Analyse diachronique de concepts politiques dans un corpus en tunisien issu de Facebook

ZAMITI ASMA

Mémoire dirigé par Mathieu Valette

Soutenu le 29 juin 2015

Master d'Ingénierie Linguistique, parcours Recherche et Développement

Université Paris III Sorbonne Nouvelle

Résumé

Le présent travail se fixe comme principal objectif l'étude d'un corpus en tunisien composé de textes issu d'Internet dans le but de s'interroger sur les difficultés concrètes qui peuvent émerger lors du traitement automatisé de cette langue. Ces difficultés sont multiples. Le tunisien est minoré par rapport à l'arabe et n'a pas le statut officiel de langue, il n'y a de fait aucune norme codifiée pour le tunisien. En l'absence de toute norme et en raison du plurilinguisme qui caractérise la situation linguistique en Tunisie, l'écriture du tunisien peut être multiple mélangeant les systèmes d'écriture et les langues. Notre corpus de travail est constitué de commentaires en tunisien postés sur la page Facebook officielle de la radio tunisienne Mosaique FM entre janvier 2011 et décembre 2011. Cette période englobe la chute de la dictature avec le départ de l'ancien président Ben Ali (le 14 janvier 2011) ainsi que l'élection démocratique d'une Assemblée constituante (23 octobre 2011). Les commentaires reflètent les principaux événements qui ont marqué l'actualité tunisienne. Pour cette première exploration, nous ne procédons pas à la normalisation des graphies. L'analyse textométrique de certains concepts politiques exprimés dans les commentaires vise ici à essayer de déterminer si le choix d'un système d'écriture ou d'une langue est significatif dans l'expression des avis et opinions.

Mots-clés : textométrie, langue tunisienne, arabizi, analyse du discours politique

Remerciements

Je tiens à remercier :

Mathieu Valette pour sa confiance, ses conseils et ses encouragements tout au long de ce travail ;

Jean-Michel Daube, Marie-Anne Moreaux, Damien Nouvel, François Stuck et André Salem de m'avoir fait l'honneur de participer à mon jury ;

Pierre Marchal et François Stuck pour leur aide technique lors de la préparation du corpus ;

Serge Fleury pour sa bienveillance, sa disponibilité et son aide tout au long de ce master;

Julien Masanès et les équipes d'Internet Memory pour leur soutien tout au long de l'année ;

et, enfin, ma très chère amie Ikram pour son dévouement et son aide documentaire dans ce travail.

A tous, très sincèrement merci.

Table des matières

Introduction 5
1. Présentation des médias de travail : Facebook et Mosaïque FM 7
1.1. Présentation générale de Facebook 7
1.2. Internet, Facebook et la cyber-censure sous Ben Ali 9
1.3. La radio Mosaïque FM 10
2. Le tunisien : statut problématique et répercussions dans le TAL 12
2.1. L'idiome tunisien : éléments de définition 12
2.1.1. La conception diglossique de la répartition arabe-tunisien 12
2.1.2. Remise en question de la diglossie 13
2.2. Traitement automatique du tunisien 20
2.2.1. L'état de l'art 20
2.3. Remarques et critiques sur l'état de l'art 22
2.3.1. Observations globales 22
2.3.2. Langue et système d'écriture 23
2.3.3. Corpus et état de langue 23
2.3.4. Subordination du tunisien à l'arabe 24
2.3.5. Choix pour la constitution du corpus 25
2.4. Bref compte rendu sur les productions écrites en tunisien 26
3. Le corpus 28
3.1. Éléments de définition d'un corpus 28
3.1.1. Définition générale 28
3.1.2. Le critère de représentativité 28
3.1.3. Le critère d'homogénéité 29
3.1.4. La notion de réflexivité 29
3.1.5. Les séries textuelles chronologiques 30
3.2. Positionnement de notre corpus par rapport à l'état de l'art 31
3.2.1. Brève présentation de notre corpus 31
3.2.2. Discussion 31
3.2.2.1. Représentativité et homogénéité 31
3.2.2.2. Notion de réflexivité 32
3.2.2.3. La dimension chronologique 34

3.3. Présentation détaillée du corpus 34
3.3.1. Choix de la période 34
3.3.2. Langues et systèmes d'écriture 35
3.3.3. Balisage du corpus 36
3.4. Constitution du corpus 37
3.4.1. Récupération des données sur Facebook 37
3.4.2. Normalisation 37
3.5. Observations générales sur le corpus 39
3.6. Unités et parties du corpus 42
Exploration textométrique du corpus 45
3.7. Brève présentation de la textométrie 45
3.8. Segmentation et découpage des textes 46
3.9. Méthodes textométriques 46
3.10. Exemples d'applications 47
3.10.1. Témoignages dans les forums 47
3.10.2. Commentaires d'articles de presse 48
3.10.3. Posts sur une page Facebook 48
3.11. Exploration textométrique du corpus 49
3.11.1. Nouvelles formes : nouveaux thèmes et/ou nouveaux commentateurs 49
3.11.2. Deux périodes distinctes : politique et non politique 52
3.11.3. Les commentaires Facebook entre spam-attack et [banalités ?] du quotidien 61
3.11.4. Quelques exemples de cooccurrences 63
3.11.5. Les émoticônes : expression non verbales des émotions 86
Conclusion et perspectives 90
Bibliographie 92

Introduction

Entre décembre 2010 et janvier 2011 en Tunisie une série de manifestations et de sit-in mènent au départ de l'ancien président de la République tunisienne, Zine el-Abidine Ben Ali, et à la chute de la dictature qu'il avait instaurée depuis son arrivée au pouvoir en 1987. Au cours de ces événements les médias tunisiens, aussi bien publics que privés, n'ont joué aucun rôle alors que la répression se faisait de plus en plus violente. Internet, et notamment le réseau social Facebook, était alors devenu la principale source d'information pour le grand public en Tunisie et à l'étranger. Les médias étrangers se servaient des images et des vidéos prises sur les lieux par les manifestants pour relayer l'information, contribuant ainsi à donner plus de visibilité à ces contenus et à accroître leur propagation à l'intérieur même du pays. Facebook, qui occupait déjà une place centrale dans la vie sociale des Tunisiens bien avant ces événements, a continué de gagner en importance après le départ de Ben Ali. Aujourd'hui, c'est un lieu incontournable de la vie socio-politique tunisienne. Les partis politiques, grands absents de la révolution, ont dû s'approprier ce média pour construire leur image, véhiculer leurs idées et attirer les électeurs. Les médias traditionnels aussi ont investi ce réseau devenu incontournable pour développer et étendre sa communication. Facebook offre ainsi un aperçu non négligeable des débats socio-politiques qui animent une partie de la population tunisienne. Par ailleurs, ces conversations se déroulent le plus souvent en langue tunisienne. Le tunisien est la langue maternelle en Tunisie. Il est employé au quotidien dans tous types d'échanges entre individus. Habituellement parlé, son utilisation s'étend de plus en plus à l'écrit. Les réseaux sociaux, les sites communautaires, les espaces de commentaire des sites de presse, constituent autant de lieux où se déploie aujourd'hui l'écriture tunisienne avec un système d'écriture spécifique : le *arabizi*, combinant lettres latines et chiffres.

Le tunisien tel qu'on peut le lire sur les réseaux sociaux, n'est pas une variante "arabisée" ou "littérarisée" de la langue quotidienne, mais bien la langue telle que l'on peut l'entendre à l'oral. Par rapport à l'arabe, la langue tunisienne est catégorisée comme une "variante basse" (Laroussi 2002) et est de fait marginalisée et peu étudiée. Nécessaire pour l'analyse des réseaux sociaux, son traitement automatique est aujourd'hui très insuffisant. En TAL, il résulte, en effet, de la subordination du tunisien à l'arabe, un retard dans le développement d'outils et de ressources pour cette langue alors que l'arabe est richement outillé. La quantité grandis-sante de contenus en tunisien, disponible notamment grâce à l'essor des réseaux sociaux, a mis en évidence la réalité des

différences substantielles phonologiques, morphologiques, syntaxiques et lexicales entre le tunisien et l'arabe. Lorsqu'il a fallu traiter des données en tunisien, les ressources et les outils développés pour l'arabe se sont avéré inapplicables. De fait, des publications sur le tunisien commencent à paraître mais restent insatisfaisantes car fortement influencées par la minoration linguistique.

Le présent travail se fixe comme principal objectif de mettre à plat les difficultés qui peuvent émerger lors d'un traitement automatisé d'un corpus de textes écrits dans une langue peu dotée. Il s'agit d'un premier travail exploratoire via des méthodes textométriques sur un corpus en tunisien. Nous proposons pour cela d'analyser quelques concepts politiques avec l'outil de textométrie Lexico 3 (Salem et al., 2003). Notre corpus de travail est constitué de commentaires Facebook publiés sur la page officielle de la radio Mosaïque FM entre janvier 2011 et décembre 2011. Les sujets abordés sont presque exclusivement relatifs à l'actualité tunisienne et la langue employée est en majorité le tunisien même si quelques commentaires sont rédigés en français et en arabe. Outre le multilinguisme, le corpus intègre trois systèmes d'écritures : français, arabe et *arabizi*. Cette hétérogénéité, qui reflète la situation linguistique tunisienne, constitue à la fois un handicap pour l'application des méthodes textométriques en raison de la grande variété de graphies qui en découlent mais présente en même temps un véritable intérêt pour un premier travail exploratoire dans la mesure où les choix d'une langue plutôt qu'une autre ou d'un système d'écriture plutôt qu'un autre pourrait, ou non, significatif par rapport aux opinions exprimées. Nous commencerons par présenter les médias sur lesquels nous travaillons à savoir Facebook et Mosaïque FM. Après avoir présenté les problématiques liées au statut du tunisien (langue/dialecte), nous proposerons un compte rendu critique de l'état de l'art du traitement automatique du tunisien. Nous présenterons ensuite brièvement les différents critères de définition d'un "bon corpus" et nous essayerons de situer notre corpus de travail par rapport à ces critères. Enfin, la dernière partie de ce mémoire se consacrée aux résultats de l'analyse textométrique à proprement parler après un bref état de l'art de ce domaine.

1. Présentation des médias de travail : Facebook et Mosaïque FM

En 2014, 5.81 millions de Tunisiens, soit 53% de la population, utilisent internet. Parmi eux, 4.6 millions sont sur Facebook[1]. Le taux de pénétration classe la Tunisie 10ème au niveau africain et 66ème au niveau mondial. Le régime Ben Ali déjà œuvrait activement pour la promotion de l'internet et des technologies de l'information et de la communication (TIC) mais sa politique s'inscrivait "dans le paradoxe de la modernisation autoritaire, où l'État se mobilise en faveur de la diffusion des TIC, mais avec des modes de mobilisation qui entravent celle-ci" (Bras, 2007). En effet, avant la chute de la dictature en 2011, la Tunisie figurait régulièrement dans les classements des pays "ennemis d'internet" établis notamment par Reporters sans frontières qui dénonçait "la politique du président Zine el-Abidine Ben Ali en matière de Réseau" comme étant "l'une des plus liberticides de la planète"[2]. Dans ce climat liberticide, Facebook arrivera pourtant à se maintenir, gagnant en popularité chaque année, jusqu'à devenir le catalyseur de la contestation qui fera chuter la dictature. Dans cette partie, nous présenterons brièvement les principales caractéristiques du réseau social Facebook. Nous évoquerons la place de Facebook en Tunisie notamment face à la cyber-censure. Enfin, nous justifierons le choix de la page Facebook de Mosaïque FM en récapitulant le parcours de cette radio et sa situation dans le paysage radiophonique tunisien.

1.1. Présentation générale de Facebook

Facebook est un réseau social en ligne fondé par Mark Zuckerberg en 2004. Initialement destiné aux seuls étudiants de l'université de Harvard, puis ouvert aux étudiants d'autres universités et écoles, Facebook est rendu accessible à tous en 2006 à la condition d'être âgé d'au moins 13 ans et de posséder un compte de messagerie valide.

Selon le classement Alexa, entreprise appartenant au groupe Amazon qui fournit des statistiques sur le trafic du web mondial, Facebook est le site le plus visité après Google.com. Les chiffres publiés par Facebook indiquent que le réseau a enregistré une moyenne quotidienne de 936 millions

[1] Etude par l'Institut Tunisie Sondage
[2] RSF – La liste des 13 ennemis d'internet : http://bit.ly/1Iv5h8r

et une moyenne mensuelle de 1.44 milliards d'utilisateurs actifs sur le mois de mars 2015[3] ce qui en fait virtuellement le pays le plus peuplé au monde devant la Chine et l'Inde.

Jusqu'en 2014, les règles d'utilisation de Facebook exigeaient des utilisateurs de s'enregistrer sous leur nom civil. Depuis cette date, en raison de nombreuses critiques, Facebook a assoupli ses règles en autorisant les noms d'usage. Il n'est cependant pas rare que des utilisateurs s'enregistrent avec un pseudonyme. Chaque utilisateur dispose d'un profil personnel sur lequel il peut partager des publications de différents types : texte, image, vidéo. Les publications peuvent être publiques, visibles par tous les "amis" ou certains d'entre eux ou totalement privées.

Les "amis" sont les utilisateurs du réseau qui peuvent accéder au contenu du profil d'une personne. Cette relation, mutuelle, est supposée être le prolongement virtuel d'une relation, ou amitié, réelle, hors-ligne, mais c'est loin d'être systématique. Il existe un autre type de relation, unilatérale qui est "l'abonnement" et qui permet à des utilisateurs de recevoir uniquement les publications publiques d'un profil.

Les utilisateurs peuvent aussi créer ou rejoindre des groupes et des pages qui sont des espaces communautaires crées autour d'un centre d'intérêt : thème, activité, marque, personnalité, etc. Les pages peuvent être officielles, c'est à dire tenues par des représentants officiels d'une entité publique comme une marque ou une personnalité. Les contenus des pages peuvent être publics ou privés, accessibles à tous ou à une liste restreinte de personnes sélectionnées par le créateur/administrateur. Un utilisateur peut commenter le contenu des pages via son profil personnel ou en tant qu'autre page s'il en est l'administrateur.

Le contenu des pages est organisé dans un ordre antéchronologique par année. Les publications sont affichées au fur et à mesure que l'utilisateur déroule la page. Par défaut, ce sont les publications à la unes qui sont affichées, l'utilisateur doit le spécifier s'il veut afficher l'intégralité des publications. L'utilisateur peut aussi naviguer entre les années grâce à une frise temporelle ou *Timeline* qui remonte jusqu'à la fondation de la page. Les commentaires sous les publications ne sont pas tous affichés par défaut, et l'utilisateur peut appliquer des filtres pour gérer leur disposition par pertinence, popularité ou par date.

[3] Facebook – Facebook Reports First Quarter 2015 Results : http://bit.ly/1HVQYe6

1.2. Internet, Facebook et la cyber-censure sous Ben Ali

Avant la chute de la dictature, la cyber-censure en Tunisie s'inscrivait dans un processus plus global visant à contrôler la diffusion et l'échange d'informations et à empêcher les contenus hostiles au pouvoir en place. En plus du filtrage de l'accès au web, de la surveillance des contenus des courriers électroniques, des campagnes de *phishing* et de piratage de comptes, les autorités tunisiennes avaient mis en place un arsenal législatif pour réprimer les voix dissidentes comme le code de la presse qui interdit la diffusion d'informations que le pouvoir en place juge diffamatoires ou portant atteinte à l'ordre public ; ou encore la loi des "traitres" adoptée en 2010 sanctionnant l'atteinte à la sécurité économique de la Tunisie notamment en véhiculant une "mauvaise image" du pays à l'étranger. Les intimidations et les violences physiques étaient également fréquentes.

Si les sites d'opposition, les blogs de "cyberdissidents" (Lecomte, 2009), les sites de certaines ONG ou certains sites de presse critiques envers le régime sont facilement repérables et immédiatement censurés, les contenus circulant sur les sites de partages restent incontrôlables malgré le dispositif de surveillance massive. Sous-prétexte de combattre le terrorisme et la pornographie, le gouvernement s'attaque aux sites les plus populaires : YouTube et Dailymotion en 2007 puis Facebook, qui est rendu inaccessible le 24 août 2008. Mais face au tollé, le gouvernement est contraint de reculer. Le 2 septembre 2008, soit moins de deux semaines après sa censure, Facebook est de nouveau accessible en Tunisie. Ben Ali était intervenu personnellement pour demander la réouverture du réseau.

Facebook est donc très vite devenu un véritable phénomène en Tunisie, avec une popularité grandissante d'année en année, et, malgré d'autres tentatives, le gouvernement n'arrivera jamais à l'enrayer définitivement. Les cyberdissidents, dont la portée du discours était réduite par la censure, y avaient certes trouvé refuge pour pouvoir véhiculer plus amplement leurs idées, mais dans une société conservatrice, verrouillée par les valeurs religieuses et le poids des traditions, Facebook offrait surtout un lieu de rencontre, d'échange et de partage pour la jeunesse tunisienne.

Ce brassage marque, après la démocratisation des blogs, une nouvelle étape du "décloisonnement de la critique en ligne" (Lecomte, 2011) en permettant à des internautes qui n'étaient pas nécessairement sensibilisés à l'activisme politique d'être confronté au discours des opposants au régime de Ben Ali via les réseaux de contacts (ou "amis") sur Facebook. Ce qui, sans être à l'origine

des événements de 2011 comme le fait croire le mythe de la "révolution 2.0", jouera un rôle de catalyseur (Lecomte, 2011) en permettant de relayer et d'amplifier le soulèvement des populations dans les régions les plus déshéritées, notamment Sidi Bouzid, Kasserine et Gafsa (où un mouvement contestataire fortement réprimé avait déjà éclaté en 2008), qui, délaissées par le pouvoir en place, revendiquent leur droit au travail et à la justice sociale.

