitter :

1. des formes linéaires sans caractéristiques technolangagières autres que l'inscription sur support informatique ;
2. des émoticones ;
3. des liens (URL) qui permettent d'accéder à des sites ;
4. des technomots comme le hashtag (précédé du croisillon #) qui permet l'organisation de l'information par la mise en réseau de plusieurs messages et le pseudo (précédé de @) qui renvoie au compte du twitteur ;

Dans les sections ci dessous nous exposons la terminologie ainsi que les définitions que nous avons adopté pour ces différentes "formes langagières" (ibid.).

2.2.1 Pictogramme

(BECCUCCI 2018) définit un *émoticône* comme « un signe graphique 'ressemblant' à une émotion. » *l'emoji* quant à lui consiste en des symboles listé d'une banque de données. *Emoji* est un mot d'origine japonaise qui se traduit par « caractère-image ». Désormais, nous utiliserons le terme « *pictogramme* » afin de désigner à la fois les émoticônes et les émojis. Les pictogrammes peuvent être insérés à différentes positions dans la chaîne syntaxique (MAGUÉ, ROSSI-GENSANE et HALTÉ 2020) :

1. antéposé : au début de la chaîne syntaxique (exemple (3)),
2. interphrastique : entre deux phrases « syntaxiques » au sein d'un même tour de parole (exemple (4)),

3. postposé : en fin de chaîne syntaxique (exemple (5)).

(3) ➡️ Sergi Roberto et Jordi Alba compléteraient un milieu de terrain très encombré et à l'avant seraient Messi et Luis Suárez

(4) RT @X : Nouvelle bio : « fan du Puy du Fou » 🤪 merci le CM de l'@X

(5) + @X ma vie (je le rajoute après il pleure sinon) ❤️ ❤️

2.2.2 Technomots

Mot-dièse Nous utilisons la traduction du terme anglophone « hashtag » : « mot-dièse ». Plusieurs définitions lui sont attribuées, parmi ces dernières nous retenons celle du (JORF numéro 19 du 23 janvier 2013) qui définit un mot-dièse comme :

> « une suite signifiante de caractères sans espace commençant par le signe # (dièse), qui signale un sujet d'intérêt et est insérée dans un message par son rédacteur afin d'en faciliter le repérage. »

Il « est constitué du signe "#" suivi de caractères alphanumériques contigus, formant une ou plusieurs unité de sens. » Le mot-dièse sert à « "indexer" ses publications, en insérant simplement un dièse devant un mot ou un groupe de mots de son choix, pour que le réseau Twitter (Facebook ou encore Google+) considère que sa publication n'est pas isolée, mais faisant partie d'une thématique ou d'un sujet sur lequel parlent d'autres personnes. [...] Il est important de noter que l'usage du mot-dièse n'est ni encadré, ni normé. Il fait partie du processus même de rédaction. » (Jackiewicz et Vidak 2014).

Mention d'URLs La nature « hybride » des tweets est notamment caractérisée par la présence d'URLs qui sont généralement insérés en fin de tweet. L'abréviation « URLs » vient de l'anglais « Uniform Resource Locators » défini par (Berners-Lee, Masinter, McCahill et al. 1994) comme :

> « [...] a compact string representation for a resource available via the Internet. These strings are called "Uniform Resource Locators" (URLs) »

Les URLs ont une forme formatée par Twitter et sont plus courts que les URLs classiques.

(6) Moi à 10 ans qui cherche sur YouTube les vidéos qui démontrent que Michael Jackson n'est pas mort. **https ://t.co/wi6rxODMcD**

(7) Vazy envoie le pognon #Cdiscount2emeDemarque **https ://t.co/P-UQeZXu2CB**

Mention de l'identifiant de l'utilisateur Lorsqu'un twitteur veut interpeller un autre twitteur il peut mentionner son identifiant en le faisant commencer par un « @ ».

(8) **@X** Mdrrr même moi je veut qu'on me rajoute tqt dès que je fait un groupe je t'ajoute!!!!!

Abréviation spécifique de Twitter Différentes abréviations peuvent être mentionnées dans les tweets. Ces abréviations désignent des fonctionnalités proposées par Twitter :

1. ReTweeter : souvent utilisé lorsqu'un twitteur retweet, c'est à dire publie un tweet d'un autre utilisateur. Sa forme abrégée, « RT », est mentionnée en début de tweet (exemple (9)).
2. Trending Topics : souvent mentionné sous sa forme abrégée, « TT », met en avant un sujet tendance à un moment donné sur Twitter. Il peut également être mentionné dans un mot-dièse « #TT » (exemples (10) et (11)).
3. LT : Live Tweet souvent mentionné sous sa forme abrégée, « LT », est utilisé lorsque le twitteur crée et publie un tweet d'un évènement auquel il assiste. Il peut également être mentionner dans un mot-dièse « #LT » (exemples (12) et (13)).

(9) **RT** @X : PPPPPTDDRRR on dirait un goss jvais canner https ://t.co/THbhTQROYC

(10) Comment ça on est en **TT** la veille à 9h du matin ? Ça m'avais manqué #10YearsOneDirection

(11) Je profite que "Saddam Hussein" soit en #**TT** pour rendre hommage au refus de Jacques Chirac de partir en guerre dans une folle coalition en 2002. Qui qu'il soit par ailleurs, il avait senti et compris les innombrables conséquences et cruautés que ça aurait engendré. #GuerreEtPaix

(12) Charlie Hebdo / début du procès des #AttentatsJanvier2015 à #CharlieHebdo #Montrouge #HyperCacher. Les accusés arrivent, sous escorte armée. De très nombreux policiers et gendarmes., ⏩ pour la radio @X, @X et @X couvrent l'événement en **LT** https ://t.co/JBQn3Ao1hd

(13) #**LT** EN DIRECT 🔴 #Webinar #RETAIL ! "L'expérimentation c'est quelque chose de fondamental ! La culture de l'expérimentation, pour une entreprise, c'est déjà d'être explicite sur le droit à l'échec d'une initiative" Guillaume Patin d'@X nous partage sa vision. https ://t.co/NyssM00o33

2.3 Constitution du corpus global

Une extraction automatique de tweets est faite. Cette extraction extrait sans *a priori* linguistique quant aux registres de langue. Aussi, elle ne part pas d'une théorie linguistique mais suppose qu'en relevant les tweets, qui mentionnent les termes les plus utilisés à un instant t dans une zone géographique donnée (section *"Trends"* dont l'exemple est donné figure 3), la diversité des productions sera représentative des différentes fonctions du langage.

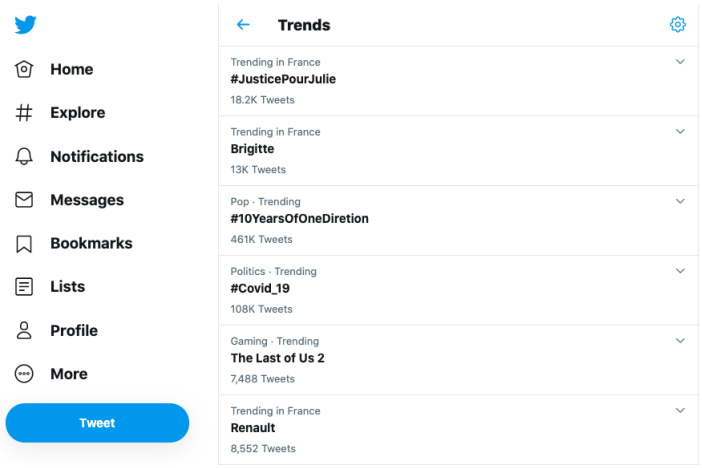

FIGURE 3 – Exemple de tendances dans la section *"Trends"* sur twitter

Nous utilisons l'API proposée par Twitter [6]. Les outils permettent, à partir d'un identifiant de localisation, de récupérer automatiquement les 50 tendances associées à Paris. Pour chacune de ses tendances une extraction, qui recherche tous les tweets qui mentionnent ces tendances, est faite : la recherche a une limite maximale de 10 000 résultats. L'extraction des mots clefs les plus utilisés est spécifiée par :

1. la date,
2. la zone géographique (*Where On Earth IDentifier - WOEID*).

Afin de couvrir le plus d'usages différents et donc de sujets différents 10 extractions à 10 dates différentes sont effectuées. Ces extractions couvrent un durée totale d'un mois. Le lieu ne change pas et reste Paris dont le WOEID est 615702.

La table 1 présente le nombre de tweets recueillis pour toutes les extractions faites.

2.4 Pré-traitement du corpus

Le texte de chaque tweet est isolé. Pour chacun d'entre eux sont vérifiées :

1. leur langue : nous excluons les autres langues que le français,

6. `https://developer.twitter.com/en/docs`

extract	nbr tweets
0	233 080
1	205 474
2	27 562
3	220 769
4	162 916
5	109 689
6	24 917
7	43 310
8	69 997
9	173 189
total	1 270 903

TABLE 1 – Synthèse des extractions automatiques de tweets

2. leur intégrité : nous excluons les tweets tronqués.

Les tweets non français ont été repérés grâce à un module python [7] qui, pour un texte donné, prédit une langue à une certaine probabilité P. Si $P \leq 0.90$ pour le français, alors le texte est conservé dans le corpus. La valeurs de P est fixée afin de garder des textes avec la présence de quelques termes non français intéressants tels que « *lol* », « *dead* », « *stan* »... (GADET 2003) relève en effet que le registre familier emprunte des termes lexicaux aux langues non françaises. Ces tweets hétérogènes sont pertinents et seront conservés dans le corpus (exemple (14)).

(14) @X Alors explique à quoi sa sert de dénigrer le corps de tootatis même si c'est fesse son **fake** .

Quant aux tweets tronqués, ils ont été repérés grâce à un signe de ponctuation spécifique de Twitter : trois points de suspension resserrés différents des « ... » classiques. Une règle symbolique a écarté les tweets qui se terminaient par ce signe de ponctuation particulier.

Finalement, après l'exclusion des tweets non français ou tronqués, le corpus compte 228 505 tweets (6 201 339 mots).

3 Principes généraux de l'annotation

3.1 Catégories employées

Nous utilisons le terme « registre » défini *comme une variation des formes linguistiques, à différents niveaux d'analyse de la langue, par rapport à un standard donné*. Ce standard correspond à l'intersection d'une « norme objective » (les règles grammaticales) et d'une « norme subjective » (les règles d'usage, c'est-à-dire « *la contrainte collective qui donne lieu à des jugements de valeurs*

7. langdetect

constitutifs de l'attitude courante quelle que soit la façon de parler des locuteurs ») (Gadet 2007).

Notre travail partitionne l'espace linguistique en trois registres principaux : familier, courant et soutenu. Bien que nous admettions sans difficulté qu'il existe un continuum entre ces trois registres, cette partition découle du besoin d'un découpage en valeurs discrètes pour un traitement automatique. Les catégories employées pour l'étiquetage des textes sont donc les mêmes. La catégorie « Poubelle » est ajoutée pour les tweets mal encodés ou incompréhensibles.

Ainsi, suivant notre définition d'un registre, un texte est considéré comme :

1. « Familier » lorsque la norme objective n'est pas suivie,
2. « Courant » lorsqu'il se conforme partiellement aux normes objectives et subjectives,
3. « Soutenu » lorsqu'il se conforme complètement aux normes objectives et subjectives,
4. « Poubelle » lorsqu'il est de mauvaise qualité.

3.2 Segmentation du corpus

Chaque contenu textuel complet des tweets extraits constitue l'entité à annoter. Les entités à annoter peuvent donc être :

1. une phrase avec ou sans ponctuation (exemples (15) et (16)),
2. un ensemble de phrases avec ou sans ponctuation (exemples (17) et (18)).

(15) #MonPireDate il a pas arrêté de me dire que son ex était tout le contraire de moi...

(16) #MonPireDate tu peux pas avoir de pire date quand tu n'as JAMAIS eu de Date https ://t.co/ArzLDO6tE4

(17) @X @X Oui thauvin une vraie recrue. Mais faut vite gommer ce genre de match ou un relégable te met la pression comme ça. En L1 faut avoir plus de caractère. Et je dis ça vraiment pour être objectif. Pas le supp de base parisien. On est pas rival.

(18) @X @X @X @X Si ils le font pas c'est psk a part de la promo ya pas d'intérêt pas besoin de collab pour pesé dans le game damso sans aucune collab avec une grosse industrie il est dans le top 5 des rappeur les plus streamé en france pnl sont forts mais faut pas généralisé ya bien mieux ajd

Pour un texte donné l'annotateur doit attribuer de un registre à quatre registres de langue.

4 Annotation

4.1 Protocole d'annotation

Pour un texte donné l'annotateur doit classer les registres selon leurs prédominances dans le texte. Ce classement s'illustre par l'attribution d'un « rang », noté r, pour chaque registre présent : 1, 2, 3 ou 4. Si un seul registre est présent alors il n'y aura qu'une valeurs numérique, « 1 », qui correspond au registre attribué au texte (figure 4).