L'usage de Facebook a donc surtout permis de contourner les médias traditionnels qui n'ont pris leurs distances le régime pour dénoncer les violences policières qu'après le départ de Ben Ali. C'est via cette plateforme que les informations sur les événements qui secouaient les régions de l'intérieur ont pu être relayées grâce à la diffusion virale des témoignages, des photos et des vidéos amateurs réalisées au moyen de téléphones portables ainsi qu'à un important travail d'agrégation et de centralisation de l'information par des cyber-activistes. Ces contenus diffusés sur internet étaient également exploités par les médias étrangers notamment des chaînes satellitaires informant aussi bien la population tunisienne, notamment les personnes les moins connectées, qu'un vaste public à l'étranger des événements en cours. En effet, avant la fuite de Ben Ali et la chute de son régime, les médias traditionnels relaient exclusivement la parole du gouvernement. Une emprise totale qui n'épargnait pas les médias privés même les plus populaires comme la station de radio Mosaïque FM.

1.3. La radio Mosaïque FM

En 2003, le 7 novembre, date symbolique marquant l'anniversaire de son arrivée au pouvoir, le président Ben Ali annonce la libéralisation de l'espace audiovisuel avec l'ouverture de la première radio privée du pays : Mosaïque FM. Mais cette privatisation n'est "absolument pas garante de liberté d'expression, c'était même une manifestation de la mainmise du régime de Ben Ali" (Zeineb, 2012). Le choix "des personnes privées tunisiennes, obéit principalement aux critères d'allégeance au pouvoir politique". Le directeur de Mosaïque FM, connu pour "son allégeance totale et zélée à l'égard du pouvoir", était dénoncé par la Ligue Tunisienne de défense des droits de l'Homme[4] pour ses "écrits calomnieux contre les défenseurs des droits de l'Homme" (Larbi, 2007). L'autre personnage fort de cette station était Belhassen Trabelsi, frère de l'ancienne première dame Leïla Trabelsi, qui en détenait 13% du capital. Par ailleurs, la convention signée

[4] Ligue Tunisienne pour la Défense des Droits de l'Homme – Médias sous surveillance : http://bit.ly/1IhrMQ0

entre le gouvernement et les particuliers à l'époque stipulait que "le responsable, ainsi que le directeur de l'information de la station 'sont nommés en accord avec le gouvernement'". En ce qui concerne les programmes à proprement parler, la station ne devait pas diffuser des informations comportant des éditoriaux ou des commentaires ni "des nouvelles de nature à perturber l'ordre public et à porter préjudice à l'image de marque du pays" (Larbi, 2007).

Mais ce qui fera le succès de Mosaïque FM, qui est encore aujourd'hui la première radio privée de Tunisie[5], ce n'est pas la qualité de l'information qu'elle propose mais la légèreté du ton qu'elle adopte, en rupture totale avec le style affecté des médias publics. Les programmes sont présentés par une équipe jeune et dynamique en langue tunisienne ce qui apporte à l'époque un souffle nouveau au paysage audiovisuel plombé par la langue de bois qui se matérialise dans l'usage si peu naturel de l'arabe, ou de langues étrangères comme pour RTCI (Radio Tunis Chaîne Internationale).

Diffusant en FM et via son site internet. Mosaïque FM occupe aussi une place importante sur Facebook puisque sa page est la troisième parmi les pages de médias tunisiens en nombre de "fans" avec plus de 2.2 millions d'abonnés[6]. Avant la fuite de Ben Ali, le contenu qui y était diffusé concernait essentiellement la musique et le sport, notamment le football. A partir du 14 janvier 2011, les publications de la page, reflet de la programmation, se sont diversifiées et la politique y a pris une place importante. Il subsiste cependant des pratiques douteuses que nous avons noté lors de la collecte de notre corpus, par exemple la suppression des publications relatives à des périodes sensibles comme les élections législatives et présidentielles de 2014, phénomène que nous avons retrouvé dans toutes les autres pages de médias tunisiens que nous avons pu consulter. Malgré de nombreuses sollicitations aucun de ces médias n'a souhaité nous répondre.

Récapitulatif

- Facebook est le principal réseau social en Tunisie. Sa popularité est telle que même le régime autoritaire de Ben Ali n'est pas parvenu à le censurer.
- Facebook n'est pas à l'origine de la révolution tunisienne mais a surtout servi de catalyseur des mouvements de contestation.
- Mosaïque FM est la première station de radio tunisienne, sa popularité est notamment due à sa programmation en tunisien.
- Comme tous les médias tunisiens, Mosaïque FM était sous le contrôle du pouvoir en place sous Ben Ali, mais traite de l'actualité politique depuis la révolution.

[5] Sigma – SigMag des medias, de la communication et du marketing en Tunisie : http://bit.ly/1LtDc64
[6] Socialbakers – Media in Tunisia : http://bit.ly/1eLww53

2. Le tunisien : statut problématique et répercussions dans le TAL

"Fhimtkom" ("Je vous ai compris"), le 13 janvier 2011, pour la toute première fois en vingt-trois ans de pouvoir, Zine El Abidine Ben Ali s'adressait au peuple non plus en arabe mais en tunisien, "la langue de tous les Tunisiens et Tunisiennes". Si cette allocation marquera un tournant historique en raison des événements politiques qui ont suivi, avec la chute de la dictature et le début de ce que les observateurs occidentaux ont appelé le Printemps Arabe, elle amorcera aussi en Tunisie une certaine réhabilitation du tunisien dans la parole institutionnelle. Jusque-là adepte des discours prononcés en arabe littéral, dans une posture rigide et une attitude austère, Ben Ali tentait pour la première fois un rapprochement avec ce peuple qu'il avait tyrannisé et qui maintenant se soulevait. Le choix du tunisien, est puissant, à la fois dans le fait et dans la symbolique. La révolution tunisienne s'est en effet concrétisée dans le passage du culte de la personnalité (Geisser et Gobbe, 2008) à la surprenante humilité du "je ne suis pas le soleil qui brille sur toute chose", ultime aveu avant le départ. Mais ce passage s'est surtout fait dans la langue : de l'arabe au tunisien, élevé pour l'occasion au rang de "langue". Dans ce chapitre nous rendrons compte des problématiques sociolinguistiques liées au statut du tunisien. Nous présenterons par la suite un aperçu de l'état de l'art du traitement automatique du tunisien. Enfin, nous proposerons un bilan critique de ces travaux.

2.1. L'idiome tunisien : éléments de définition

2.1.1. La conception diglossique de la répartition arabe-tunisien

Ne jouissant pas du statut officiel de "langue", l'idiome tunisien se voit attribuer une grande variété d'étiquettes selon les diverses prises de position – car, tout du moins en linguistique, "toute étiquette est une prise de position" comme le souligne (Laroussi, 2002) – dans les travaux consacrés à la situation linguistique en Tunisie. Mais le plus souvent, c'est l'idée d'une hiérarchisation dépréciative et dévalorisante par rapport à l'arabe littéral qui est consciemment ou inconsciemment perpétuée. Que ce soit à travers l'étiquette "arabe vulgaire", "arabe local", "arabe parlé", "arabe maternel", "arabe dialectal" ou encore "arabe tunisien", la position traditionnelle et dominante définit le tunisien comme la variante régionale ou le versant dialectal d'une langue unique, l'arabe. Ces étiquettes (hormis "arabe tunisien") sont d'autant plus dévalorisantes qu'elles amalgament différentes langues en ne permettant aucune distinction entre les pays, la même étiquette pouvant

être appliquée aussi bien au tunisien, qu'à l'égyptien ou au libanais etc. alors même que les partisans de cette vision s'accordent à distinguer différents groupes de parlers ou "dialectes arabes", le tunisien étant rattaché au groupe des "dialectes arabes maghrébins".

Dans tous les cas, la coexistence du littéral et du dialectal est justifiée par une répartition fonctionnelle claire entre les deux idiomes selon le modèle de diglossie établi par Fergusson. On retrouve par exemple cette position chez (Baccouche, 1998) qui identifie "deux pôles apparentés à deux niveaux d'une même langue dont les registres possèdent un champ de dispersion très vaste", le niveau littéral touchant "aux confins du classique" et le niveau dialectal allant "jusqu'à l'idiome de l'analphabète". A côté de ce modèle diglossique binaire, (Laroussi, 2002) évoque le modèle de continuum proposé par M. Chaïeb avec "*al-fusha* (l'arabe littéraire), en variété acrolectale, *al-wusta* (l'arabe intermédiaire) en variété mésolectale et *al-dârija* (l'arabe tunisien) en variété basilectale". Mais, comme le note (Laroussi, 2002), même s'il s'agit d'une tentative de dépasser la distinction entre variété "haute" et variété "basse" qu'implique le modèle de Fergusson, ce concept de continuum "reconduit implicitement la hiérarchie qu'il tente de récuser" en plaçant "l'arabe littéraire ancien en haut de l'échelle et l'arabe tunisien en bas".

2.1.2. Remise en question de la diglossie

A contre-courant, il existe un autre positionnement, pas forcément défendu par des linguistes, comme le fait remarquer (Laroussi, 2002), selon lequel l'apparente dualité paisible dans la diglossie fonctionnelle cache en réalité des enjeux socio-politiques capitaux. (Laroussi, 2002) cite ainsi les exemples de (Ben Achour, 1995) qui voit dans cet arabe érigé constitutionnellement en langue officielle "une négation de soi" et de (Balegh, 1998) pour qui cette domination de l'arabe sur le tunisien, que l'on voudrait exclure de toute production intellectuelle, est "le pire des apartheids, l'apartheid linguistique".

La thèse la plus audacieuse est cependant soutenue par A. (Elimam, 1997, 2009 et 2012) qui conteste la vision dominante selon laquelle les idiomes du Maghreb seraient des dialectes arabes. Selon lui "les trois pays de l'Afrique septentrionale présentent un profil sociolinguistique quasi identique" avec "deux aires linguistiques distinctes : l'une chamito-sémitique et, l'autre sémito-méditérranéenne" où "se profilent deux langues vernaculaires à la fois natives et naturelles". La première, majoritaire, réuni les parlers aujourd'hui désignés, à tort, comme "dialectes arabes". La

deuxième, minoritaire, "est traditionnellement désignée par le générique "berbère" ou "tamazight", appellation préférée par les militants de la berbérophonie.

Réunissant ainsi les "dialectes arabes" du Maghreb sous le nom de "maghribi" – nom inspiré de l'appellation "maghrébi" utilisée, entre autres, par Fergusson et Marçais pour désigner ces idiomes – (Elimam, 1997, 2009 et 2012) affirme que cette langue unique est en réalité punique et non arabe, qui toutes deux sont sémitiques. La confusion, entre "arabe" et "sémitique", avait été créée, selon lui, par les orientalistes qui ont étiqueté cette langue comme "dialecte arabe", "sans précaution méthodologique rigoureuse (...) comme si l'on pouvait dire par exemple que l'hébreu est un 'dialecte arabe'".

Pour (Elimam, 1997, 2009 et 2012), le substrat punique constitue "environ 50% de l'actuelle langue vernaculaire majoritaire du Maghreb". En effet, le punique, étant une langue "native et maternelle", "a traversé le temps en empruntant aux autres formations langagières que le Maghreb a pu porter. Il s'est enrichi d'apports variés (berbère, latin, grec, turc, arabe, etc.)". Mais, s'il y a eu "des apports lexicaux, stylistiques et même, parfois, morphologiques – ce qui n'est pas étonnant en domaine sémitique", il n'y a pas eu substitution. (Elimam, 1997, 2009 et 2012) affirme d'ailleurs que c'est "ce patrimoine sémitique qui sera mis à profit dans la rencontre avec le texte coranique et le message islamique qui lui est associé". Il n'y aurait donc pas eu "arabisation" du Maghreb mais "islamisation" et affirmer le contraire "n'est qu'une vue de l'esprit".

Selon (Elimam, 1997, 2009 et 2012) donc "le fait punique ne saurait être incontournable" et "vouloir démunir l'histoire du Maghreb de son passé punique revient à lui spolier la mémoire". Or pour l'auteur "imposer une langue extérieure au corps social" est une entreprise vouée à l'échec car comme il le rappelle "les langues natives se reproduisent et traversent l'histoire quand bien même elles sont minorées et que les différents pouvoirs les marginalisent". Dénoncer l'entreprise de minoration que subit le "maghribi" devient alors nécessaire pour "repousser le concept de diglossie au profit de celui d'un bilinguisme d'où les vernaculaires ne sont pas exclus".

(Laroussi, 2002) – qui cependant opte pour l'étiquette "arabe maternel" – dénonce lui aussi le processus de minoration linguistique qui selon lui vise à faciliter la mise en place du système diglossique en dévalorisant la langue maternelle. Ce concept de "minoration linguistique", emprunté à Jean-Baptiste (Marcellesi, 1980), se définit comme le processus "par lequel des systèmes virtuellement égaux au système officiel se trouvent cantonnés par une politique d'Etat

certes, mais aussi par toutes sortes de ressorts économiques, sociaux dans lesquels il faut inclure le poids de l'histoire, dans une situation subalterne, ou bien sont voués à une disparition pure et simple".

Aussi bien (Laroussi, 2002) que (Elimam, 1997, 2009 et 2012) dénoncent les stratégies déployées dans certains discours, qui sous couvert de scientificité, sont en réalité mus par des idéologies à la fois linguistiques, politiques et religieuses. (Laroussi, 2002) qui qualifie ces idéologies tour à tour de "pro-arabe littéraire", "négatrice de l'arabe maternel", "nationnaliste panarabique" et "salafiste", démontre que si les détracteurs de "l'arabe maternel" lui dénient "le statut même de langue" – en lui reprochant notamment d'être non-scientifique, non-normé, non-prestigieux, non-national et profane – leurs arguments ne se basent pas réellement sur un état de faits strictement linguistiques mais reposent plutôt sur "une illusion fantasmatique puisant ses motivations dans des considérations idéologiques historiquement construites". L'effet pervers de cette posture idéologique étant qu'elle permet à ses tenants de "justifier la minoration par ses propres effets".

2.1.2.1. Argument linguistique

D'un point de vue linguistique, si le tunisien est déprécié au profit de l'arabe, c'est en raison de l'absence de norme d'une part et de sa nature "non-scientifique" d'autre part. Or, l'absence de "norme codifiée", et non l'absence de norme tout court comme le nuance (Laroussi, 2002), est "la conséquence d'un processus de minoration historiquement et socialement situé" et aucunement "une donnée intrinsèque" à la langue. Ce n'est donc pas le système linguistique du tunisien qui ne se prête pas à la codification mais l'absence de volonté de la part des chercheurs et des académiciens qui prive cette langue d'un système codifié et normé. Or, (Laroussi, 2002) rappelle que "la codification de la grammaire de la variété minorée est une entreprise qui ne peut avoir lieu que si les usagers de la langue non seulement en sont conscients mais aussi en voient l'utilité". De même, l'idée selon laquelle le tunisien serait "exclusivement oral" est entièrement fausse, et nous reviendrons sur ce point ultérieurement, puisqu'il existe des productions écrites dans cette langue.

Cependant, s'il est vrai qu'il n'existe pas de publications scientifiques en tunisien, (Laroussi, 2002) rappelle que ce critère "doit être observé en diachronie, sans perdre jamais de vue les potentialités de développement et d'enrichissements dont toute langue est porteuse". Pour appuyer son propos, il cite les exemples de l'allemand, qui n'était pas considéré comme une langue scientifique au 18ème siècle, ainsi que du français, concurrencé par le latin, langue savante par excellence, durant

des siècles. Le caractère scientifique d'un idiome, sa richesse, ou sa pauvreté, "reflètent fidèlement les représentations des locuteurs qui les parlent". L'enjeu réel étant d'accepter de "s'approprier une langue dite "pauvre" pour que celle-ci s'enrichisse de toutes sortes de notions".

N'étant ni scientifique ni codifié, le tunisien serait par rapport à l'arabe une langue basse ou non-prestigieuse. Cet argument fondé sur la "notion de base dans le dispositif théorique de Fergusson" qui fait la distinction entre "langue de prestige" et "langue de moindre prestige", relève en réalité, selon (Laroussi, 2002), d'un "discours émotionnel fonctionnant dans la plupart des cas sous forme de connotation". Cette dépréciation n'est jamais basée sur des "arguments linguistiques fiables" mais est plutôt motivée par des "considérations extralinguistiques" à la fois "historiques, sociales [et] idéologiques". (Elimam, 1997, 2009 et 2012) de son côté dénonce la "confusion endémique" entretenue par les acteurs de l'arabisation qui, "en assimilant la langue sémitique du Maghreb septentrional à une "dégénérescence" de l'arabe", au point que "l'équivalent arabe de "dialecte" signifie "sous-langue", voire langue-fille", maintiennent injustement cette hiérarchie.

(Elimam, 1997, 2009 et 2012) dans le même ordre d'idées note "le peu (pour ne pas dire la quasi-absence) de compétitivé de la langue arabe sur le marché mondial de productions scientifiques et technologiques". En effet, en se référant à la réalité des pratiques dans les pays du Moyen-Orient, il constate que "l'enseignement (scientifique) supérieur se fait en anglais et/ou en français, et les ouvrages disponibles en langue arabe sont généralement des traductions d'ouvrages étrangers". Vu sous cet angle, l'arabe ne serait pas plus "scientifique" que ses "dialectes".

(Elimam, 1997, 2009 et 2012) souligne même que le maghribi, langue naturelle, préexiste à l'arabe "qui est une élaboration (in vitro) relativement récente (Xe-XIe siècle J-C). L'élaboration de la norme arabe qui a eu lieu au VIIe siècle avait pour but de "doter les musulmans d'un même code de lecture du Livre Saint". Or "la forme linguistique du Coran n'est pas la réplique d'une langue unique mais fait fonds sur un ensemble de langues" soit comme l'écrit (Hadj-Salah, 1978), la fusion de "35 idiomes appartenant à des tribus différentes". D'ailleurs, (Elimam, 1997, 2009 et 2012) rappelle que "la forme linguistique du Coran n'a jamais, au grand jamais, donné corps à une langue maternelle et naturelle". En effet, "il n'est attesté nulle part d'enfant qui soit né avec la forme linguistique du Coran comme langue maternelle. Et cela malgré les 15 siècles d'efforts permanents d'arabisation".