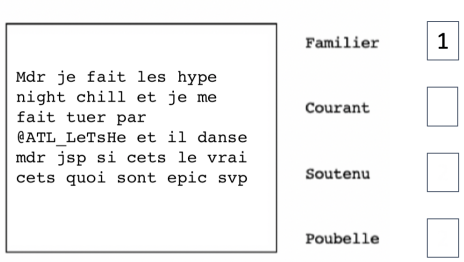

FIGURE 4 – Exemple d'un texte étiqueté en registres de langue

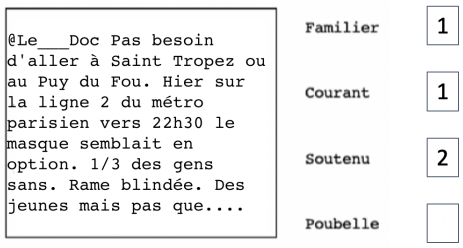

FIGURE 5 – Premier exemple d'un texte étiqueté en registres de langue avec deux registres

Si dans un même texte plusieurs registres sont présents :

1. Leurs présences dans le texte sont égales, dans ce cas là nous attribuons la même place dans le classement.
   (a) Dans l'exemple de la figure 5 les registres dominants sont le familier et le courant, le soutenu et moins présent que ces deux premiers registres.
   (b) Dans l'exemple de la figure 6 le registre dominant est le familier, le courant et le soutenu sont moins présents que le familier mais aucun ne domine l'autre : leurs présences sont égales.

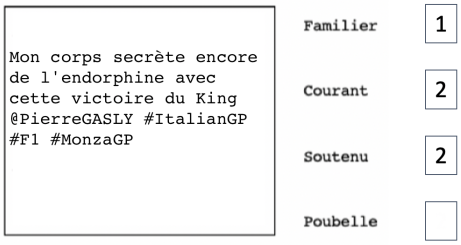

FIGURE 6 – Second exemple d'un texte étiqueté en registres de langue

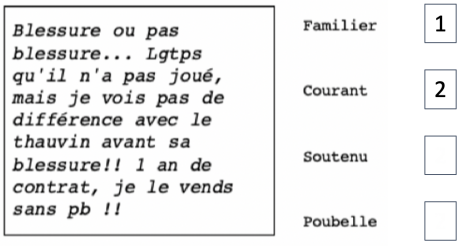

FIGURE 7 – Second exemple d'un texte étiqueté en registres de langue

2. Leurs présence dans le textes sont inégales, dans ce cas là nous attribuons des places différentes dans le classement (voir figure 7 où le registre familier est plus présent que le registre courant dans le texte)

L'annotateur doit justifier l'étiquette attribuée avec au moins la présence d'un descripteur linguistique. Pour ce faire l'annotateur peut sélectionner dans une liste un descripteur issu d'une étude linguistique préliminaire faite sur le corpus. Cette étude reprend des descripteurs déjà identifiés dans la littérature scientifique sur les registres de langue (MEKKI et al. 2018) et y rajoute des descripteurs issus de l'analyse linguistique du corpus d'écrits numériques (tweets) constitué pour ce projet. La liste complète des descripteurs linguistiques est détaillée section 5.

4.2 Transformation des annotations en proportions de registres

l'annotateur doit ordonner les registres en fonction de leur prédominance dans un texte en leur attribuant un qui doit être justifié par la présence d'au moins un descripteur de la liste issue de l'analyse linguistique préliminaire.

Chaque rang est ensuite transformé en proportion de registre. Soit :

— R un ensemble de registres ayant obtenu un rang et $Card(R)$ son nombre d'éléments,

— $r_i \in R$ un registre de R ayant obtenu le rang i,

— $inv(i)$ le rang inversé du rang i défini par $inv(i) = Card(R) - i + 1$

— srg la somme des rangs définie par $srg = \sum_{i=1}^{Card(R)} i$

La proportion du registre r_i est alors définie par $Prop_{r_i} = \frac{inv(i)}{srg}$

Ainsi, pour un texte annoté comme ceci, r_1=familier, r_2=soutenu et r_3=courant, on obtient :

— $R = \{r_1, r_2, r_3\}$ et $Card(R) = 3$,

— $inv(1) = 3$, $inv(2) = 2$, $inv(3) = 1$

— $srg = 6$

Les proportions en registre sur cet exemple sont donc familier 50% ($\frac{3}{6}$), soutenu 33% ($\frac{2}{6}$) et courant 17% ($\frac{1}{6}$).

5 Liste des descripteurs linguistiques

Les descripteurs sont regroupés pas niveau d'analyse de la langue (détails table 2) : niveau syntaxique, lexico-syntaxique, discursif, lexical, morphologique, phonologique.

Niveau d'analyse	**Nombre de descripteurs**
Syntaxique	15
Lexico-syntaxe	5
Discursif	6
Lexical	9
Morphologique	9
Phonétique	3
Total	**47**

TABLE 2 – Détails quantitatifs des descripteurs par niveaux d'analyse de la langue

Chaque exemple donné pour illustrer les descripteurs linguistiques constitue le tweet intégral dans lequel est présent le descripteur. Les éléments en gras (textuels) ou bien soulignés (les pictogrammes) sont les objets précis sur lesquels portent les descripteurs.

5.1 Niveau syntaxique

Mention de l'identifiant de l'utilisateur dans un syntagme L'identifiant de l'utilisateur, par exemple « @X » est utilisé comme sujet d'un verbe conjugué. Autrement dit l'identifiant de l'utilisateur est intégré syntaxiquement.

(19) Les gars dites vous bien que **@X veut** faire du trading avec des prêts sans OA. Campos en tremble déjà... 🤔 #MercatOM #VenteOM

Mot-dièse intégré syntaxiquement Lorsqu'un mot-dièse [8] est intégré syntaxiquement, c'est à dire qu'il assume une fonction syntaxique au sein de la phrase.

(20) Essor de la #**télémédecine**, attractivité des métiers et établissements, enseignements et perspectives de la crise #**sanitaire**, bilan du #**SegurDeLaSante** ... Autant de sujets présents aux conférences de #**SANTEXPO** 2020. RDV du 7-9 oct à Paris ! 👉 le programme 👉 https ://t.co/ofnMjmDtLU https ://t.co/sJf5qR1z2I

Mot-dièse indépendant syntaxiquement Lorsqu'un mot-dièse [9] n'est pas intégré syntaxiquement à la phrase, c'est à dire qu'il n'assume pas de fonction syntaxique au sein de la phrase.

(21) RT @X : Caleb Ewan est tombé dans cette chute ! Un des favoris du jour. #**TDF2020**

Mot-dièses sans rapport syntaxique entre eux Lorsque plusieurs mot-dièses [10] sont écrits côte à côte sans avoir de rapport syntaxique entre eux. Cela crée un effet de juxtaposition c'est à dire qu'il n'y a pas de rapport syntaxique mais que la liaison syntaxique se fait « par simple rapprochement excluant la coordination et la subordination » [11].

(22) Marche arrière toute direction la Terre ! 🧐 #**espace** #**astronomie** https ://t.co/YWYL88OSIf

Expression modalisatrice Une expression modalisatrice est une expression qui fait parti de l' « Ensemble des faits linguistiques (mode, forme assertive, interrogative ou injonctive de la phrase, adverbes ou auxiliaires modaux) traduisant l'attitude du sujet parlant par rapport à ce qu'il énonce » [12]. Par exemple les expressions telles que « *genre* » (exemple (23)), « *franchement* » (exemple (24)) ou bien « *imagine* » (exemple (25))...

(23) @X @X **Genre** toi tu préfères Kaguya Sama à SBR ? ?

(24) **Franchement** ça fait plaisir un joueur français sur Fifa 21 vraiment https ://t.co/nM2N1DWOyj

8. défini sous section 2.2.2
9. défini sous section 2.2.2
10. défini sous section 2.2.2
11. `https://www.cnrtl.fr/definition/juxtaposition`
12. `https://www.cnrtl.fr/definition/modalit%C3%A9`

(25) Mdr **imaginez** vous êtes le p'tit frère de Tootatis. Eh la défaite miskine.

Absence de ponctuation classique Lorsqu'il n'y a aucun signe de ponctuation classique (« . », « *?* », « *!* », « ... », « , », « *;* ») présent dans le texte.

(26) C'est une hécatombe aujourd'hui sur la route du tour Thierry

Procédé de reprise de parole Un procédé de reprise de parole d'autrui, c'est lorsqu'un premier locuteur rapporte les paroles exactes d'un second locuteur. Généralement les paroles sont rapportées entre guillemets et introduites par un verbe de parole : « *"J'adore les chats" dit-elle* » (exemple (27)). Elles peuvent également être introduites après le signe de ponctuation « : », par exemple : « *Marie : "j'adore les chats"* » (exemple (28)). La reprise de parole peut également se faire en introduisant une subordonnée complétive après un verbe de parole « *Elle m'a dit qu'elle adorait les chats.* » (exemple (29)).

(27) 🚨 BREAKING : Kylian #Mbappé **confirme** qu'il reste au #PSG 🔴🔵🔥 **"Je suis là. Je suis dans le projet pour une quatrième année. Je serai là quoi qu'il arrive et je vais essayer de ramener des trophées avec l'équipe en donnant le meilleur de moi-même."**

(28) Hamilton **:"Enormes félicitations à Pierre"** https ://t.co/-rVjPzWKNYm

(29) @X Tootatis elle même assume que ce n'est pas son vrai corp sur les photos hein, **elle dit que son corp es 100% photoshop**

Construction syntaxique : « *vu* » suivi d'un groupe nominal La forme au participe passé du verbe « *voir* » suivi d'un groupe nominal.

(30) @X @X oui tout à fait, il semble qu'elle a des envies pressantes. Oui au fond à droite...c'est là où il y écrit : For Ladies... Mais **vu l'urgence**, Rachida Dati peut y aller quand même !

(31) @X @X Baba un grand merci a toutes l équipe @X qui a encore mis la barre très haute cette saison, sa promet **vu le déroulement de la première**, le meilleur reste à venir hâte de venir en plateau ça fait un bail gros bisous à toi @X ainsi qu a toutes l équipe

Répétition contiguë d'items Lorsqu'un item identique se répète au moins deux fois à la suite. Par exemple des signes de ponctuations qui se répètent (exemple (32)), ou bien des pictogrammes (définis dans la sous section 2.2.1) (dans l'exemple (33) une séquence de répétitions contiguës de pictogrammes), des adverbes d'intensité (exemple (34))...

(32) gaelle gd à la crémaillère de lena **? ? ? ? ? ?** wsh

(33) @X Si on rajoute les gens qui crient pendant qu'ils parlent (non j'ai pas été voir BTS en concert mais j'ai déjà vue des concerts) 😭 😭 😭 😭

(34) Comment vous dire que non seulement il a l'air meilleur que le 1er mais surtout il à l'air GÉNIALE SUR MA VIE J'AI TROP HÂTE **TROP TROP** #AfterWeCollidedMovie

Élément doublé Lorsqu'un item est doublé par un second item qui crée une redondance sémantique. Par exemple, le rajout de « *à lui* » « *à elle* » après les pronoms « *son* » et « *sa* » [13] (exemple (35), le syntagme lexical redoublé en position sujet ou bien lorsqu'un pictogramme a une fonction référentielle et vient après un terme qui fait référence au même référent extralinguistique (exemple (36)).

(35) LE BIRTHDAY DE L'HOMME que **son** jour **à lui** ma big life j'l'aime #JungkookDay https ://t.co/oHEeJqMwJm

(36) Westbrook a quasi 32, il a autant d'énergie qu'à 25. Je ne sais pas ce qu'il bouffe. Du **feu** 🔥 peut-être. #nbaextra

Mise en commun du sujet pour plusieurs verbes successifs Lorsque plusieurs verbes qui se suivent partagent le même sujet [14].

(37) Je viens de voir la performance de BTS, ils étaient aux US ou alors avec le Corona **ils sont** rester en Corée et **on fais** une performance en simultané ? L'avantage ? On vois la chorégraphie et pas le publique toutes les 30sc. Et qu'ils sont beaux ! Olalala #VMAs2020BTS #VMAs

(38) #GGRMC une chose qui m'a toujours "étonnée" avec Dieudonné c'est savoir ce qui s'est passé reellement entre lui et Elie Seimoun ? **ils ont travaillé** et **étaient** "amis" durant des années et d'un seul coup il serait devenu "antisémite" ? ! plutôt bizarre ! vous connaissez la raison ? 🤔 🇫🇷

13. Gadet 2003
14. Bilger et Cappeau 2004

Non inversion sujet/verbe dans une phrase interrogative Lorsque le sujet précède le verbe dans une phrase interrogative alors que ce dernier devrait être placé après le verbe et non devant [15] : « *Comment **tu vas ?*** » vs. « *Comment **vas tu ?*** ».