Pour (Elimam, 1997, 2009 et 2012), même l'arabe moderne, simple "aménagement bureaucratique de la norme arabe", qui "se cristallise dans les médias et la littérature", "n'est jamais parvenu – pour des raisons différentes – à devenir la langue maternelle de quiconque". Car, "une fois la zone de l'écrit franchie, les locuteurs arabes reviennent à leurs langues natives et maternelles". Cette idée se retrouve aussi chez (Ben Achour, 1995) pour qui la question ne se limite pas à une affaire de "double registre linguistique", la distance, selon lui, n'étant pas entre "deux niveaux du réel" mais entre "l'utopie et le vécu". Le locuteur de l'arabe "peut discourir, non point vraiment parler. Il se met en scène pour pouvoir s'exprimer, il se dédouble deux fois. L'acteur souffre et les auditeurs entendent une déclamation. Celle-ci peut être bouffonne ou tragique. Elle passera rarement comme la respiration".

La notion de prestige est aussi remise en question par (Gibson, 2002) qui démontre que, contrairement à l'hypothèse communément admise, l'arabe littéral ne constitue pas "the main influence in the changes that are going on in Arabic dialects today". En étudiant l'évolution de phénomènes phonologiques et morphologiques en tunisien, il arrive à la conclusion que dans tous les cas "the shift is toward the variety of Tunis". Ainsi, même si le déclin du phonème /g/, caractéristique du parler bédouin, au profit du /q/, forme privilégiée à la fois dans le parler citadin et par les locuteurs lettrés, car proche de l'arabe littéral, peut laisser penser qu'il s'agit de la manifestation de l'influence de l'arabe littéral sur le tunisien et l'alignement de ce dernier sur la forme standard que représente l'arabe littéral, l'observation d'autres variables, moins ambigües, montre qu'il n'en est rien. Le traitement de la voyelle finale dans les formes verbales au pluriel permet de constater que le parler bédouin, bien que plus proche de l'arabe standard, tend à s'aligner sur le parler citadin qui lui s'en éloigne. De même, la non distinction en genre pour la seconde personne, aussi bien dans le système pronominal que verbal, spécifique au parler citadin, tend à s'imposer alors même que le parler bédouin marque cette distinction tout comme l'arabe littéral. Enfin, la réalisation des diphtongues /ay/ et /aw/, de l'arabe littéral, en /ii/ et /uu/, tous deux caractérisitques du parler de Tunis, et l'évaluation négative des locuteurs employant ces deux diphtongues selon la prononciation originelle, va également dans le sens d'une normalisation globale au profit du parler tunisois. Selon l'auteur, ce constat met à mal l'idée que l'arabe littéral serait l'unique norme et l'unique variété de prestige, notamment en ce qui concerne l'usage quotidien de la langue.

Notons, enfin, que dans plusieurs publications portant sur les "dialectes arabes", indépendamment de la thèse qui sous-tend ces travaux, il revient souvent l'idée que ces "dialectes" ne se limitent pas en réalité à de simples variantes régionales à peine déformées d'une langue principale, l'arabe. (Baccouche, 1998) écrit par exemple que "si nous examinons de près d'une manière comparative les structures du littéral et du dialectal, à tous les niveaux, nous concluons qu'il s'agit typologiquement de deux langues différentes bien que nettement apparentées". La distinction langue/dialecte dépend ici clairement d'un postulat de départ qui ne repose pas principalement sur des considérations purement linguistiques.

2.1.2.2. Argument politico-religieux

L'examen strictement linguistique de la question des "dialectes arabes" invalidant les arguments des partisans de l'arabe littéral pour justifier son hégémonie, il faut adopter une autre grille de lecture pour comprendre les motivations et les enjeux du débat. (Elimam, 1997, 2009 et 2012) pose ainsi la question : "le refus de prendre (officiellement) en considération ne serait-ce que l'existence de l'être social langagier (sans parler de sa nécessaire croissance) n'est-il pas motivé par les ambitions de pouvoir exclusif ?"

D'un point de vue strictement politique, ce que les "nationalistes panarabiques", comme les identifie (Laroussi, 2002), reprochent à "l'arabe maternel" c'est de concurrencer l'arabe littéral "symbole de l'unification de la nation arabe". D'ailleurs, seul l'arabe littéral est "considéré comme la variété nationale" où "nation désigne ici non l'Etat-nation (Tunisie, Maroc, Algérie, par exemple) mais la "Grande nation" arabe, une supra-nation en quelque sorte." La langue maternelle est alors vue comme une "langue anti-nationale" qui entrave "la constitution de la Grande patrie arabe". Cette idéologie panarabique dans son "projet de dépassement des différences nationales" ambitionne donc d'abolir les spécificités linguistiques, et donc culturelles et civilisationnelles, pour que "l'Etat-nation [perde] toute sa signification" et "se [fonde] dans l'Etat supranational". Pour (Elimam, 1997, 2009 et 2012), la "vision généreuse" qui a motivé l'élaboration de la norme linguistique arabe a été "vite subsumée par la mécanique de la reproduction du pouvoir temporel qui, précisément prend appui sur cette perspective (illusoire) d'unicité linguistique".

(Elimam, 1997, 2009 et 2012) cite l'exemple du berbère "particulièrement en Kabylie" comme étant "le témoignage de la répression plus ou moins voilée, plus ou moins douce des langues authentiquement nationales… dans le cadre d'applications de politiques se voulant nationalistes".

L'exemple peut aussi s'appliquer à la Tunisie où l'Etat a toujours adopté une politique de marginalisation jusqu'au déni des berberophones (Pouessel, 2012). Mais, comme le souligne (Elimam, 1997, 2009 et 2012) si le berbère peut prétendre à une émancipation juridique, les "dialectes arabes" ne le peuvent pas en raison du "raccourci lexical" qui les assimile à l'arabe, langue nationale, rendant de fait leur reconnaissance redondante, d'où l'importance cruciale du choix de l'étiquette pour désigner ces idiomes. La politique d'arabisation intensive continue donc imperturbablement notamment via l'enseignement que (Elimam, 1997, 2009 et 2012) identifie comme "la principale agence d'arabisation". Une définition qui s'applique particulièrement au contexte tunisien comme on peut le lire dans l'Article 39 de la Constitution tunisienne de 2014 qui stipule que "l'Etat veille aussi à enraciner l'identité arabo-musulmane et l'appartenance nationale dans les jeunes générations et à ancrer, à soutenir et à généraliser l'utilisation de la langue arabe".

L'idéologie panarabique se double selon (Laroussi, 2002) d'une autre idéologie qu'il qualifie de "salafiste" ou "traditionnaliste". Celle-ci, rejoignant "le panarabisme dans la dévalorisation de la variété maternelle", plaide "pour le retour aux sources, c'est-à-dire au mythe de l'âge d'or et de l'apogée de la civilisation arabo-islamique". Bien loin des observations linguistiques, ce discours religieux rejette le tunisien au profit de l'arabe littéraire qui, "en tant que langue du Coran, est considéré comme une variété sacrée" qu'il faut "contempler" et non "modifier, car on risquerait de la dénaturer". Le tunisien, comme tous les "dialectes arabes" d'ailleurs, serait alors une atteinte, quasi-blasphématoire, à un idéal immuable. Mais comme le rappelle (Laroussi, 2002), "la grammaire du littéraire [ayant] été codifiée au Moyen âge (...) on peut se rendre compte facilement du fossé qui existe entre ladite grammaire et la réalité des pratiques langagières".

On peut cependant déplorer que ni (Laroussi, 2002) ni (Elimam, 1997, 2009 et 2012) n'évoque dans ses travaux le cas du maltais qui rend bien compte du poids du contexte géopolitique et religieux dans les questions linguistiques ainsi que de l'importance des initiatives personnelles pour y faire contrepoids. Le maltais, que (Vanhove, 1997) définit comme étant "originellement un dialecte arabe de type maghrébin citadin, vraisemblablement proche de celui des vieilles cités tunisiennes" et dont "le statut (...) jusqu'au début du 20ème siècle est celui d'une langue parlée" (Vanhove, 1999), est devenu "la seule langue nationale de l'Archipel", depuis l'Indépendance en 1964, et "co-existe avec l'anglais comme langue officielle" ce qui est, comme le souligne (Vanhove, 1994), un "cas unique pour un dialecte arabe". L'enseignement en maltais à l'école

avait, lui, commencé dès 1934 après l'adoption officielle d'un "alphabet maltais en caractères latins", "œuvre d'un groupe d'écrivains et de grammairiens" (Vanhove, 1999).

2.2. Traitement automatique du tunisien

2.2.1. L'état de l'art

Les arguments en faveur de la reconnaissance du tunisien – et plus généralement des "dialectes arabes" – en tant que langue distincte, nous venons de le voir, sont nombreux. Il n'en demeure pas moins que la position dominante catégorise le tunisien comme une variante régionale, déformée du fait de l'oralité et de l'analphabétisme, d'une langue unique, l'arabe. C'est majoritairement sur ce postulat que se fondent aujourd'hui les nombreux travaux en TAL qui s'intéressent à l'arabe. Il résulte de cette situation – qui fait écho à l'amalgame fait et entretenu entre l'arabe et ces langues mentionné supra – un retard dans le développement d'outils et de ressources pour le traitement automatique des "dialectes arabes" dont le tunisien. Or, lorsqu'il a fallu traiter ces "dialectes", le domaine du TAL s'est trouvé confronté à un obstacle inéluctable : les outils et les ressources développées pour l'arabe littéral sont inutilisables pour les "dialectes arabes". La réalité des différences substantielles phonologiques, morphologiques, syntaxique et lexicales entre ces derniers et l'arabe littéral s'est alors imposée comme une évidence. L'intérêt pour ces langues dans leurs spécificités commence donc à croître. Pour le tunisien, quelques rares travaux existent avec différentes approches.

2.2.1.1. Traduction arabe-tunisien

(Zribi *et al.*, 2013) exploitent un analyseur morphologique de l'arabe pour l'adapter au "dialecte tunisien". La méthode proposée suit deux étapes. Dans la première étape les auteurs génèrent un lexique tunisien à partir d'un lexique arabe. La deuxième étape consiste à extraire les racines et les patterns morphologiques sur le modèle arabe. Le lexique enrichis avec ces informations morphologiques peut ensuite être traité avec un analyseur morphosyntaxique initialement constitué pour l'arabe.

(Hamdi *et al.*, 2013) présentent "un système de traduction de verbes entre arabe standard et arabe dialectal" qui repose essentiellement sur le lexique et la morphologie. La méthode proposée relève d'une architecture de transfert au niveau morphologique. Le lexique est formé par les occurrences verbales extraites du corpus de l'Arabic Tree Bank, composé de transcriptions d'émissions d'actualité en arabe littéral diffusées par différentes chaînes arabes. Ce lexique en

arabe littéral est ensuite traduit en tunisien pour créer un nouveau lexique en tunisien. Au final, à chaque entrée du lexique, en arabe et en tunisien, est associé un couple (racine, MBC), où MBC désigne une classe morphologique ou "Morphological Behavioural Class".

(Boujelbane *et al.*, 2013) suivent une approche similaire à celle de (Hamdi *et al.*, 2013) et (Zribi *et al.*, 2013) pour la création d'un lexique et la génération d'un corpus en tunisien. Les formes verbales sont extraites à partir de l'Arabic Tree Bank puis lemmatisées. Ces lemmes sont ensuite traduits vers le tunisien et pour chaque lemme un pattern (correspondant au "MBC" chez (Hamdi *et al.*, 2013) et une racine suivant le modèle arabe sont construits. Ce travail permet la constitution d'un dictionnaire bilingue dans lequel chaque forme arabe et tunisienne est associée à un lemme, un pattern et une racine. Enfin des règles syntaxiques permettent de réaliser une traduction automatique afin de transformer un corpus en arabe littéral en un corpus en tunisien.

2.2.1.2. Normalisation orthographique

(Zribi *et al.*, 2014) proposent, dans la continuité de (Habash *et al.*, 2012), une orthographe standardisée pour le traitement automatique du tunisien. L'idée étant de convertir les productions écrites en tunisien vers cette norme orthographique afin de pouvoir les traiter par la suite. Pour pouvoir appliquer les normes orthographiques préétablies les auteurs décortiquent les spécificités syntaxiques et morphologiques du tunisien afin d'adopter une conversion cohérente. Les auteurs listent cinq visées majeures à ce système, appelé TUN CODA pour Tunisian Conventional orthography for Dialectal Arabic : le système attribue à chaque mot une seule orthographe, il a été développé pour une utilisation en TAL, il utilise l'alphabet arabe uniquement, il permet d'unifier tous les dialectes arabes, il chercher à atteindre un équilibre optimal entre le maintien des spécificités du dialectal et l'établissement de conventions basées sur les similarités entre le littéral et le dialectal.

2.2.1.3. Création d'ontologies

(Graja *et al.*, 2011) proposent une méthode basée sur les ontologies pour la compréhension du tunisien dans le cadre d'un système de dialogue homme-machine. En partant d'un corpus oral existant (TuDiCol ou Tunisian Dialect Corpus Interlocutor, constitué à partir d'enregistrements de dialogues entre voyageurs et agents dans les stations de train tunisiennes) transcrit manuellement, les auteurs construisent une ontologie spécialisée pour l'annotation sémantique et l'interprétation du tunisien dans le cadre restreint des voyages ferroviaires. Pour chaque émission, les mots-clés

relatifs au lexique du domaine ont été extraits, annotés, regroupés thématiquement et reliés entre eux pour identifier les relations sémantiques qui les lient.

Dans une approche similaire, (Bouchlaghem *et al.*, 2014) présentent TunDiaWN, une ressource lexicale pour le tunisien visant à enrichir la base de données lexicale WordNet. Le corpus utilisé ici est construit à partir de la collecte de données sur diverses sources comme des sites web, les réseaux sociaux, des dictionnaires du tunisien, des transcriptions phonétiques, etc. Afin de traiter les grandes variations orthographiques dues à l'absence de norme codifiée, les auteurs ont pris soin d'enrichir la structure de leur base de données pour rassembler les différentes graphies possibles d'un même mot. Les variantes régionales ont également été prises en compte dans la structuration des données.

2.2.1.4. Etiquetage morpho-syntaxique

(Hassoun et Belhadj, 2014) s'intéressent spécifiquement à l'écriture *arabizi*, sur laquelle nous reviendrons ultérieurement, en constituant un corpus à partir de publications collectées sur les réseaux sociaux. Dans un objectif d'analyse de sentiments ils proposent un étiquetage morphosyntaxique, et donc une analyse morphosyntaxique approfondie, avec un effort de normalisation à travers une "correction orthographique". Il est par ailleurs intéressant de voir que les auteurs ont tenu compte, pour l'analyse de sentiments, de la signification de certaines pratiques spécifiques, même si elles contrarient la norme orthographique établie, telles que la répétition des voyelles qui correspond à un allongement du phonème à l'oral et qui indique une intensité dans le propos, qu'il soit positif ou négatif.

2.3. Remarques et critiques sur l'état de l'art

Ce bref compte rendu de l'état de l'art du traitement automatique du tunisien appelle quelques remarques.

2.3.1. Observations globales

Tout d'abord, si à l'échelle de l'état de l'art du TAL, les publications concernant le tunisien restent mineures, il est intéressant de constater l'intérêt grandissant pour cette langue surtout depuis la révolution tunisienne. Ces publications sont en effet toutes postérieures à 2011.

Ensuite, nous relevons de nombreuses approximations linguistiques dans la grande majorité de ces travaux notamment dans la définition du tunisien et donc de l'objet d'étude même de ces travaux. Or il nous semble qu'en TAL le volet linguistique n'a pas simplement valeur d'accessoire. Plutôt

qu'un "pré-texte", les questions linguistiques sont centrales en cela qu'elles peuvent orienter la réflexion, conditionner la rigueur méthodologique et légitimer les parti-pris techniques.

2.3.2. Langue et système d'écriture

(Hassoun et Belhaj, 2014) par exemple définissent "le dialecte arabe" comme étant "une langue mélangée avec de nombreuses autres langues" et l'écriture *arabizi* ou *arabish* comme "une nouvelle langue" ou encore "une langue proche du dialectal" dont ils attribuent l'apparition à l'émergence des réseaux sociaux dans le monde arabe. Or, cette écriture est pratiquée depuis les années 90s avec la généralisation de l'usage des téléphones cellulaires et notamment de la communication via les SMS. Les appareils téléphoniques n'étant le plus souvent pas dotés de claviers arabes, les utilisateurs ont simplement transcrit les lettres arabes en lettres latines en compensant les phonèmes qui n'avaient pas d'équivalent dans l'alphabet latin par des chiffres. Il s'agit donc uniquement d'un système d'écriture où les "dialectes" mêmes sont orthographiés au moyen de la transcription plutôt qu'avec l'alphabet arabe. Les auteurs confondent donc arabe, tunisien, emprunt et bilinguisme, d'une part – la langue tunisienne étant riche de nombreuses influences bien antérieures à l'arabe et les tunisiens étant pour la plupart bilingues en français – et, d'autre part, langue et système d'écriture. Pour la tâche de détection de la langue, ils isolent donc d'emblée les messages rédigés en lettres arabes qu'ils considèrent comme étant ipso facto des messages en arabe. Cependant, le tunisien peut aussi bien être écrit en *arabizi* qu'en alphabet arabe. Dans ce dernier cas, les utilisateurs ont tendance à appliquer le même principe de l'écriture *arabizi*, mais dans l'autre sens, en écrivant les termes "non-arabes" qu'il s'agisse d'emprunt ou de bilinguisme, en lettre arabes. Parfois, les utilisateurs peuvent mélanger, dans un même message, les deux écritures et/ou les deux langues, tunisien et français. Il s'agit ici de la réalité des pratiques langagières qui ne peuvent échapper à une observation neutre sans *a priori* sur l'idée que l'on se fait de la langue étudiée et qui est trop souvent tributaire des idéologies dont nous avons fait l'inventaire précédemment. En ignorant cette réalité, les auteurs se privent de données précieuses pour constituer et étoffer leur corpus tout en évacuant des difficultés qui sont pourtant inévitables si l'on s'intéresse à l'écriture tunisienne sur les réseaux sociaux.