(39) @X16 @X **Tu connais** pas Alain Soral mdr **? T'es née** en 2009 **?**

Absence de l'accord au pluriel du syntagme « *c'est* » devant un syntagme pluriel Lorsque l'expression « *c'est* » précède un groupe de mots au pluriel [16].

(40) Chez @X nous savons que l'avenir **c'est les jeunes**. C'est pour ça que nous avons recruté 10% d'alternants. Agir aujourd'hui pour demain #rentreescolaire2020 #alternance @X https ://t.co/RcISUyc4Id

Absence d'un item attendu Lorsqu'un item devrait être présent mais ne l'est pas et entraîne une construction « non standarde ». Par exemple l'absence du pronom « *il* » dans une construction impersonnelle (exemple (41)), l'absence d'un verbe (exemple (42)), l'absence d'un subordonnant dans proposition relative (exemple (43)), l'absence du « *ne* » dans une construction négative (exemple (44))...

(41) au boit d'un moment ∅ faut arrêter de se victimiser

(42) ∅ *Jamais **vu** un plus gros # de fils de pute #BoycottNetflix*

(43) @X @X Je pense ∅ elle a peur des critiques étant donné qu'elle est en maillot de bain

(44) @X Ça dépend. Souvent je joue avec un autre astro donc je fais l'inverse de lui. **J∅sais pas** si c'est bien

Décumule de comparatif synthétique Lorsque un comparatif synthétique (tels que « *meilleur* », « *mauvais* », « *mieux* »...) se divise en deux éléments distincts ce qui créée une redondance sémantique [17].

(45) Adama Traoré il peut même pas reposer en paix , tous les jours ça lui invente de nouveaux délits qui sont tous les jours un peu **plus pire** , j'ai honte de la France

(46) @X Québec nous donne des leçons de discipline alors qu'il est le **plus mauvais** élève du pays ! et si le nombre de morts était dû à sa gestion désastreuse ? s'inspirer de la France où ils ont commis toutes les erreurs plutôt que choisir les bons élèves comme la Suède... #Arrudaenprison

15. Gadet 2003
16. Favart 2009
17. Gadet 1997

5.2 Niveau lexico-syntaxique

« *cette* » ou « *cet* » suivi d'un groupe nominal puis « *de* » puis d'un groupe nominal Lorsque le patron $(cette|cet) + GN + de + GN$ est présent dans le texte.

(47) j'adore Léna situations mais **cette histoire de stage**... bref no comment

(48) Merci à toi Fatima , une voix dissonante , à la Kazib , bonne continuation et on t'embrasse , t'as été un rayon de soleil pour beaucoup . Longue vie à toi ! Par contre , virez **cette salope de fromage Blanc** . C'est d'la merde lui ! #GGRMC

(49) C'est quoi **cette putain de mode de fils** de putains de faire des parallèles avec la shoah ?

« *juste* » suivi d'un adjectif ou bien d'un adverbe Lorsque le patron $juste + (ADJ|ADV)$ est présent dans le texte.

(50) go silhouette serieux, l'opening est **juste magnifique** https ://t.co/kieYeOCZVv

(51) **Juste incroyable** 😍 🔥 🔥 Hâte de le porter sur mes épaules pour FIFA 21 💪 🏆 https ://t.co/4iyoQRmqsq

Construction du futur avec le verbe « *aller* » Le futur est exprimé grâce au verbe « *aller* » conjugué suivi d'un verbe à l'infinitif.

(52) Quand je **vais voir** mon emplois du temps et que je vais faire du 8h-17h tout les jours 🤯 #RentreeScolaire https ://t.co/lAxCsDioQE

(53) Bon beh nous **allons** légèrement **souffrir** les prochain jours à Toulouse #canicule #Chaleur #chaud https ://t.co/bqCYEVSpGE

« *est-ce que* » utilisé pour formuler les phrases interrogatives Lorsque l'expression « *est-ce que* » est présent dans le texte pour former les phrases interrogatives [18].

(54) EH OH wesh **est ce que** quelqu'un sait ce qui se passe entre Ben et Léna/Suli/Sparkdise/Bilal ? ?

18. Ilmola 2012

Locution adverbiale semi-figée Une locution est un « groupe de mots constituant un syntagme figé » [19], une locution adverbiale semi-figée c'est lorsque la locution comporte un adverbe est être considéré comme de véritables mots composés caractérisés par un sens unique : « *avoir lieu* » équivaut à « *arriver* », « *se produire* » [20].

(55) @X Des snipers seront-ils positionnés sur les toits pour **faire entendre raison** aux plus récalcitrants ? #MasqueObligatoire #masquesObligatoires #coronavirus

(56) @X @X 👏 👏 👍 👍 👍 Plenel n'est pas un journaliste C'est un INQUISITEUR de la pire des espèces Un Journaliste INFORME RELATE avec Discernement Objectivité Honnêteté Intellectuelle En aucun cas il doit suggérer, **prendre parti** ou orienter J'ai appris cela en 1968 École de Photos

5.3 Niveau Discursif

Mot-dièse utilisé comme commentaire phrastique Lorsqu'un mot-dièse [21] est employé pour faire un commentaire sur la phrase qui le précède.

(57) Pas de 😁 , serrage de ✋ ... Apparemment, le respect des gestes barrières, ce n'est que pour le bas peuple. Les hierarques en sont dispensés. #rentreescolaire2020 #COVID19france #rebond #**foutagedegueule** https ://t.co/Ige8bKITJ5

(58) L'astuce du Puy du Fou pour rester dans les clous #**genie** https ://t.co/Y6N4Yzv3gd

Texte structuré par la ponctuation Lorsque la ponctuation ou bien un pictogramme [22] permet de mettre en avant une information en structurant le texte.

(59) ➡️ Vous souhaitez préparer un BTS Communication en alternance ? Notre partenaire dans le secteur automobile propose un poste de chargé.e de communication, adressez votre candidature au plus vite ! 👉 https ://t.co/EOGKLyeSar #apprentissage #formation #alternance #recrutement
https ://t.co/SbEdzqX3F1

(60) #LeSaviezVous ? Vous pouvez bénéficier d'aides financières si vous faites isoler votre logement par un professionnel certifié #RGE. Mais pas si vous l'isolez vous-mêmes ! IBC, leader gardois en #isolation, s'occupe de vous ! Plus d'infos 👉 0 805 691 777 / https ://t.co/Uf0yv63Tla https ://t.co/MNtrlKvwet

19. `https://www.cnrtl.fr/definition/Locution`
20. Brunot 1922,Ilmola 2012
21. défini sous section 2.2.2
22. défini sous section 2.2.1

Le présent comme unique temps utilisé Lorsque seul le présent de l'indicatif est utilisé comme temps verbal dans le texte.

(61) Après d'un autre côté ça **peut** être drôle si tu **veux** qu'on t'**appelle** n'**importe** comment du style : Read the text please Lionel Jospin https ://t.co/iV39wiGPtM

Diversité des temps verbaux Lorsqu'au moins deux temps verbaux différents sont utilisé dans un même texte.

(62) Jospin **fut** un bon PM dans l'ensemble, sans tchatche, ce qui lui **a coûté** d'ailleurs l'élection ! https ://t.co/OvLPY3IbZf

(63) @X Cette position **a coûté** cher à Jospin... Dire que l'insécurité **est** un fantasme n'**est** pas tenable.

Diversité des connecteurs Un connecteur « est un opérateur susceptible de faire de deux phrases de base une seule phrase transformée » [23]. Le descripteur est considéré comme présent lorsqu'au moins deux connecteurs différents sont écrits dans le texte.

(64) Le double pivot ne fonctionnera jamais **puisque** ni Neymar ni Di Maria n'auront de culture défensive Il faut un retour en 433 accompagné d'un box to box pour régler ce soucis, + que n'importe quel latéral ou défenseur qui règleront **certes** d'autres soucis, **mais** moins prioritaires

(65) @X dans une fratrie lorsque qu'un individu est homo il ya ~~3~~0% de chance pour qu'un second le soit. **cependant** ces 30% n'augmentent pas dans le cas de vrais jumeaux : cela implique que l'homosexualité n'est pas 100% inscrite dans les gènes **mais** decoule bien d'un processus biologique

Enchaînement de plusieurs phrases avec des ponctuations classiques Lorsque plusieurs phrases s'enchaînent à la suite. Les phrases doivent toutes commencer toutes par une majuscule et se terminer par une ponctuation de fin de phrase (« . », « *!* », « *?* »).

(66) **L**a disparition de Philippe Frémeaux m'attriste énormément**.** **I**l nous a tant éclairé sans nous donner de leçon, à travers ses articles et en conférences**.** **A** titre personnel, j'ai en grande partie construit ma sensibilité économique avec ses écrits**.**

23. https ://www.cnrtl.fr/definition/connecteur

5.4 Niveau Lexical

Pictogramme utilisé pour remplacer un mot de la phrase Lorsqu'un pictogramme [24] prend la place d'un mot.

(67) J'ai même pas refléchis, j'ai d'abord écouter Kaaris & Bosh, avant de vouloir écouter le rest et c'est le 🔥 🔥 🔥

(68) RT @X : La 🇫🇷 aux côtés de nos amis libanais... #Beyrouth #Liban https ://t.co/tZQbU3Q59g

« *ça* » qui désigne une entité animée Lorsque « *ça* » est utilisé pour faire référence à une entité animée.

(69) @X @X @X @X @X @X @X @X @X @X @X @X Un Ministre ,**ça** ferme sa gueule ;si **ça** veut l'ouvrir,**ça** démissionne. (Jean-Pierre Chevènement)
https ://t.co/ACvWRKERYt

(70) 5ème fois ? Ah ouais les lyonnaises **ça** rigole pas hein bravo https ://t.co/pLOQKoAWq3

« *ça* » à la place de « *cela* » Lorsque la forme contractée de « *cela* » est utilisée, c'est à dire lorsque « *ça* » est présent dans le texte.

(71) Putain y a pas à dire l'Islam **ça renforce**, **ça fait** des hommes forts

Interjection Lorsqu'une interjection, c'est à dire un « mot invariable, autonome, inséré dans le discours pour exprimer, d'une manière vive, une émotion, un sentiment, une sensation, un ordre, un appel, pour décrire un bruit, un cri. » [25], est présent dans le texte.

(72) @X @X **Bas** j'vais pas dire jeune maître

(73) #BoycottNetflix **bahaha** les victimes qui tweet sérieusement ici pour le passage d'un film ça me détruit

(74) @X **bon bah** on tente sa chance **hein**, pas pour moi mais pour mon copain, il est temps que monsieur change de portable #XiaomiMi10

(75) **Allo** c'est de la fiction ! ! !

24. défini sous section 2.2.1
25. `https://www.cnrtl.fr/definition/interjection`

Expression idiomatique Selon (Caillies 2009) « parmi les éléments du lexique, nous trouvons, en plus des mots simples et complexes, des suites de mots figées dont le sens n'est guère prévisible. Les expressions idiomatiques en font partie. Elles constituent des locutions, connues comme telles et pouvant être répertoriées dans des dictionnaires, dont la signification est supposée ne pas résulter de la composition des significations des mots qui les constituent. » Autrement dit, les expressions idiomatique sont des expressions dont le sens est figuré. Le descripteur est présent lorsque une expression idiomatique apparaît dans le texte.

(76) Jviens de voir la dernière vidéo du raptor, sa **tire à balle réel**. https ://t.co/z0xMAZzKuO #Raptor #Ecologie #PLS

(77) @X Un incroyable album ça fait partie moi aussi des premiers albums que **j'ai grave poncé** sur Houston

(78) @X @X @X @X @X Macth amical ou pas l'enjeu est le meme gagner tu crois que sochaux ils veulent **se prendre une valise** ? Autant leurs dire qu'il vont pour pedre on ira plus vite

« *tu* » préféré au « *vous* » et « *on* » préféré au « *nous* » Lorsque le tutoiement est utilisé plutôt que le vouvoiement ou bien lorsque le sujet « *nous* » est transposé en « *on* ».

(79) @X Pas ma faute si **tu** sais pas apprécier un bon gameplay et qu'à la place **tu** préfères jouer à Fate Extra.

(80) @X En effet **on** a rajouté 1h 😭 😭 😭 😭

Emprunt étranger Lorsqu'une abréviation, un mot ou syntagme de langue étrangère est présent(e) dans le texte [26].

(81) **Wtf** Selena Gomez vient faire quoi dans mon top

(82) @X Alors explique à quoi sa sert de dénigrer le corps de tootatis même si c'est fesse son **fake** .

Orthographe électronique Lorsque dans le texte est présente une utilisation de l'orthographe électronique défini comme « le langage alphaphonétique typique de l'écriture en ligne comme une compétence liée au genre/support : elle vise à l'économie dans les gestes » [27].