2.3.3. Corpus et état de langue

On retrouve ces mêmes *a priori* idéologiques chez (Graja *et al.*, 2011) qui justifient le choix de l'approche lexicale, que nous ne remettons pas en cause en tant que telle, par le fait que les principales caractéristiques du "dialecte tunisien" sont la brièveté des émissions et le non-respect

de la grammaire. Or la brièveté des émissions n'est pas imputable à la nature même du tunisien mais à la nature du corpus étudié qui compile des échanges entre voyageurs demandant des informations et des agents dans des gares ferroviaires. Il est de fait évident que ce type d'émissions ne peut pas être très développé et très élaboré. Ensuite, le non-respect de la grammaire comme "caractéristique" du tunisien est une affirmation pour le moins discutable. De quel grammaire s'agit-il ? De celle de l'arabe ? Si oui, une langue ne peut évidemment pas respecter la grammaire d'une autre langue quand elle respecte la sienne, qui lui est propre. S'il s'agit de la grammaire tunisienne, comment pourrait-il seulement y avoir une grammaire tunisienne si l'une des caractéristiques premières du tunisien est de ne pas respecter cette grammaire ? A partir de quoi ce serait alors forgée cette grammaire pour devenir un standard dont on peut s'écarter ? Ce constat n'est-il pas plutôt, encore une fois, imputable à la nature même de la situation d'énonciation caractéristique de ce corpus ? Ces affirmations pour justifier le choix de l'approche nous semble donc bien expéditifs et peu prudents. La question de la représentativité se pose pleinement ici. Si de nombreux chercheurs estiment que la représentativité est un critère du corpus (McEnery et Wilson, 2001) ou dépend des objectifs visés par la recherche (Bowker et Pearson, 2002), d'autres soutiennent l'idée qu'un corpus aussi vaste soit-il ne peut donner des résultats généralisables à toute une langue (Rastier, 2005 ; Leech, 2006). Que l'on soit partisan d'une théorie ou de l'autre, il paraît ici évident qu'un corpus aussi réduit et spécifique que le TuDiCol ne peut en aucun cas être représentatif d'une langue ou d'un état de langue. La méthode présentée comme prenant en compte les spécificités du "dialecte tunisien" en réalité prend en compte les spécificités de ce corpus. Elle permet un traitement automatisé de cette situation d'énonciation spécifique plutôt que l'automatisation de la compréhension du tunisien à l'oral.

2.3.4. Subordination du tunisien à l'arabe

La confusion, parfois consciente, entre l'arabe et le tunisien est particulièrement fréquente. Il est vrai que l'exploitation d'outils et de ressources d'une langue proche richement dotée dans le traitement de langues peu dotées pour lesquelles il n'existe pas d'outils et de ressources spécifiques comme c'est le cas pour le tunisien peut être suffisante pour certaines applications. Mais dans ce cas-là, l'idée est d'adapter ces outils et ressources et donc de les modifier pour qu'ils deviennent conformes à la langue cible. Certaines des publications présentées supra proposent de faire le contraire, à savoir adapter les caractéristiques du tunisien pour qu'elles ressemblent le plus à celles de l'arabe. (Boujelbane *et al.*, 2013) parlent même de "forcer" les mots en tunisien à avoir une

racine qui correspond aux patterns morphologiques arabes bien que ce choix s'avère infructueux pour la majorité du lexique étudié qui est à 60% totalement différent de l'arabe. Il aurait peut-être été possible de dépasser cette difficulté si les auteurs s'étaient intéressés de plus près à l'étymologie des mots plutôt que d'essayer de retrouver de l'arabe à toute force quand il n'y en a pas ? (Hamdi *et al.,* 2013) et (Zribi *et al.*, 2013) font un choix similaire en "arabisant" la morphologie du tunisien". Les erreurs dans les résultats obtenus sont de fait le plus souvent dues aux mots "non-arabes", ou "étrangers" qui ne sont pas traitables selon l'approche choisie.

Si (Zribi *et al.*, 2014) se fixent comme objectif d'atteindre un équilibre optimal entre le maintien des spécificités du "dialectal" et l'établissement de conventions basées sur les similarités entre le "littéral" et le "dialectal", lorsqu'il y a inconciliabilité et qu'il faut trancher c'est l'arabe littéral qui "l'emporte" quitte à perdre une caractéristique importante du tunisien. Les phonèmes /g/, /v/ et /p/, spécifiquement tunisiens et non-arabes sont ainsi mis de côté par commodité car il n'existe pas de lettres arabes pour les retranscrire. Or dénaturer ainsi la langue peut poser un problème de lexique et faire perdre des données importantes. D'abord parce que ces phonèmes peuvent être indicateurs de l'origine géographique ou du milieu social du locuteur (Gibson, 2002). Ensuite, parce que leur "équivalent arabe", à savoir les phonèmes /q/, /f/ et /b/ existent aussi dans le système phonétique tunisien. Il s'agit donc de représenter des mots différents selon une même graphie créant potentiellement une ambiguïté impossible à résoudre par la suite. Contrairement à l'affirmation de départ, les spécificités "dialectales" sont traitées en exceptions et ignorées alors qu'il existe une adaptation de l'alphabet arabe pour le tunisien qu'il serait possible d'exploiter : "ڤ" pour le /g/, "ڤ" pour le /v/ et "پ" pour le /p/.

2.3.5. Choix pour la constitution du corpus

Ces travaux se rejoignent aussi dans le choix de constituer un corpus non pas en tunisien directement, mais à partir de la traduction d'un corpus en arabe littéral vers le tunisien avec l'objectif d'aboutir à un corpus qui a "l'apparence" du "dialecte". Les corpus sur lesquels sont appliquées les méthodes proposées sont donc approximativement tunisiens et ce choix soulève de nombreuses interrogations. Quel est l'apport de ces méthodes lorsqu'elles ne sont pas appliquées, ni applicables, sur des corpus dans la langue étudiée ? Les auteurs font le choix de "supposer" et de "présumer" que les deux langues sont quasi identiques sans aucun appui linguistique pour étayer leur propos. Or le tunisien étant suffisamment différent de l'arabe pour nécessiter un traitement

spécifique, pourquoi faire abstraction de ces différences en constituant un corpus dans une langue "artificielle" débarrassée de ses spécificités et de fait ne correspondant en rien à la réalité des productions émises par les locuteurs tunisiens ? Les *a priori* idéologiques subordonnant le tunisien à l'arabe sont ici indéniables alors même que la démarche part du constat qu'il existe un écart entre ces deux langues tel qu'il est nécessaire de les étudier à part.

2.4. Bref compte rendu sur les productions écrites en tunisien

De ces quelques remarques, il ressort que l'influence de la minoration linguistique que nous présentions supra pèse aujourd'hui de tout son poids sur l'état de l'art du traitement automatique du tunisien. Elle met, nous semble-t-il, un frein au développement d'un véritable fonds tunisien pour le TAL. Pour expliquer l'absence de ressources pour le tunisien, les auteurs des travaux que nous avons présenté renvoient tous à la nature exclusivement orale du tunisien, variante basse de la langue arabe. Mais en réalité, il est faux de dire que le tunisien ne se réalise que dans l'oralité dans le cadre restreint des échanges informels du quotidien.

Formé par un groupe d'intellectuels et d'artistes tunisiens à l'entre-deux-guerres, le cercle Taht Essour ("sous les remparts") a doté la Tunisie d'un véritable patrimoine littéraire : contes, nouvelles, récits, pièces de théâtre, chansons. Depuis les années 70 les productions théâtrales se font quasi-exclusivement en tunisien, il en sera de même pour les productions cinématographiques et les fictions télévisuelles. Plus proche de notre époque quelques productions littéraires ont vu le jour : un recueil de proverbes tunisiens (1994), une traduction du Petit Prince de Saint-Exupery (1997) ou encore une transcription du conte traditionnel Ommi Sissi (2013).

En dehors de la sphère littéraire, le tunisien a investi le monde de la communication dès les années 90s avec l'apparition des premiers slogans publicitaires en tunisien. En 2003, l'arrivée de Mosaïque FM, première radio à émettre en tunisien, chamboule le paysage audiovisuel. Depuis, le modèle a été suivi par de nombreuses autres radios et chaînes de télévision. Avec la révolution, le tunisien est devenu la langue privilégiée par la majorité des politiques que ce soit pour les slogans de campagne, dans les discours ou lors de débats. Mais Habib Bourguiba, le père de la nation tunisienne, avait déjà pour habitude de faire ses discours en tunisien dans un style très apprécié des Tunisiens, encore aujourd'hui (Salah, 2012). La nouvelle constitution tunisienne a même été

traduite en tunisien par l'Association tunisienne de droit constitutionnel dans le but de permettre aux Tunisiens de mieux se l'approprier.

Sur un plan individuel, l'arrivée de la téléphonie mobile avec l'usage des SMS puis de l'internet a poussé chacun à reproduire le parler dans l'écrit, que ce soit à travers le système d'écriture *arabizi* ou en utilisant l'alphabet arabe. Les réseaux sociaux, les sites communautaires, les espaces de commentaire des sites de presse, constituent autant de lieux où se déploie l'écriture tunisienne. Mais internet, ce n'est pas uniquement des messages instantanés à l'orthographe approximative ou fantaisiste, les dictionnaires et les cours de tunisien en ligne sont en effet nombreux. Plusieurs blogs, rédigés exclusivement ou en partie en tunisien comme *Bent Trad*, *Chut ! Libres* ou *La Pomme Empoisonnée*, pour ne citer que quelques exemples, proposent également un contenu à l'écriture rigoureuse : poésie, récit, essai, etc.

En TAL l'absence de ressources effectives pour le tunisien ne doit pas être une excuse pour ignorer les ressources potentielles. Si les méthodes développées ambitionnent de pouvoir un jour traiter toutes ces données dans leur diversité, il faudra qu'elles se résolvent à prendre le tunisien comme véritable objet d'étude et donc à sortir de cette "illusion fantasmatique" coupée de toute réalité (Laroussi, 2002) selon laquelle le tunisien, car caractérisé par la même nature immatérielle et fugace que l'oralité dans laquelle on voudrait l'enfermer, doit s'arabiser pour être digne d'intérêt.

Récapitulatif

- La description de la situation linguistique fait débat. La thèse communément admise est celle d'une répartition fonctionnelle de l'arabe et du tunisien selon le modèle diglossique établi par Fergusson.
- D'autres estiment que le tunisien est une langue à part entière à distinguer de l'arabe et considèrent que la coexistence de l'arabe et du tunisien relève du bilinguisme.
- Le modèle diglossique a pour répercussion une forte minoration de la langue tunisienne qui est considérée comme la variété "basse" de l'arabe.
- Il résulte de cette minoration linguistique un manque d'intérêt pour la langue tunisienne dans la recherche et notamment dans le domaine du TAL. Quelques travaux commencent cependant à voir le jour.
- Les ressources et les outils pour le tunisien sont quasi-inexistants, l'approche privilégiée en TAL pour l'instant est de transformer les textes étudiés pour les rapprocher le plus de la langue arabe.
- De nombreuses ressources potentielles existent et offrent une grande variété de contenus textuels en tunisien notamment sur internet.

3. Le corpus

Si les corpus sont manipulés dans les nombreuses disciplines des Sciences du langage, des Sciences humaines et des Lettres, cette pluralité des domaines d'application génère une multitude de types de corpus, et de fait, autant de définitions. Une définition unifiée semble donc difficile et problématique. Nous proposerons d'abord un bref aperçu des critères de définition d'un corpus. Nous confronterons par la suite ces éléments de définition à notre corpus. Enfin nous décrirons de façon détaillée notre corpus ainsi que la procédure suivie pour le constituer.

3.1. Éléments de définition d'un corpus

3.1.1. Définition générale

Même s'il n'existe pas une définition unifiée de ce qu'est un corpus, il y a l'idée commune d'une collection de textes motivée par des critères spécifiques. Pour (Sinclair, 1996), le corpus est une "collection of pieces of language that are selected and ordered according to explicit linguistic criteria". On retrouve la même idée chez (Habert et al. 1997) qui définit le corpus comme "une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques explicites". (Lebart et Salem, 1994) donnent deux définitions au corpus : du point de vue linguistique, c'est "un ensemble limité des éléments (énoncés) sur lesquels se base l'étude d'un phénomène linguistique" ; en lexicométrie, le corpus est défini comme "un ensemble de textes réunis à des fins de comparaison servant de base à une étude quantitative". (Mayaffre, 2002) se réfère aux définitions données par l'*Encyclopaedia Universalis* ou le *Robert* : "Un rassemblement de textes ou une collection de textes regroupés sur la base de travail en vue de les interroger". Et (Rastier, 2005) propose de convenir d'une définition positive : "un corpus est un regroupement structuré de textes intégraux, documentés, éventuellement enrichis par des étiquetages, et rassemblés : (i) de manière théorique réflexive en tenant compte des discours et des genres, et (ii) de manière pratique en vue d'une gamme d'applications".

3.1.2. Le critère de représentativité

En amont de la constitution du corpus, il y a donc un objectif d'étude, "une préoccupation des applications" qui "détermine le choix des textes, mais aussi leur mode de "nettoyage", leur codage, leur étiquetage ; enfin la structuration même du corpus" (Rastier, 2005). Toujours selon (Rastier, 2005) cette structure peut être déterminée par deux conceptions. La première, "documentaire", considère le corpus comme un "échantillon de la langue, un réservoir d'exemples ou

d'attestations". C'est une conception "logico-grammaticale" qui ignore le "caractère textuel" des documents pour ne retenir que leurs "variables globales". On retrouve ce principe "d'échantillonage" par exemple chez (McEnery et Wilson, 2001) pour qui le corpus doit être "carefully sampled to be maximally representative of a language or language variety". Pour d'autres, il est impossible que les résultats issus de l'analyse d'un corpus puissent prétendre à la généralisation. Ainsi, (Leech, 2006) écrit "whatever is found to be true in a corpus, is simply representative in that corpus – and cannot be extended to anything else". Mais faire dépendre le corpus de l'application à laquelle il se destine permet selon (Rastier, 2005) de "dédramatiser les problèmes récurrents de la représentativité et de l'homogénéité". Ainsi, un corpus, sans jamais représenter la langue (" ni la langue fonctionnelle qui fait l'objet de la description linguistique, ni la langue historique, qui comprend l'ensemble des documents disponibles dans une langue"), peut simple être jugé "adéquat ou non à une tâche en fonction de laquelle on déterminer les critères de sa représentativité et de son homogénéité".

3.1.3. Le critère d'homogénéité

Pour ce qui est du critère d'homogénéité, il doit être respecté au sein même du corpus. En effet, selon (Rastier, 2005), "tout regroupement de textes ne mérite pas le nom de corpus". Par exemple, "une banque textuelle peut regrouper des textes numériques de statuts divers" qui, parce que dépourvus de tout critère unificateur, ne peut pas devenir un corpus. On retrouve cette idée chez (Bowker and Pearson, 2002) pour qui il est important de souligner que "a corpus is not simply a random collection of texts, which means that you cannot just start downloading texts haphazardly from the Web and then call you collection a 'corpus'". L'homogénéité dépend selon (Rastier, 2002) de trois variables globales qui doivent être partagées par les textes qui constituent le corpus : "les *discours* (ex. juridique *vs* littéraire *vs* scientifique), le *champ générique* (ex. théâtre, poésie, genres narratifs), le *genre* proprement dit (ex. comédie, roman "sérieux", roman policier, nouvelle, conte, récit de voyage). Le *sous-genre* (ex. roman par lettres) constitue un niveau encore subordonné". Ainsi, le "bon corpus" est d'abord constitué des textes qui partagent le même genre".

3.1.4. La notion de réflexivité

A côté de la conception "grammaticale" du corpus, (Rastier, 2005) identifie une deuxième conception qu'il qualifie de "philologique-herméneutique" en cela qu'elle tient compte des rapports de texte à texte". (Rastier, 2005) rapproche son propos de celui de (Mayaffre, 2002) qui, à travers la notion de "réflexivité", propose "un nouveau parcours de lecture dans lequel l'acte

interprétatif final doit être pressenti dans l'acte originel de la constitution même du corpus". En effet en soulignant les caractéristiques "sérielles" (addition de textes) et "heuristiques" (construction arbitraire qui n'a de sens que par rapport à l'intention du chercheur) du corpus, (Mayaffre, 2002) s'interroge : "comment débuter et arrêter une série ?" et "comment juger que le rassemblement établi est non seulement nécessaire (ou utile) mais suffisant ?" La question de l'interprétation est ici cruciale. Car en effet si l'intention du chercheur est déterminante pour le corpus, la phase d'interprétation nécessite de recourir à des éléments extérieurs au corpus. Le chercheur se trouve ainsi "projeté en-dehors des limites objectives du corpus, dans un-tout subjectif". Or les ressources extérieures mobilisées pour l'interprétation du corpus sont elles-mêmes le plus souvent du texte, c'est-à-dire "de la même nature que le corpus". Pourquoi alors cette "discrimination" entre, d'une part, les textes qui composent le corpus – auxquels sont appliquées des méthodes scientifiques – et, d'autre part, les textes qui forment l'intertexte – auquel le chercheur se réfère sans aucun traitement ?

Pour dépasser cette "tension dialectique" entre "le cocon objectif que constitue l'en-dedans du corpus (...) et le tout subjectif de son en-dehors textuel", (Mayaffre, 2002) propose de constituer de grands corpus réflexifs :

"Nous entendons par réflexivité du corpus le fait que ses constituants (articles de presse, discours politiques, pièces de théâtre ; de manière plus générale, sous-parties) renvoient les uns aux autres pour former un réseau sémantique performant dans un tout (le corpus) cohérent et auto-suffisant."

En intégrant au sein même du corpus et sur un pied d'égalité le texte et son environnement linguistique, ou co-texte, le chercheur n'a plus besoin de sortir du corpus pour comprendre et interpréter ses composants puisqu'ils deviennent analysables de manière contextualisée ou co-textualisée grâce à une navigation interne au corpus. Internaliser autant que possible les ressources sémantiques ou interprétatives co-textuelles implique de réfléchir à ses ressources en amont dès la constitution du corpus plutôt que d'y faire appel de façon intuitive et aléatoire, au fur et à mesure de l'avancement de l'analyse. Le chercheur rend l'acte interprétatif "si ce n'est objectif, en tout cas transparent".

3.1.5. Les séries textuelles chronologiques

Enfin, (Salem 1988 ; Salem 1991) identifie un type particulier de corpus qu'il appelle "séries textuelles chronologiques" défini ainsi :

" Nous appelons 'séries textuelles chronologiques' ces corpus homogènes constitués par des textes produits dans des situations d'énonciation similaires, si possible par un même locuteur (individuel ou collectif) et présentant des caractéristiques lexicométriques comparables."

(Salem 1988 ; Salem 1991) précise que cette définition pourrait, dans l'absolu, s'appliquer à tout corpus dans la mesure où chaque texte qui le compose possède une date de rédaction ou de publication propre donnant au tout une dimension chronologique. Cependant, quand cette dimension chronologique préside de manière évidente à la constitution du corpus, sa prise en compte permet de mettre en évidence des variations du vocabulaire qui surviennent au cours du temps, le "temps lexical".

3.2. Positionnement de notre corpus par rapport à l'état de l'art

3.2.1. Brève présentation de notre corpus

Notre corpus est composé de l'ensemble des commentaires en tunisien postés par des utilisateurs de Facebook sur la page Facebook de la radio tunisienne Mosaïque FM sur une période qui s'étend du 1er janvier 2011 au 31 décembre 2011. Cette période englobe la chute de la dictature avec le départ de l'ancien président Ben Ali (le 14 janvier 2011) ainsi que l'élection d'une Assemblée constituante (23 octobre 2011), première élection démocratique en Tunisie.

3.2.2. Discussion

A la lumière des quelques éléments de définition présentés supra et des problématiques qui en découlent, notre corpus soulève de nombreuses questions. Il semble en effet difficile de trouver des caractéristiques qui le rattachent complètement aux critères évoqués supra. S'il y a clairement un objet d'étude (analyse textométrique de concepts politiques dans des textes informels en tunisien issus des réseaux sociaux), qu'en est-il de la représentativité et de l'homogénéité ? La notion de réflexivité est-elle applicable à ce type de corpus et dans quelle mesure ? Enfin, peut-on parler de série textuelle chronologique dans notre cas ?

3.2.2.1. Représentativité et homogénéité

Les textes qui constituent notre corpus sont produits par un grand nombre d'internautes distincts avec des styles d'écriture différents, dans au moins trois langues (arabe, français, tunisien) selon plusieurs systèmes d'écriture (*arabizi*, alphabet latin, alphabet arabe), parfois au sein d'un même commentaire, ou d'un commentaire à un autre pour une même personne.

Si ces textes rendent compte de la situation linguistique en Tunisie – plurilinguisme et diglossie (Mejri et al. 2009) – et d'une certaine pratique de l'écriture tunisienne sur internet – *arabizi* et multiplicité des systèmes d'écriture – il n'est pas pour autant possible d'affirmer qu'ils remplissent le critère de représentativité de la langue. Tout d'abord, pour l'écriture du tunisien, il n'y a pas de norme codifiée (cf. 2.1), hormis quelques règles de base pour les correspondances entre chiffres et lettres en *arabizi*. Ensuite, la multiplicité des formes est non seulement due à l'absence de contrainte de code (mis à part la nécessité de se faire comprendre) mais aussi à la nature même des textes de notre corpus : l'écriture spontanée de commentaires sur les espaces de conversation virtuels est caractérisée par l'irrégularité (fautes d'orthographe et de langue, absence de signes diacritiques, abréviations, etc.). Or, comme nous l'expliquions précédemment (cf. 2.4) il existe de nombreux domaines où le tunisien est employé avec une écriture rigoureuse (cohérente) et un style élaboré (poésie, théâtre, conte, etc.). Nos textes sont donc surtout représentatifs d'un certain style d'écriture sur les réseaux sociaux.

Par ailleurs, les commentaires qui composent notre corpus ont été publiés par un grand nombre d'internautes. Il n'y a donc pas homogénéité au niveau de l'auteur. Ils présentent également un grand mélange au niveau du style, on peut donc dire qu'il s'agit de "genre conversationnel". Sur le plan de la forme, il y a une hétérogénéité évidente. Pour ce qui est du fond, ces commentaires traitent de sujets variés, orientés par les publications de la page. Leur point commun est cependant qu'ils traitent en très grande majorité de l'actualité socio-politique tunisienne et qu'ils reflètent en cela une partie de l'opinion publique. Une partie, car il s'agit évidemment ici de l'opinion exprimée par des internautes, possédant un compte sur le réseau social Facebook et abonnés à la page Facebook de Mosaïque FM. Notre corpus n'inclut donc pas la population non connectée, par choix ou non ; ni celle connectée mais non sur Facebook ; ni celle connectée, présente sur Facebook mais pas abonnée à la page Facebook de Mosaïque FM. Cette opinion est, par ailleurs, prise dans sa multiplicité, dans sa diversité et dans son évolution depuis la dictature jusqu'aux premières élections démocratiques du pays.

3.2.2.2. Notion de réflexivité

Même si (Mayaffre, 2002) parle de "ligne d'horizon" plus que d'un "objectif atteignable", il serait intéressant de s'interroger sur les moyens d'intégrer la dimension réflexive dans un corpus constitué de textes informels issus des espaces conversationnels sur internet.

Si le contenu textuel reste incontournable pour la majorité de ces espaces, c'est notamment le cas de Facebook, il est fortement concurrencé par d'autres types de contenus. Pour notre corpus par exemple, les commentaires peuvent être postés en réponse à un statut publié par la page, mais aussi à des images, des vidéos, des séquences audio ou même des sondages (une des nombreuses applications proposées sur les pages Facebook).

En se limitant strictement au contenu textuel, il ne faut pas perdre de vue que les échanges entre les internautes, organisés en communautés, s'inscrivent le plus souvent dans la continuité d'une histoire commune qui dépasse largement le cadre restreint de l'élément commenté dans l'immédiat. Les références à des publications ou conversations passées sont alors fréquentes. Les internautes peuvent également faire référence aux "métadonnées" de leurs interlocuteurs ou aux leurs : image de profil, détails biographiques, opinions religieuses ou politiques.

La référence à des éléments extérieurs à la conversation peut dépasser le cadre d'un même site web : les frontières entre les différents réseaux sociaux sont extrêmement poreuses et de nombreuses applications permettent de basculer d'un espace à l'autre ou de centraliser les publications pour une diffusion multi-réseaux. Dans ce cas-là de nombreux internautes peuvent commenter sur un réseau social une publication lue sur un autre, ou publier un même commentaire instantanément sur différentes plateformes. C'est par exemple le cas pour le trio Instagram - Facebook - Twitter. Les internautes peuvent par ailleurs faire référence à d'autres types de sites web en partageant des liens via des sites d'information en ligne par exemple.

Elle peut même dépasser le cadre de l'espace virtuel. Là encore les frontières entre monde virtuel et monde réel ne sont pas étanches et l'interférence entre les deux est constante. Dans le cadre d'un débat autour des opinions politiques des uns et des autres par exemple on peut s'attendre à ce qu'il y ait des références à des discours prononcés par des responsables politiques, à une interview publiée dans un journal ou à un débat dans le cadre d'une émission télévisée. A titre d'exemple nous trouvons dans notre corpus un grand nombre de références à des "rumeurs". Dans ce cas-là les internautes rapportent des "histoires" ou "anecdotes" entendues dans leur quotidien et les diffusent pour que d'autres internautes s'en emparent à leur tour. Cette interférence entre le réel et le virtuel et donc problématique pour la contextualisation des commentaires.

3.2.2.3. La dimension chronologique

Même si notre corpus présente un intérêt chronologique évident, peut-on pour autant parler de série textuelle chronologique ? En effet, le critère d'homogénéité, comme nous l'avons montré, est discutable. Ensuite, la notion de situation d'énonciation soulève là encore des questions. Faut-il considérer la situation d'énonciation de chaque commentaire, de chaque internaute ? Si oui, celle-ci est non seulement différente d'un internaute à un autre mais diffère aussi dans le temps pour un même internaute. Si l'on considère le locuteur comme une entité collective, peut-on dire que la situation reste la même d'un bloc de commentaires (publications de la page + commentaires qui y sont attachés) à un autre sachant que les sujets abordés sont variés ? Enfin, la longueur des commentaires peut aussi varier allant d'un mot à plusieurs paragraphes.

Parce que les sites communautaires et les réseaux sociaux sur internet transposent les codes et les caractéristiques de l'oralité à l'écrit, les textes qui en sont issus et de fait les corpus qui les réunissent, diffèrent par de nombreux aspects des corpus textuels "classiques". Pour cette raison, il ne nous semble pas possible de leur appliquer les critères formulés supra. Nous n'avons pas la prétention d'en formuler de nouveaux ici, nous nous en tiendrons donc aux points de convergence entre ces deux "types" de corpus tout en ayant conscience des limites d'une telle approche.

3.3. Présentation détaillée du corpus

3.3.1. Choix de la période

Notre corpus réuni un ensemble de commentaires de longueurs variées écrits et publiés par des utilisateurs du réseau social Facebook en réaction aux publications de la page officielle sur Facebook de la radio tunisienne Mosaïque FM. Il s'agit donc d'une réunion d'un ensemble de textes produits au cours d'une période de temps que nous avons limitée à une année.

Le choix de la période (de janvier 2011 à décembre 2011) est motivé par le fait que nous souhaitons étudier l'évolution de la prise de parole publique à partir de la révolution tunisienne tout en ayant un point de comparaison entre un "avant" et un "après" Ben Ali. Or c'est en janvier 2011 que le mouvement de révolte a pris une ampleur nationale, aboutissant rapidement au départ de Ben Ali (Lecomte, 2011). De nombreux médias, dont Mosaïque FM ont très vite supprimé la grande majorité de leurs publications datant d'avant 2011 et il n'est plus possible aujourd'hui d'accéder à leurs archives (nous constaté ce fait notamment lors d'un précédent travail dans le cadre d'un cours de M1). Concernant la page Facebook de Mosaïque FM, la plupart des publications qu'on trouve

aujourd'hui sont du type "Bonjour et bonne journée à tous", ce qui ne présente pas un très grand intérêt textométrique. Sur les années suivantes, de nombreux mois ont également été entièrement ou en grande partie supprimés des publications de la page. Nous avons remarqué par exemple que les mois des périodes électorales et post-électorales de 2014 ont été supprimés. Sur notre corpus, on voit clairement que les mois de novembre et de décembre au cours desquels l'actualité était centrée sur les résultats de l'élection de l'assemblée constituante et la formation d'un nouveau gouvernement présente un nombre de publications bien inférieur au reste de l'année.

3.3.2. Langues et systèmes d'écriture

Pour la plus grande part, les commentaires sont rédigés en *arabizi*. L'*arabizi* est un système d'écriture (utilisé aussi bien pour le tunisien que pour les autres langues communément identifiées comme "dialectes arabes") pratiqué depuis les années 90s avec la généralisation de l'usage des téléphones cellulaires et notamment de la communication via les SMS. Les appareils téléphoniques n'étant le plus souvent pas dotés de claviers arabes, les utilisateurs ont transcrit les lettres arabes en lettres latines en compensant les phonèmes qui n'avaient pas d'équivalent dans l'alphabet latin par des chiffres.

Chiffres	**Lettres arabes**
2	ء
3	ع
4	ض
5	خ
6	ط
7	ح
8	غ
9	ق

Tableau 1 Correspondances entre chiffres et lettres arabes

Mais le tunisien peut aussi bien être écrit en *arabizi* qu'en alphabet arabe. Dans ce dernier cas, les utilisateurs ont tendance à appliquer le même principe de l'écriture *arabizi*, mais dans l'autre sens, en écrivant les termes "non-arabes" qu'il s'agisse d'emprunt ou de bilinguisme, en lettre arabes avec une adaptation de l'alphabet arabe pour transcrire les phonèmes tunisiens qui n'existent pas

en arabe : "ڤ" pour le /g/, "ڤ" pour le /v/ et "پ" pour le /p/ (cf. 2.3.4). Parfois, les utilisateurs peuvent mélanger les deux écritures et/ou les deux langues, tunisien et français.

Bien que majoritairement écrit en tunisien, le corpus contient des passages en français ou en arabe littéral. Si nous faisons le choix, du moins à ce stade-là dans le cadre d'une première exploration, de ne pas isoler les différentes langues et/ou les différents systèmes d'écriture c'est qu'il nous semble plus pertinent d'étudier le corpus dans son hétérogénéité étant donné que ce "mélange" correspond à une des caractéristiques de l'écriture sur les réseaux sociaux en Tunisie. Nous pourrons, ultérieurement, constituer des groupes de formes avec Lexico 3 pour comparer l'emploi de termes relatifs à une même notion dans les différentes langues et/ou systèmes d'écriture.

3.3.3. Balisage du corpus

Le corpus est d'abord divisé en 12 parties qui correspondent chacune à un mois de l'année. Le mois de janvier est lui subdivisé en deux sous-parties correspondant à l'avant et à l'après 14 janvier, date du départ du président déchu Ben Ali. Pour isoler de ces deux parties, nous avons ajouté une troisième balise qui englobe le reste du corpus. Enfin, à l'intérieur de chacune de nos parties, nous avons fait le choix de considérer chaque commentaire comme une partie à part en le faisant précéder d'une balise.

Afin d'afficher les mois dans leur ordre chronologique, nous avons fait précéder chaque nom de mois d'une lettre de l'alphabet car cela permet de représenter l'ordre chronologique par l'ordre alphabétique : A_Janvier, B_Fevrier, C_Mars, etc. Les parties du corpus correspondant aux mois sont définies par la balise "mois", les parties du corpus correspondant à une période à l'intérieur du mois sont définies par la balise "partie" et les commentaires sont définis par la balise "commentaire". Dans le contenu de ce dernier type de balises, nous avons utilisé le terme employé pour la partie supérieure qui englobe le groupe de commentaires concernés suivis, chacun, d'une numérotation afin de les distinguer entre eux et de garder une trace du nombre total de commentaires pour chaque partie.

Exemples :
<mois=A_Janvier>
<partie=AA_JanAV14>
<commentaire=JanvierAV_1>

Enfin, nous avons ajouté le caractère "§" à la fin de chaque commentaire afin de délimiter les commentaires et de les séparer les uns des autres. Nous utilisons ainsi un découpage initialement destiné à la délimitation des paragraphes dans un même texte car étant donné la nature du média sur lequel nous travaillons – réseau social où les messages sont brefs et ne respectent que rarement les règles de ponctuation – il n'est pas possible de réaliser des découpages à l'intérieur de chaque commentaire.

3.4. Constitution du corpus

3.4.1. Récupération des données sur Facebook

Pour constituer notre corpus, nous avons écrit un programme en python qui utilise notamment la bibliothèque *Beautiful Soup*[7]. Celle-ci permet de *parser* un document HTML afin de récupérer le contenu de certaines balises spécifiquement.

Le document HTML ici est celui de la page Facebook de Mosaïque FM. Les pages Facebook étant dynamiques, c'est à dire que leur contenu n'est pas directement codé en HTML mais généré dynamiquement, nous avons choisi d'enregistrer en local le code HTML généré par le navigateur web (ici Chrome). Nous avons par la suite extrait le contenu de cette page HTML locale en utilisant la bibliothèque python urllib2 comme si c'était une page web statique.

Une partie du balisage a été automatisée dans le script. Avec *Beautiful Soup*, nous avons d'abord isolé des blocs dans le code HTML, chaque bloc correspondant à une publication de la page suivie des commentaires des internautes. Ensuite, pour chaque commentaire, nous avons ajouté la balise "commentaire", le contenu correspondant au mois (ou à la partie du mois) (que nous avons repéré au moyen d'expressions régulières), et le numéro du commentaire (que nous ajoutons au moyen d'une variable incrémentée à chaque commentaire). Nous obtenons ainsi tous les commentaires de notre corpus, balisés selon le mois de leur publication. Nous avons par la suite, manuellement, ajouté les balises "mois" et "partie" avant chaque premier commentaire (portant le numéro "1") de la partie en question.

3.4.2. Normalisation

Etant donné que nous utilisons une version de Lexico qui ne supporte pas encore l'encodage utf-8 alors que notre corpus contient des lettres latines accentuées et des lettres arabes, nous avons dû

[7] Documentation Beautiful Soup : http://bit.ly/O2H8iD

procéder à quelques normalisations afin de dépasser les incompatibilités entre les deux alphabets et de permettre l'affichage du corpus en windows-1256. Nous avons donc pour cela modifié les lettres accentuées en les faisant précéder de divers caractères afin d'en garder une trace. Par ailleurs, les émoticônes affichées en dessin dans le navigateur sont écrits en toutes lettres dans le code HTML, chaque type d'émotion étant précédé du mot "émoticône". Afin de pouvoir traiter ces éléments, nous avons collé le mot "emoticone" (débarrassé des accents) avec le mot exprimant l'émotion dont il était suivi.