(83) @X mais **g** appris comme ça

(84) @X non vrmt ma soeur, ma mère et moi **c** pareil. juste **c** dans mes gènes **c** la vie, **c** dieu il veut pas que jfasse comme mes darons

26. Gadet 2003
27. Paveau et Rosier 2008

(85) On va pas se mentir les VMA cette année ils ont frappé fort **ct** vrm bien entre Lady Gaga Ari BTS et grv tt le monde **gt** grv plaiz

Insulte Lorsque dans le texte est présente une insulte. Une insulte est définie comme « paroles ou attitude (interprétables comme) portant atteinte à l'honneur ou à la dignité de quelqu'un (marquant de l'irrespect, du mépris envers quelque chose) » [28].

(86) La dernière fois qu'on a vu Liverpool comme ça bah c'était à l'an 100 av J-C c'est vraiment **des fils de pute**

(87) #GGRMC gratuité des transports ,une mesure de gauche la **pouffiasse** de barbara vomit sa bile , elle préfère coucher avec le facho d' onfray

5.5 Niveau morphologique

Répétition de caractère Lorsque dans un seul mot un caractère est répété plusieurs fois (au moins deux fois), par exemple « *j'aiiiiiime les chats* ».

(88) @X **Ptdrrr** tu l'a vraiment fait

(89) Ma Lena tu me manque **trooop**, j'ai hâte de pouvoir revenir à Paris et passer de jolies moments avec toi. Plein de bisous ❤️

Agglutination Lorsqu'un terme qui résulte d'une agglutination. Une agglutination « consiste en ce que deux ou plusieurs termes originairement distincts, mais qui se rencontraient fréquemment en syntagme au sein de la phrase, se soudent en une unité absolue ou difficilement analysable » [29]. Par exemple : « *pomme de terre* » → (« *PommeDeTerre* » | « *pommedeterre* »).

(90) Après #**Boycott_dieuoff** on a #**BoycottNetflix**. Après les SJW les LGBT et les féministes (extrémistes) Je décerne donc officiellement la licence du FC Ouin Ouin aux religieux bienvenue chez nous.

(91) Les jeunes sont prêts à relever le défi et à rejoindre le #**ClubParis2024** 💪 https ://t.co/nBL7Qc3KXl

28. `https://www.cnrtl.fr/definition/insulte`
29. `https://www.cnrtl.fr/definition/agglutination`

Majuscule utilisée en dehors de son usage conventionnel Caractère mis en majuscule en dehors de leur utilisation conventionnelle, c'est à dire en dehors de la première lettre d'un nom propre ou bien en début de phrase [30].

(92) Les dernières secondes de la vidéo de lena quand Seb l'a tire vers **LUI PK PERSONNE MAVAIS PRÉVENU DJDJDKDKDKDJDB JE REPLAY LE MOMENT X1515** c **TROP** jmeurs de ouf je ship c trop

Subjonctif qui s'aligne sur le présent Une forme verbale « non standarde » du subjonctif qui mime la forme présent [31]

(93) Communiqué pour mon Fan Club : inscription au KIKADI réalisée... On y croise les doigts les types on y **croive** ! #RMCLive

Redoublement syllabique dans un mot Lorsqu'un mot est composé d'une même syllabe répétée deux fois, par exemple « dodo » [32].

(94) La vidéo de Léna situation elle m'a donné envie de partir en week-end avec mon **doudou** mais je suis broke et mineur donc Chill... anyways.... so...

(95) Franchement le son Kaaris Bosh il est gentil de fou connaissant les 2 rappeurs c'est pas **foufou**

Raccourcissement de mot Terme qui peut être plus court que leur version standarde [33] : par exemple « *cela* » → « *ça* », « *quoi* » → « *koi* », « *qu'il* » → « *kil* »... Ce raccourcissement peut parfois être dû à une troncation finale du mot [34] : « *bourgeois* » → « *bourge* » ou bien les exemples (97) et (98), ou bien à un procédé d'abréviation c'est à dire un « procédé par lequel on obtient une représentation graphique tronquée, mais suffisamment claire, d'un signe plus long » [35] comme dans l'exemple (96) : par exemple « *rendez-vous* » → « *rdv* ». Ce raccourcissement peut aller jusqu'à une lettre unique [36] : « *c toi ?* » ou bien exemple (99).

(96) @X @X @X @X Suis un peu plus le PSG tu verras k'il se débrouille pas mal et il a dit oui et sisi tu veux k'il rajoute **koi** de plus

30. Bilger et Cappeau 2004
31. Gadet 2003
32. Gadet 1997
33. Sommant 2005
34. Gadet 1997
35. `https://www.cnrtl.fr/definition/abr%C3%A9viation`
36. Sommant 2005

(97) @X @X @X @X Le portrait de Libé nous dit qu'elle touche 3.000 euros/mois pour deux mi-temps, on en connait qu'un : dirigeante de "Paris sans sida", **assoce** financée par la ville, pour 1682 euros/mois. Quel est ce second mi-temps ? On en sait toujours rien... https ://t.co/HsZoskzc9L

(98) @X Non, ils sont pareils. Ils détruisent ce qui ne leur plaît pas. On peut comparer ça aussi à des enfantillages. Et lorsque la destruction d'un tel symbole implique une trentaine d'**anar** débiles, je me fais vite mon opinion sur ce que pense les Martiniquais de cette action

(99) @X @X Ok, donc vous êtes apparues sur cette terre d'un seul parent. Je sais pas **pquoi** vous tenez absolument à être noires. Et oublier vos gênes blancs. **C** plutôt une richesse d'avoir deux gênes.

Terminaison de mot discriminante Lorsqu'une terminaison contribue à la connotation du mot[37]. Par exemples les termes qui se terminent pas « *-asse* » (exemple (100)), « *-iotte* » (exemple (101)), « *-o* » (exemple (102)), « *-ou* » et « *-ouze* » (exemples (103) et (104)).

(100) Le problème c'est que si les maires ne prennent pas ce genre de décision, les **gauchiasse** iront se plaindre et insulter ont les maires en disant " mais vous avez rien fait pr nous protéger !" #GGRMC

(101) Russell Westbrook qui sort du corona virus vient de faire gagner les Rockets pendant que Giannis a choke comme une sale **fiotte** Cheh sale chien https ://t.co/5k03fYr4VN

(102) @X Désolé, le viol est un crime #AdamaViolent est donc un criminel que soutiennent les pastèques **écolo gaucho collabo**. #JusticePourKevin #LaRacailleTue

(103) quand c'est qu'epic compte bouger leurs cul à régler le bug du pad ? ça ruine toutes mes cups la ça deviens **relou**

(104) Wine, boyfriend, job.. Je dis pas non mais j'aurais préféré dla **binouze** https ://t.co/pIBhACmXOz

37. Ilmola 2012

Dérivation d'un nom ou bien d'un adjectif en adverbe Lorsqu'un adverbe vient d'un nom commun [38] (exemple (105)) ou bien d'un adjectif (exemple (106)).