À	#A	à	#a
Â	*A	â	*a
Á	~A	á	~a
Ä	^A	Ä	^a
Ç	#C	ç	#c
É	#E	é	#e
È	~E	è	~e
Ê	^E	ê	^e
Ë	*E	ë	*e
Î	^I	î	^i
Ï	#I	ï	#i
		ì	*i
Í	~I	í	~i
Ô	^O	ô	^o
		ó	*o
Ø	~O	ø	~o
Ö	#O	ö	#o
Ù	#U	ù	#u
		ú	~u
Û	^U	û	^u
Ü	*U	ü	*u
		ş	*s
		ñ	~n
		ß	~b

		¿	~?

Tableau 2 Lettres accentuées et leur équivalent sans accent

émoticône smile	emoticoneSmile	
émoticône grin	emoticoneGrin	
émoticône tongue	emoticoneTongue	
émoticône wink	emoticoneWink	
émoticône heart	emoticoneHeart	
émoticône kiki	emoticoneKiki	
émoticône kiss	emoticoneKiss	
émoticône cry	emoticoneCry	
émoticône unsure	emoticoneUnsure	
émoticône upset	emoticoneUpset	
émoticône frown	emoticoneFrown	
émoticône gasp	emoticoneGasp	
émoticône squint	emoticoneSquint	
émoticône colonthree	emoticoneColonthree	
Emoticône like	emoticoneLike	

Tableau 3 Normalisation des émoticônes

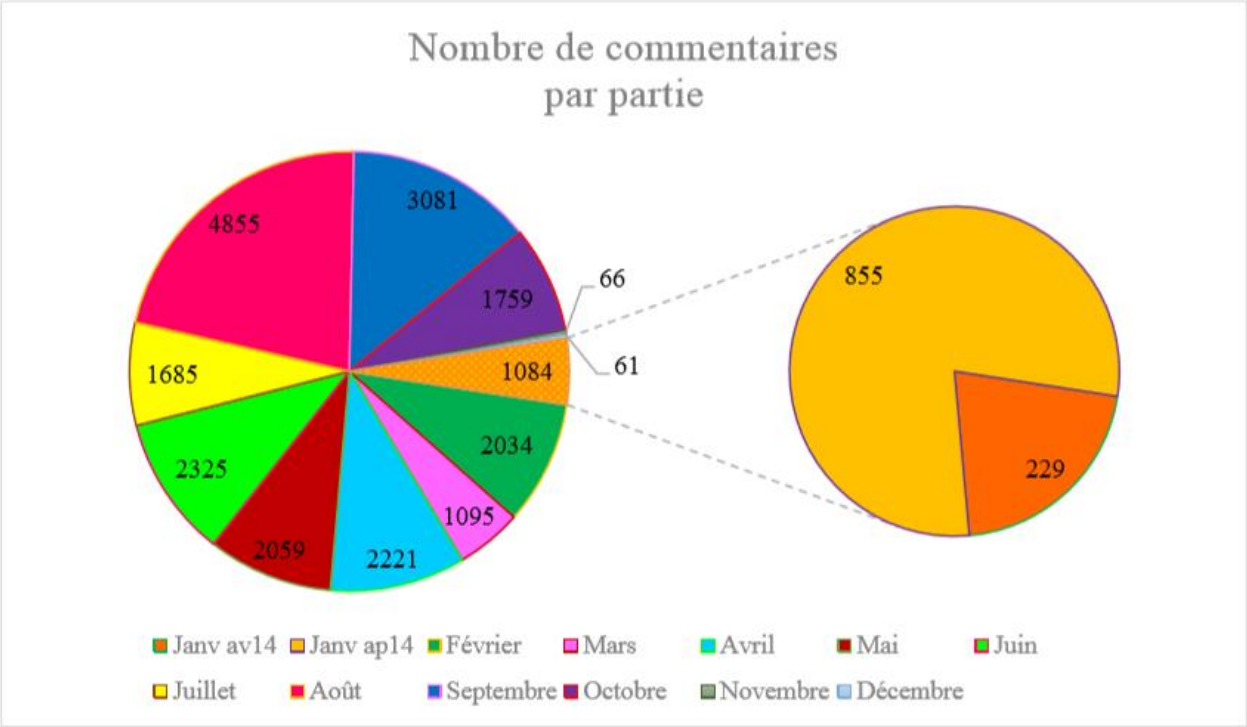
3.5. Observations générales sur le corpus

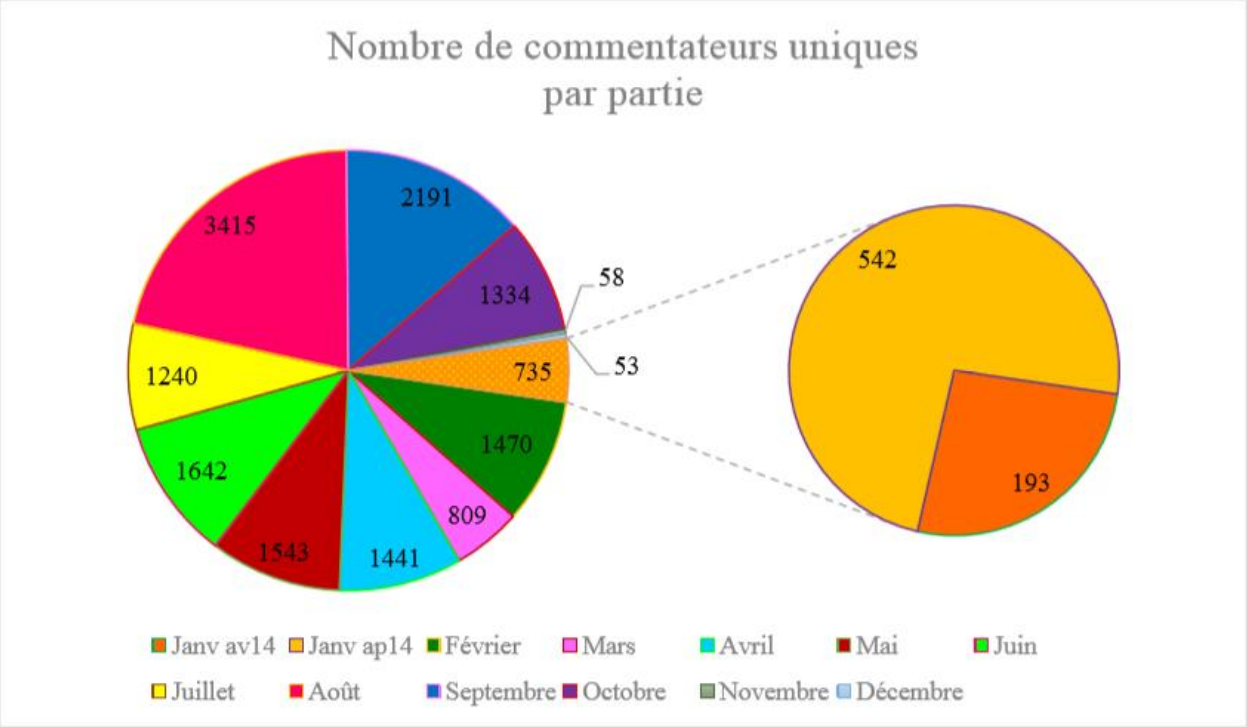
L'observation de notre corpus appelle quelques remarques. Tout d'abord, nous constatons une certaine disparité en nombre de commentaires entre les différents mois. Les mois de décembre et de novembre sont ceux qui présentent le plus grand écart avec le reste des parties puisqu'ils ne contiennent respectivement que 66 et 61 commentaires. En observant le document HTML brut, nous constatons que le nombre de publications de la page pendant ces deux mois est bien inférieur aux autres mois. Nous savons que les médias tunisiens présents sur Facebook ont tendance à supprimer régulièrement certaines publications de leur page. Lors de la constitution du corpus nous avons par exemple constaté que les principaux médias tunisiens ont supprimés de leurs pages Facebook et de leurs sites web l'intégralité ou la majorité de leurs publications datant d'avant le départ de Ben Ali et des périodes électorales ou post-électorales de 2011 et de 2014. Nous ne connaissons cependant pas les motifs de ces pratiques car aucun de ces médias n'a souhaité nous répondre lorsque nous les avons sollicités à ce sujet.

Janvier	1084
Février	2034
Mars	1095
Avril	2221
Mai	2059
Juin	2325
Juillet	1685
Août	4855
Septembre	3081
Octobre	1759
Novembre	66
Décembre	61

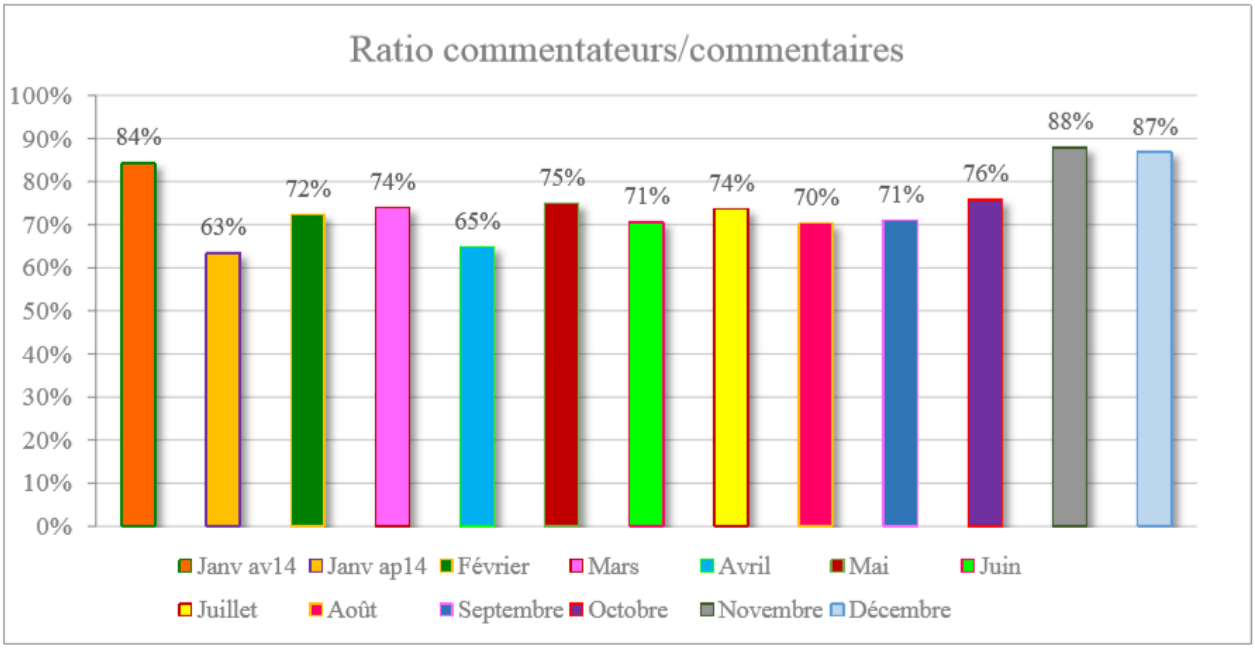
Tableau 4 Nombre de commentaires par mois

Ensuite, outre le nombre de commentaires, nous avons également calculés le nombre de commentateurs uniques. Le calcul a été fait au moyen d'un script en python dans lequel nous avons récupéré pour chaque partie (repérée au moyen d'expressions régulières) non plus le texte du commentaire mais l'URL de l'auteur du commentaire. En effet, contrairement au nom d'utilisateur pour lequel il peut exister de nombreux homonymes, l'URL du profil est toujours unique. Avec ces Url, nous avons constitué une liste dont nous avons calculé la taille hors doublons. Les chiffres obtenus restent cohérents avec le nombre de commentaires.


Nombre de commentaires
par partie
3081
1759
66
61
1084
2034
1095
2221
2059
2325
1685
4855
855
229
Janv av14
Janv ap14
Février
Mars
Avril
Mai
Juin
Juillet
Août
Septembre
Octobre
Novembre
Décembre


Nombre de commentateurs uniques
par partie
2191
1334
58
53
735
1470
809
1441
1543
1642
1240
3415
542
193
Janv av14
Janv ap14
Février
Mars
Avril
Mai
Juin
Juillet
Août
Septembre
Octobre
Novembre
Décembre

Nous avons calculé pour chaque partie la moyenne de commentateurs uniques par rapport au nombre de commentaires. Les résultats montrent que pour tous les mois sauf septembre et octobre la moyenne se situe entre 60% et 70% de commentateurs uniques. Pour les deux derniers mois de l'année, la moyenne est presque à 90% ce qui laisse supposer que la page a effectivement supprimé la majorité de ses publications et ainsi un grand nombre de commentaires potentiellement postés par les mêmes utilisateurs. Pour le mois de janvier, on constate un écart de deux points entre la partie avant le 14 janvier et la partie après le 14 janvier. La partie après le 14 janvier présente par ailleurs le ratio le plus bas de tout le corpus, ce qui pourrait indiquer un plus grand engagement dans des conversations entre les internautes.



Enfin, autre constat intéressant : la présence en quantité non négligeable de caractères ou symboles appartenant à différents alphabets comme "¿", "ş", "ñ", "ß". Ces mélanges peuvent s'expliquer aussi bien par l'absence de norme codifiée pour le tunisien qui n'a pas d'alphabet officiel que par la volonté de "décorer" le texte en recourant à des caractères "fantaisistes" comme il est habituellement d'usage dans les sites communautaires et les réseaux sociaux.

3.6. Unités et parties du corpus

Pour la segmentation du texte, nous avons choisi la liste des délimiteurs proposée par défaut. En effet, les règles orthographiques sont très peu respectées du fait de la nature même du média à partir duquel nous avons constitué notre corpus. Sur Facebook (et plus généralement sur les espaces de

discussion), l'écriture reproduit le plus souvent l'oralité, beaucoup d'éléments orthographiques comme les tirets dans les mots composés ne sont pas utilisés. Il ne nous semblait donc pas nécessaire de complexifier la segmentation surtout qu'il est possible par la suite de constituer des groupes de formes. Les formes liées sont de plus repérables par des répétitions similaires dans le corpus. Il sera par ailleurs important de prendre en considération le caractère parfois agglutinant de la langue tunisienne. En effet, certains éléments comme les articles définis, les conjonctions de coordination et les prépositions sont parfois collés au mot qui en dépend, parfois non, sans règles particulière (le tunisien ne disposant d'une norme orthographique codifiée comme nous l'évoquions supra). Pour traiter ces cas-là, nous privilégierons également l'utilisation des groupes de formes pour regrouper les différentes graphiques d'un mot qu'il soit employé seul ou avec une conjonction de coordination, par exemple.

Le dépouillement de notre corpus en formes graphiques délimitées par la liste des délimiteurs proposée par défaut donne les résultats suivants :

Nombre d'occurrences	245363
Nombre de formes	60856
Nombre d'hapax	43178
Fréquence maximale	3784

On note notamment le nombre élevé de formes qui constituent presque 25% des occurrences ainsi que le nombre élevé d'hapax qui constituent plus de 17.5% des occurrences. A titre de comparaison, en nous référant à (Salem et Fleury, 2009), dans le corpus *Duchn,* le rapport entre nombre de formes et nombre d'occurrences est de 7.84% et le rapport entre nombre d'hapax et nombre d'occurrences est de 3.58%. Dans le corpus *Monde/Insécurité,* le rapport entre nombre de formes et nombre d'occurrences est de 4.31% et le rapport entre nombre d'hapax et nombre d'occurrences est de 1.75%. On constate aussi des écarts non négligeables pour la fréquence maximale : 3784 dans notre corpus (1.54 % du nombre d'occurrences) contre 6130 dans le corpus *Duchn* (4.34% du nombre d'occurrences) et contre 44194 dans le corpus *Monde/Insécurité* (5.09% du nombre d'occurrences). Il y a donc un grand nombre de formes différentes ainsi qu'un grand nombre de formes qui n'apparaissent qu'une seule fois ce qui a pour conséquence de réduire le nombre de formes qui se répètent à l'identique. Ceci reflète la difficulté que constitue le traitement

de notre corpus en raison à la fois de l'absence de norme orthographique pour le tunisien et du contexte de production des textes (écriture peu rigoureuse sur les sites communautaires et les réseaux sociaux).

Récapitulatif

- La définition de ce qu'est un corpus dépend des domaines d'application. Il existe cependant certains critères communs qui définissent ce qu'est un "bon corpus" pour les traitements statistiques.
- Notre corpus ne remplit pas intégralement tous ces critères mais l'essor des réseaux sociaux et des sites communautaires sur internet oblige à repenser ces critères établis pour des corpus "classiques".
- L'hétérogénéité rend compte à la fois des problématiques liées au traitements de corpus de textes issus du web (fautes d'orthographes, abréviations, etc.) ainsi que des problématiques liées au traitement d'un corpus en tunisien (absence de norme codifiée, plurilinguisme et multiplicité des systèmes d'écriture).
- A ce stade et pour une première exploration, nous choisissons d'explorer notre corpus dans son hétérogénéité.

Exploration textométrique du corpus

Née de la rencontre entre la linguistique et la statistique grâce aux avancées technologiques réalisées dans le domaine informatique, la textométrie réunit un ensemble d'approches mathématiques dans le but d'analyser des données textuelles numérisées réunies sous forme de corpus. Ces approches permettent de délinéariser les textes (Maingueneau, 1991) et de mettre ainsi en évidence des régularités imperceptibles pour une lecture non-outillée (Pincemin, 2011). Ou, comme l'exprime (Rastier, 2005), "la textométrie fait partie des techniques capables de plonger dans des dimensions profondes du matériau textuel". Dans cette partie nous proposerons une brève définition de la textométrie. Nous présenterons ensuite les unités nécessaires à l'application des méthodes textométriques. Nous évoquerons sommairement ces méthodes. Enfin, nous proposerons quelques exemples d'applications.

3.7. Brève présentation de la textométrie

Comme son nom l'indique, la textométrie place le texte au centre de ses procédures. Bien qu'essentiellement quantitative, elle intègre "des moyens de parcours et d'interprétation qualitatifs" et "accorde [de fait] une place fondamentale au 'retour au texte'". Le chercheur, à chaque étape de son analyse, est amené à faire des choix, à contrôler et à valider : "la dynamique de l'interprétation procède par ajustement progressif des données des calculs" (Pincemin, 2011). Selon (Söze-Duval, 2011) :

"Le retour systématique au texte s'impose alors comme seule possibilité de contrôle et de validation des différences entrevues au plan statistique. Pour cette raison, les logiciels de textométrie articulent en général plusieurs ensembles de méthodes : certaines sont destinées à produire des synthèses statistiques ; les autres sont mobilisables pour obtenir des restitutions du contexte, organisées autour des points saillants du texte mises en évidence par les premières."

Le terme "textométrie" est issu d'une évolution du terme "lexicométrie". Cette évolution exprime l'idée que l'analyse du corpus ne se limite plus à l'étude du lexique mais s'élargit pour investir l'ensemble du texte (Pincemin, 2011). De fait, l'analyse textométrique s'intéresse à tous les éléments qui composent le matériau textuel. Le texte y est vu comme un ensemble structuré par des contraintes "qui ne ressortent pas au système linguistique mais aux positionnements de ses énonciateurs" (Maingueneau, 1991). Il est donc nécessaire de réfléchir à cette structure et aux

éléments qui la composent. Pour (Söze-Duval, 2011), "la démarche textométrique repose sur l'hypothèse, vérifiée à partir de très nombreuses expériences, que pour comparer les différentes parties d'un ensemble de textes, que l'on peut considérer comme autant de *contenants*, il est utile d'observer, au sein de ces textes, les variations de fréquence de systèmes d'unités textuelles : lexèmes, graphèmes, etc., que l'on peut considérer comme des *contenus*."

3.8. Segmentation et découpage des textes

Les "contenus" ou "unités" textuelles résultent de la fragmentation du texte. Cette opération permet de découper ou "d'émietter" le texte en unités minimales sur la base d'un ensemble de caractères délimiteurs choisis parmi l'ensemble des caractères du textes (Lebart et Salem, 1994). A cette opération s'ajoutent l'annotation (chaque unité se voit affecté une étiquette, par exemple une catégorie grammaticale) et le typage (regroupement sous un même type d'un ensemble d'unités selon des propriétés communes). Le texte résultant de la procédure de segmentations est une suite d'unités isolées qui permettent de reconstituer le texte de départ potentiellement enrichi de nouvelles informations apportées par les étiquettes et les types (Söze-Duval, 2011). C'est sur ce texte segmenté que sont appliquées les mesures quantitatives.

Les "contenants" ou "partitions" du texte résultent du découpage du corpus en différentes parties. Ces parties constituent des cadres, des repères, dans lesquels sera analysée la répartition ou la distribution des unités textuelles. La partition du corpus dépend aussi des besoins et des objectifs de l'analyse et résulte d'un processus de réflexion en amont. Par exemple, pour une analyse diachronique, la partition du corpus doit restituer les informations chronologiques. En plus de cette délimitation des textes en "zones", il est possible de rendre compte de l'organisation spatiale du texte en opérant une segmentation selon cette organisation (paragraphes, chapitres, livres, volumes, etc.). Le découpage du corpus peut combiner les deux approches.

3.9. Méthodes textométriques

Contenus (types) et contenants (zones) sont les éléments sur lesquels s'appliquent les opérations textométriques. (Söze-Duval, 2011) propose une classification des méthodes sur lesquelles reposent ces opérations :

Méthodes zones-types : calcule les types les plus caractéristiques d'une zone donnée (ex. calcul des spécificités qui permet de déterminer si une forme et en sur-emploi ou en sous-emploi dans une partie donnée du corpus).

Méthodes types-zones : identifie les zones dans lesquelles un type est en sur-emploi ou en sous-emploi (ex. ventilation des formes dans les partitions, carte des sections).

Méthodes types-types : fait ressortir des relations entre types (ex. les segments répétés, deux formes ou plus qui se répètent à l'identique à différents endroits du corpus ; les cooccurrences, formes associées de façon récurrentes dans le corpus sans pour autant appartenir à une structure unique).

Méthodes zones-zones : localise, pour une zone donnée, les zones avec lesquelles elle entre en relation (ex. analyse factorielle des correspondances qui permet d'évaluer la distance entre deux parties sur la base de leurs distributions de types).

3.10. Exemples d'applications

Les domaines d'applications des méthodes textométriques sont nombreux. Si à ses débuts, la textométrie s'est fait une spécialité du traitement des questions ouvertes dans les enquêtes, elle compte aujourd'hui des utilisateurs dans des domaines aussi variés que la politique, la littérature ou la philologie. L'objectif étant pour les chercheurs de se doter de techniques permettant de renouveler la lecture de leurs corpus (Pincemin, 2011). L'essor des réseaux sociaux et des sites communautaires sur internet fait apparaître de nouveaux types de corpus et offrent de nouveaux lieux d'exploration pour les méthodes textométriques notamment pour l'analyse de la subjectivité. Nous proposons ici trois exemples d'études sur la base des types de sites dont les corpus ont été extraits.

3.10.1. Témoignages dans les forums

(Eensoo et Valette, 2012) montrent les apports de la textométrie à l'analyse des sentiments sur un corpus de témoignages issus de forums de discussion. La méthode proposée, combinant analyse textométrique et classification automatique basée sur l'apprentissage supervisée, permet de pallier le manque de performances des méthodes de classification pour les tâches d'analyse de la subjectivité. La difficulté réside dans le fait qu'il existe plusieurs niveaux de description pour la subjectivité. Celle-ci ne relève pas seulement du lexique mais peut s'exprimer dans l'organisation temporelle du récit, la structure argumentative, etc. Les méthodes textométriques permettent de prendre en compte ces niveaux de description dans leur multiplicité et donc de les implémenter pour améliorer les résultats de la classification automatique.

3.10.2. Commentaires d'articles de presse

(Eensoo et Valette, 2014) proposent, dans la continuité de (Eensoo et Valette, 2012), une méthode pour caractériser les opinions exprimées dans des commentaires d'articles de presse sur le web au sujet de la communauté Rom en France. Cette étude part du constat que la communauté Rom souffre d'une image fortement négative dans l'opinion française. Une hostilité partagée par toutes les sensibilités politiques (gauche et droite confondues) basée sur des stéréotypes racistes courants comme l'insécurité et la délinquance. La méthodologie s'inscrit dans le cadre de la classification de documents. Il s'agit d'une méthode hybride combinant l'analyse linguistique et l'apprentissage automatique. Les critères linguistiques de classification reposent ici sur une grille d'analyse contrôlée d'un point de vue théorique en articulant la textométrie et la sémantique textuelle. L'analyse du corpus et l'extraction des critères ont été effectuées avec Lexico 3. Cette étude montre qu'il est possible de caractériser le discours évaluatif sur la base d'un ensemble de marqueurs sémantiques dont le sens inhérent n'est pas évaluatif et qui sont sélectionnés grâces aux méthodes textométriques.

3.10.3. Posts sur une page Facebook

(Ducos et al., 2013) utilisent le logiciel Iramuteq (P. Ratinaud) pour réaliser une analyse lexicométrique sur un corpus hétérogène constitué du contenu extrait d'une page Facebook (ici, la page "Soutien au bijoutier de Nice" en référence à un fait divers). L'extraction des données s'est faite au moyen de l'API de Facebook qui permet notamment d'obtenir les commentaires associés à chaque publication de la page. L'écriture sur Facebook étant caractérisée par une grande hétérogénéité dans la graphie (fautes d'orthographes, de syntaxes, etc.), les auteurs font le choix d'en rétablir l'homogénéité (Ratinaud et Marchand, 2011). Ce choix est argumenté par le fait qu'une trop grande variabilité des formes rend toute analyse automatique du corpus peu stable. Le spam est aussi évité en supprimant les URL et certains termes jugés indésirables. Le corpus corrigé améliore les possibilités d'analyse en regroupant en une seule graphie des formes disparates. Ceci permet de classifier les textes du corpus sur la base du lexique et d'analyser le profil de chaque classe.

3.11. Exploration textométrique du corpus

3.11.1. Nouvelles formes : nouveaux thèmes et/ou nouveaux commentateurs

L'outil accroissement de vocabulaire de Lexico3 permet d'observer la courbe du vocabulaire du corpus. La caractéristique globale de cette courbe est de croître rapidement au début du corpus pour finir par ralentir au fur et à mesure qu'on avance. On constate cependant des différences entre les parties, ainsi que des zones d'accroissement plus fort ou de stagnation dans une même partie.

Pour la partition en mois, la courbe connaît au tout début du mois de février un pic puis prend directement après une forme concave avant de croître à nouveau. La courbe connaît également un ralentissement à la fin du mois d'avril. Les parties dans lequel il y a le plus grand nombre de formes nouvelles sont du mois de février, du mois d'avril et août. C'est ce que confirme l'observation de la courbe de chaque partie isolée du corpus entier. On constate que la courbe la plus développée est celle du mois d'avril suivie de celle de février qui présente une partie concave entre deux pics. La courbe du mois d'août connaît d'abord une croissance supérieure à celle du mois de février avant de passer en dessous. Il est intéressant de noter que la page Facebook de Mosaïque FM a connu une attaque de spam (cf. 4.8.1.3) lors des mois de janvier et d'avril.

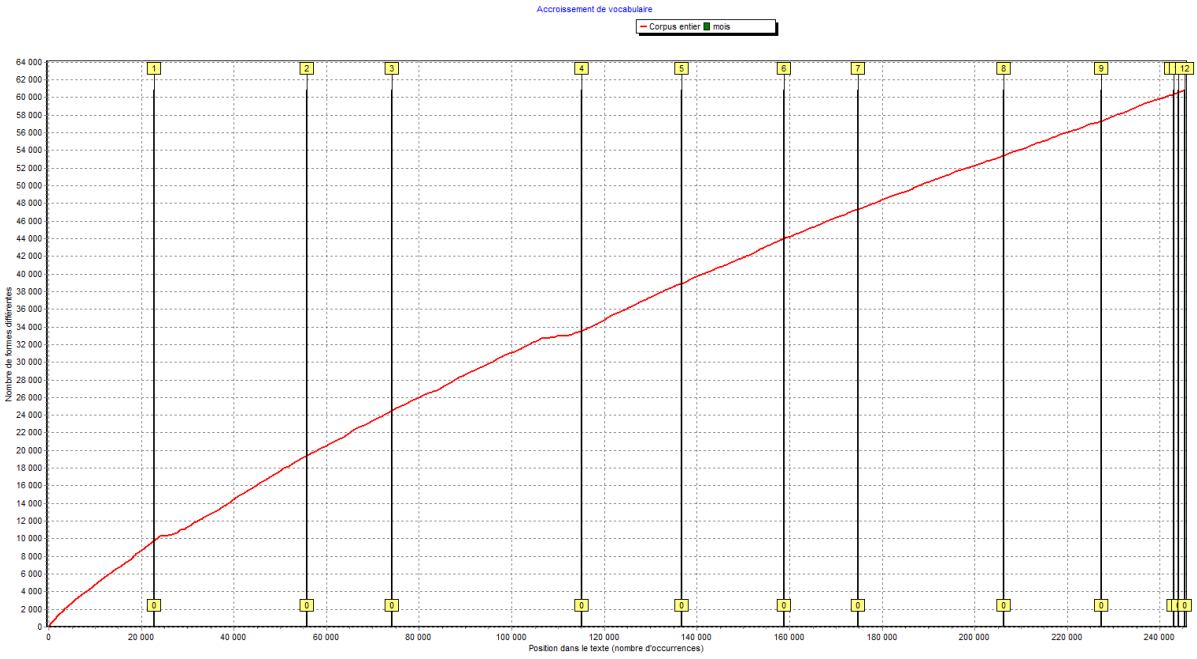


Figure 1 Accroissement du vocabulaire par mois

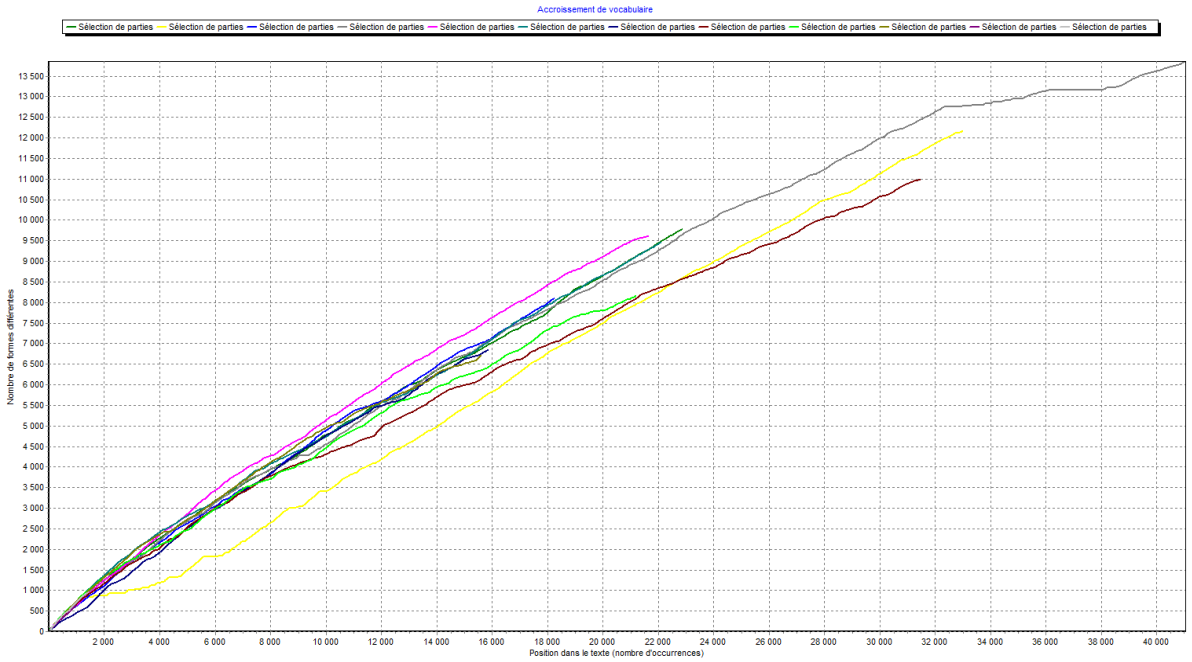


Figure 2 Accroissement du vocabulaire pour chaque mois

Pour le mois de janvier, on constate que la courbe de la partie avant-14 est minime par rapport à la courbe de la partie après-14 qui elle est quasi identique à la courbe de tout le mois de janvier. Ce qui reste cohérent avec le déséquilibre entre les deux parties. Lorsqu'on on observe la courbe par partie, on constate qu'elle est presque à la verticale ce qui indique que les nouvelles formes sont apparues rapidement et en grand nombre.

Figure 3 Accroissement du vocabulaire pour le mois de janvier

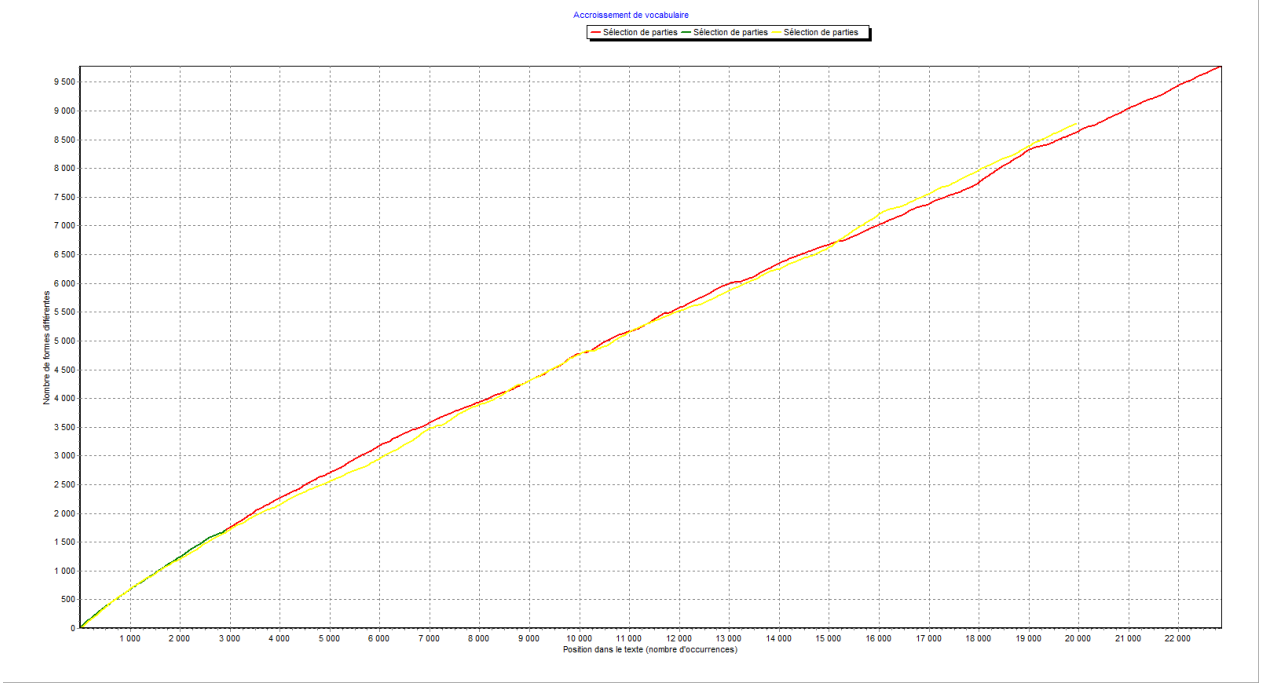


Figure 4 Accroissement du vocabulaire pour les parties du mois de janvier

Dans le diagramme de Pareto, on constate que le mois qui contient le plus grand nombre de formes dont la fréquence est la plus élevée est le mois d'avril, suivi des mois de février et d'août. Pour le

mois de janvier, là encore, la courbe de la deuxième partie du mois est presque identique à celle du mois entier tandis que celle de la première partie du mois est bien inférieure.