(105) @X #Cdiscount2emeDemarque ça serait quand même **vachement** chouette de gagner !

(106) @X @X @X Pourquoi mauvais exemple il y a qui a part lui en Suède **sérieusement** meme si oui il en fait trop

Verlan Le verlan est un « procédé de codage lexical par inversion de syllabes, insertion de syllabes postiches, suffixation, infixation systématique ; type particulier d'argot qui en résulte » [39]. Le verlan entraîne une augmentation du nombre de syllabes en /oe/ [40].

(107) @X non ca c'est thierry.. (le nouveau **keum** de mams au cas où tu saurais pas..)

(108) RT @X : Mdrr par exemple aucune entreprise du cac 40 n'est dirigée par un mec/ une **meuf** de stmg

(109) La F1 est **relou** on connais toujours les podiums. Après Monza les RB ne sont pas autant performantes mais par contre les McLaren sont hyper fortes bordel Sainz et Norris trop fort avec une super voiture.

5.6 Niveau phonologique

« *il* » remplacé par « *y* » Lorsque le pronom « *il* » est remplacé dans le texte par « *y* ».

(110) @X @X On me disait : Non **y faut** un 9 /ailier avant un Latéral , voir un DC ou milieu mdr mais trk moi je savais que j'avais raisons mdr

(111) Incroyable la twittosphere footix , tellement ça veut décrédibiliser la saison de Benzema je vois écris des " suarez 2015 ou kb9 2020 " mdrrrrrrr donc en 2015 **y doit** y avoir un idiot qui a du écrire " suarez 2015 ou torres 2009 " Naaaaaaannnn zidane 2006 ou fontaine 1978

38. Ilmola 2012
39. `https://www.cnrtl.fr/definition/verlan`
40. Gadet 2003

Suppression de certaines lettres due à l'élision ou l'apocope Phénomène de suppression de voyelles et de consommes dans certains termes mimétiques de phénomènes oraux tels que l'élision ou l'apocope [41].

(112) @X Heyy je te conseille japscan a utiliser avec AdBlock et c'est lourd **t'as** juste deux petites pubs sur le haut mais c'est pas gênant

(113) À **vot'** bon coeur, M'sieur dame @X @X @X, une cagnotte pour des sucettes à l'anis #AdamaViolent #AssaTraore #LesMiserables https ://t.co/g5PvgeCJpt

Onomatopée Lorsqu'une onomatopée est présente dans le texte [42]. L'Onomatopée est définie comme la « création de mots par imitation de sons évoquant l'être ou la chose que l'on veut nommer » [43].

(114) Après les SJW les LGBT et les féministes (extrémistes) Je décerne donc officiellement la licence du FC **Ouin Ouin** aux religieux bienvenue chez nous.

(115) RT @X : Leto et Ninho ils sont trop connectés **wesh** ils prennent leur douche ensemble ou quoi

41. Favart 2009
42. Ilmola 2012
43. `https://www.cnrtl.fr/definition/Onomatop%C3%A9e`

Références

Beccucci, Laurène (2018). “Pierre HALTÉ, Les émoticônes et les interjections dans le tchat. Limoges : Éditions Lambert Lucas, 2018”. In : *Communication et organisation* 54, p. 253-255.

Berners-Lee, Tim, Larry Masinter, Mark McCahill et al. (1994). “Uniform resource locators (URL)”. In :

Bilger, Mireille et Paul Cappeau (2004). “L’oral ou la multiplication des styles”. In : *Langage et société* 3, p. 13-30.

Brunot, Ferdinand (1922). *La pensée et la langue : méthode, principes et plan d’une théorie nouvelle du langage appliquée au français*. Masson et cie.

Caillies, Stéphanie (2009). “Descriptions de 300 expressions idiomatiques : familiarité, connaissance de leur signification, plausibilité littérale,«décomposabilité» et «prédictibilité»”. In : *LAnnee psychologique* 109.3, p. 463-508.

Domenget, Jean-Claude (2013). *La visibilité sur Twitter : un enjeu professionnel*.

Favart, Françoise (2009). “La représentation de" l’oralité populaire" dans quelques romans du second XXème siècle (1966-2006)”. Thèse de doct. Paris 10.

Gadet, Françoise (1997). “La variation, plus qu’une écume”. In : *Langue française*, p. 5-18.

— (2000). “Français de référence et syntaxe”. In : *Cahiers de l’Institut de Linguistique de Louvain* 26.1-4, p. 265-283.

— (2003). “Is there a French theory of variation ?” In : *International journal of the sociology of language* 2003.160, p. 17-40.

— (2007). *La variation sociale en français*. Editions Ophrys.

Ilmola, Maarit (2012). *Les registres familier, populaire et vulgaire dans le canard enchaîné et charlie hebdo : étude comparative*.

Jackiewicz, Agata et Marko Vidak (2014). “Étude sur les mots-dièse”. In : *shs Web of Conferences*. T. 8. EDP Sciences, p. 2033-2050.

Jacques, Anis (1999). “Internet, communication et langue française”. In :

Magué, Jean-Philippe, Nathalie Rossi-Gensane et Pierre Halté (2020). “De la segmentation dans les tweets : signes de ponctuation, connecteurs, émoticônes et émojis”. In : *Corpus* 20.

McQuail, Denis (2010). *McQuail’s mass communication theory*. Sage publications.

Mekki, Jade et al. (2018). “Identification de descripteurs pour la caractérisation de registres”. In :

Paveau, Marie-Anne (2013). “Genre de discours et technologie discursive. Tweet, twittécriture et twittérature”. In : *Pratiques. Linguistique, littérature, didactique* 157-158, p. 7-30.

— (2017). *L’analyse du discours numérique. Dictionnaire des formes et des pratiques*. Hermann.

Paveau, Marie-Anne et Laurence Rosier (2008). *La langue française. Passions et polémiques*.

Rao, Sudha et Joel Tetreault (2018). "Dear sir or madam, may i introduce the gyafc dataset : Corpus, benchmarks and metrics for formality style transfer". In : *arXiv preprint arXiv :1803.06535*.

Rebourcet, Séverine (2008). "Le français standard et la norme : l'histoire d'une «nationalisme linguistique et littéraire» à la française". In : *Communication, lettres et sciences du langage* 2.1, p. 107-118.

Sommant, Line (2005). "Impact des nouvelles technologies sur l'écrit : la rédaction du courriel et le langage SMS 1." In :