En se référant au ratio commentateurs/commentaires calculé plus haut, on voit que le mois d'avril est le mois où le nombre de commentateurs uniques est le plus bas. Le mois d'avril semble donc avoir été particulièrement riche en débats : de nombreux commentateurs se sont exprimés à plusieurs reprises mais ont employé un vocabulaire varié ou des graphies variées pour les mêmes termes.

3.11.2. Deux périodes distinctes : politique et non politique

Le calcul des spécificités, à partir du nombre d'occurrences d'une forme donnée dans une partie donnée en prenant en compte la fréquence totale de la forme, la taille de la partie et l'effectif total, permet de dégager un modèle qui prédit les effectifs pour chaque partie du corpus. On parle de spécificités positives lorsque les effectifs observés dépassent les prévisions que l'on pouvait faire selon ce modèle et de spécificités négatives lorsque les effectifs observés sont inférieurs à ces prévisions.

Sur notre corpus, le calcul des spécificités rend d'abord compte du fait que les termes qui apparaissent en spécificités positives sont assez représentatifs des principaux événements qui ont marqué chaque mois. Ainsi même si notre corpus se limite aux commentaires postés sous les publications de Mosaïque FM, il reste néanmoins assez représentatif de l'actualité tunisienne de l'année 2011, sans pour autant l'englober entièrement.

On note également deux périodes dans l'année : de janvier à mai où les termes qui apparaissent en premier dans les spécifiés positives réfèrent à l'actualité politique et de juin à octobre où les spécificités positives réfèrent à des événements non politiques. Les deux derniers mois de l'année sont plutôt mélangés et leur interprétation pause problème étant donné que la majorité des publications et des commentaires ont été supprimés.

Pour la première partie de l'année on retrouve donc des références à "Ghannouchi", "Ben Ali", "Trabelsi", ainsi que des formes issues des attaques de spam que nous évoquions plus haut. Pour la deuxième partie de l'année on retrouve plutôt des émoticônes, des formules de salutation, ainsi que des références à l'actualité sportive, à la fête des pères ou encore au baccalauréat.

- Janvier

Dans les spécificités de la première partie du mois de janvier, on retrouve de nombreuses références à dieu comme “Hamdoullah” (dieu merci), “rabbi” (mon dieu), “inchallah” (si dieu le veut), qui sont autant de formules communément utilisées en Tunisie dans des propos qui se veulent rassurant ou optimistes lorsque l’on fait face à une situation difficile ou pour indiquer que l’on “remercie dieu malgré tout” (dans l’espoir que la situation n’empire pas, l’idée étant que même si ça va mal, on remercie quand même dieu car les choses pourraient être pire.). Plus loin on retrouve “compris” et “discours” (en référence au “je vous ai compris” du dernier discours prononcé par Ben Ali avant son départ), “président” (qui fait référence au président Ben Ali), “couvre” (qui ici fait référence au couvre-feu instauré dans de nombreuses régions).

Dans la deuxième partie du mois, outre de nombreuses références à Ben Ali et à sa belle-famille, Trabelsi, on retrouve des références à “Rached Ghannouchi”, le président du parti islamiste Ennahdha qui est retourné en Tunisie après plus de 20 ans d’exil à Londres. On ne trouve sur la page Facebook de Mosaïque FM qu’une seule publication en rapport avec cet événement, mais celle-ci réunit un très grand nombre de commentaires. Le débat est déjà très vif entre ses partisans et ses opposants.

eli mouch 3ajbou elghannouchi ynajem ytiiiiiiiiiiir mel bled w ya5let 3la sidou ben ali fi jeddah 5ater eli 9oulou ghannouch mojrem w terroriste houma jme3et el tajamo3 eli tounes kemla 3arfethom 3la 79i9ethom eli houma elmojrmine w el erhebiyne w thaherli mouch elghannouchi eli 9tal 100 we7ed w mouch houya eli b3ath 9anasa to9tel fel 3bed !!! w zid ardougan 9a3ed yo7kom fi torkiya b2afkar el ghanouchi za3ma ardougan erhebi ???!!! fi9ou ya bheyem ya bgar 9a3dine t9oulou fi klem mta3 etajamo3 el kleb
➢ Celui qui n’aime pas Ghannouchi peut s’envoler loin du pays et rejoindre son maître ben ali à Jeddah parce que ceux qui disent que Ghannouchi est un criminel et un terroriste c’est la bande du RCD, maintenant toute la Tunisie les a vu sous leur vrai jour (...)

De nombreux commentaires expriment un certain tiraillement entre passé et avenir qu’on retrouve dans les termes “tarja3” (tu/elle retournes/e) et “narj3ou” (“nous retournons”). Dans “période s3iba tit3adda w tarja3 la Tunisie 5ir milli kénit” (“une période difficile qui passera et la Tunisie

redeviendra mieux qu'elle n'était"), le verbe "tarja3" exprime le retour à un état passé mais le syntagme "5ir milli kénit" ("mieux qu'elle n'était") exprime lui une comparaison qui valorise l'avenir par rapport au passé. D'autres commentaires expriment cette idée : "tet7assen bledna w tarja3 kifma kénit" ("notre pays s'améliore et redevient comme avant"), "nkounou yed wa7da bch tarja3 tounes kifma 9bal w 5ir bmalyoun mara" ("soyons main dans la main pour que la Tunisie redevienne comme avant et un million de fois mieux"), "8dwa 5ir min lyoum w touns tarja3 kif inwara" ("demain sera mieux qu'aujourd'hui et la Tunisie reviendra comme une fleur"). Dans le même ordre d'idée on retrouve aussi bien "جديد" ("nouveau") – dans "وجه جديد" ("nouveau visage"), "دستور جديد" ("nouvelle constitution"), "نقومو البلاد من جديد لازمنا نحاولو" ("nous devons essayer de relever le pays à nouveau") – que "سابق" – dans "النظام السابق" ("l'ancien régime"), "بقايا النظام السابق" ("rebuts de l'ancien régime"), "وجوه النظام السابق" ("les visages de l'ancien régime"), " وزير الداخلية السابق" ("l'ancien ministre de l'Intérieur").

On pourrait trouver dans le rejet de l'ancien régime et des figures de l'ancien régime d'une part, et l'engouement pour le renouveau qui s'annonce d'autre part, des pistes pour expliquer que les références à Ben Ali et à sa belle-famille sont en sur-emploi au mois de janvier mais disparaissent assez rapidement des conversations (cf. 4.8.1.4).

- Février

Les premiers termes qui apparaissent dans les spécificités positives sont issus d'un même message posté à répétition.

> L'appel à l'attaque du ministère de l'intérieur survenu le 26 février 2011 etait commandité par ses ADMINS des jeunes ont payés leurs vies à causes de cet acte inadmissible , ces ADMINS doivent comparaitre devant la justice tunisienne pour incitation à la haine, meme s'ils sont des résidents à l'étranger nous nous organiseront pour qu'ils soient éxtradiés et jugés pour la barbarie qu'ils ont causés. ralliez-vous à notre cause .

Ce message fait référence à un sit-in organisé à la Kasbah de Tunis pour demander la démission de Mohammed Ghannouchi, premier ministre du gouvernement de transition formé après le départ de Ben Ali. Ce sit-in a été marqué par des violences qui ont fait trois morts. De nombreuses personnes ont mis en cause le groupe de cyberdissidents Takriz qui fait partie des organisateurs de ce sit-in. On retrouve aussi dans les spécificités positives "9asba" et "القصبة" (Kasbah).

Les termes appartenant à un champ lexical patriotique ou révolutionnaire sont également nombreux. Le mois de février semble marqué par une certaine prise de conscience révolutionnaire et politique. Les termes exprimant le refus, l'opposition et le rejet sont également en sur-emploi : "ظد" (contre), "لا" (non), "démission", "DEGAGE". Ces spécificités se recoupent avec les cooccurrents des formes relatives à la révolution (cf. 4.8.1.4).

Le mot "فرحة" (joie) est majoritairement employé non pas au sens propre mais sarcastiquement dans un message répété à plusieurs reprises.

<commentaire=Fevrier_152>فرحوا . . . غنّوا دقّوا الطبول وارقصوا أخرجوا الى الشوارع النساء الرجال الشباب اللصوص المخربين المجرمين الكل يخرج الكل يغني ويزغرد الكل يمرح يشرب الخمر في الشارع يغتصب الفتيات يسرق ما يشاء . . . فرح#e . . . فرح#e . . . فرح#e . . . . افرحوا فنحن بلا دول#e و . . . . . قريبا جدًا سنصبح بلا وطن عاش المخربين عاش المساطيل عاشت الكلوشارات عاشت البانديّ#e التي ستقرض سياساتها بالقو#e . . . فلا جند اليوم ولا شرطة#e افعل ما تشاء فأنت حر . . . دام عزّكم أيّها المواطنون الّذين لا وطن لكم ولا حماي#e لكم اليوم سنحيّش في حكم الغاب وسنحتكم للبوني#e الصحيح#e . . . شكرا شباب تونس يا من حقّقتم الثّور#e فركب عليها البانديّ#e يهدّدون هذا ويحكمون على ذاك بالموت . . . فاليوم لا يجب أن تفكّر لأنّك إن فكّرت ستتهم بالتجمّع اليوم سيغيب العقل وتتقدّم البوني#e . . . اليوم سنسير في مسيرات ينظّمها معاقي زجاجات الخمر في شارع الحبيب بورقيب#e وسنمشي وراءهم طوعا أو كرها . . . ستحكمنا الكلوشارات التي تقوّت قبضتها بعد أن شوّهنا نحن صور#e الأمن . . . . . شكرا أيّها المثقفون فقد خدمتم تونس أفضل خدم#e لقد طهرتموها من النظام . . . . عاش الفساد وعاشت الديكتاتوريّ#e §

- Réjouissez-vous… chantez, jouez du tambour, dansez, sortez dans les rues, les femmes, les hommes, les jeunes, les voleurs, les vandales, tout le monde sort, tout le monde chante et pousse des youyous, tout le monde s'amuse, boit du vin dans la rue, viole les jeunes filles, vole ce qu'il veut... joie… joie… joie (…)

- Mars

La principale thématique du mois est le débat sur la notion de "laïcité" outre des formes rattachées à ce terme on retrouve dans les spécificités positives du mois de nombreuses références à la religion "musulman", "islem" ("Islam"). Ce sont les formes que l'on retrouve également dans les cooccurrents des groupes de formes rattaché à

La forme "nous" a un score de spécificité de 6 alors que la forme "vous" a un score de spécificité de 3. En se référant aux concordances on voit qu'un grand nombre de commentaires opposent le "nous" et le "vous" en employant notamment des structures négatives qui expriment le dédain ("rien ne vous empêche", "personne ne vous oblige", "ne vous inquiétez pas", "ne vous prenez pas pour") ou la colère ("nous ne vous autorisons pas", "l'histoire ne vous pardonnera jamais").

On trouve aussi de nombreuses références à l'ancien régime : "RCD" et "التجمع" (pour Rassemblement constitutionnel démocratique, le parti de l'ancien Président Ben Ali), "leila" (le prénom de l'ancienne Première Dame), trabelsia (Trablesi), "voleurs", (employé dans les commentaires en référence aux "corrompus de l'ancien régime" ou "ceux qui ont profité du système" mais de façon vague sans désigner des personnes en particulier). Il est intéressant de voir que "معارضة" (opposition) est employé pour désigner tous les politiques ou activistes par opposition au RCD (ancien parti de Ben Ali). Ce parti perd pourtant son rôle de parti gouvernemental dès janvier et ses activités sont suspendues dès début février. Ce terme désigne donc de façon tardive les opposants historiques à Ben Ali ou rend compte de la logique démocratique de "gouvernement Vs opposition" dans laquelle se plaçaient déjà les internautes bien que le contexte soit encore très confus.

- Avril

Les spécificités positives indiquent une nouvelle attaque de spam. Le message " يا إعلام يا حقير أخرج شوف الجماهير" ("médias minables sortez et regardez le peuple") est posté en réaction au terme "anti-laïcité" employé par les journalistes de la radio pour décrire une manifestation organisée par les islamistes. D'autres commentateurs, sans reprendre cette phrase, manifestent leur mécontentement et accusent la radio d'être baisée dans son traitement de l'information. On voit notamment apparaître la formule "إعلام العار" (médias de la honte). Il s'agit donc ici d'un prolongement du débat sur la laïcité qui a débuté au mois de mars,

Le terme "الشيخ" (le Cheikh) est employé en référence à Rached Ghannouchi dans le cadre des débats sur la laïcité. Selon les concordances et la carte des sections, cette forme n'apparaît qu'une seule autre fois dans le reste du corpus (pour parler d'Abdelfattah Mourou, autre grande figure du parti Ennahdha) alors qu'elle apparaît dix fois dans le mois d'avril. Les commentaires dans lesquels apparaît cette forme sont hostiles à la laïcité. Ils évoquent notamment un débat public entre Rached Ghannouchi et une universitaire sur le sujet de la laïcité. Les termes employés pour décrire le discours de l'universitaire sont très négatifs : "ضعيف" (faible) et "متحامل" (partial) "كذب" (mensonge), "حقد" (malveillance, rancœur). Mais même en dehors des commentaires dans lesquels apparaissent la forme "الشيخ", les opinions exprimées sont très majoritairement défavorables voire très hostiles à la laïcité. Il n'y a pas réellement de débat sur la page, mais plutôt une sorte de débat

en différé entre les commentateurs d'un côté et l'actualité publiée sur la page de l'autre (interview, articles, etc.)

On voit apparaître clairement une opposition entre "العلمانية" (la laïcité) et "النهضة " (Ennahdha). La forme "contre" est beaucoup employée dans des commentaires pro-Ennahdha : "contre l'islam", "contre la laicité", "contre al islem wil 9oren" ("contre l'Islam et le Coran"), "propagande contre Ennahdha", "acharnement contre Ennahdha", "contre hel 7aywanet hedhy eli t7ib tjaridna min huwiyitna el islamia" ("contre ces animaux qui veulent nous dépouiller de notre identité musulmane"). On trouve dans une moindre mesure : "contre ces salauds" (en référence aux "terroristes"), "contre les Obscurantistes Islamistes", "contre les intégristes". Enfin on trouve aussi "contre l'UGTT", "contre la Vie Chère", "contre l'exclusion et la pauvreté".

Pour la forme "démocratie", contrairement aux autres mois où les commentaires appelaient surtout à "défendre" la démocratie, on trouve de nombreux commentaires défaitistes.

(…) On est encore tres loin de la democratie pour avoir un Minstere de l interieur honnete et qui applique la democatie comme si Rajehi (…)
(…) Et parceque malheureusement il avere que 50% de peuple Tunisien ne sont pas ni civilises ni habitues au democratie (…)
ca fé mal au coeur ces commentaires d'insultes!!!!!!!!! c'est ca la democratie, la liberté!!!! vous montrez bien que vous ne meritez pas !!!!
(…) es ce notre chaine tv et radios sont pret pour la democratie? ca se discute

"mra" (femme), "femme" et "femmes" sont des formes particulièrement présentent dans les commentaires du mois d'avril. Les commentaires sont cependant assez disparates aussi bien dans le sujet que dans l'opinion véhiculée. On y trouve des références aux propositions d'Ennahdha d'instaurer la polygamie ou d'interdire aux femmes de travailler pour combattre le chômage. On y trouve également des références à la policière qui a giflé Mohamed Bouazizi (le vendeur ambulant dont le suicide est à l'origine des émeutes qui finiront par faire tomber Ben Ali) et qui avait été acquittée.

- Mai

On voit apparaître au mois de mai la seule référence à une actualité internationale qui est l'annonce du décès d'Oussama Ben Laden. La majorité des commentaires émettent des doutes sur la version officielle et considèrent que Ben Laden fait partie d'une conspiration américaine.

ben laden made in usa
ben laden est la grande mythe que l'USA a cree pour faire ces gaire ds le monde au nom de lutte contre le terrorisme !
Hedhi kedba mil american 5ater ben laden hiya san3eto
➢ Ça c'est un mensonge des Américains car ils ont créé Ben Laden.

Les autres formes qui apparaissent dans les spécificités positives font référence à un débat sur la censure des sites web pornographiques. Beaucoup d'internautes expriment leur crainte d'un retour de la censure.

le probleme c pas le porno : c le fait kils peuvent dorénavant censurer n'importe koi en te donnant un pretexte débile

D'autres soutiennent la censure au nom de la religion.

brabbi barra lawej 3al porno b3id 3alina a7na bled meslma
➢ S'il te plait va chercher du porno ailleurs, nous sommes un pays musulman.

D'autres enfin estiment que la censure est inutile car inefficace.

protéger les ptits enfants looooooooool, g vu mon premier porno a lage de 8 ans sur la télé, le probleme c pas le net, c le CONTROLE PARENTAL !!!!

La forme “hamma” fait référence à Hamma Hammami, homme politique tunisien de la gauche radicale. Les commentaires sont hostiles en majorité et accusent Hamma Hammami et Chokri Belaïd (homme politique de gauche et avocat tunisien, assassiné en 2012) d’être “responsables” de la situation actuelle en Tunisie

balid t hamma sont les premiers responsable de ce qui se passe en ce moment c facile de critiquer montrez nous ce que vous pouvez faire pour sauver le pays
ana chlaghmi yerkez 3liha ettir w hamma el hammami w chokri bel3id ma ymathlounich
➢ Les oiseaux peuvent se poser sur ma moustache et Hamma Hammami et Chokri Belaïd ne me représentent pas.
ils accusent le gouvernement alors qu’ils sont les premiers responsables
bel3id et hamma 2 tirants épouvantables
➢ Belaïd et Hamma deux tirants épouvantables.

Enfin, les formes “farhat” et “الراجحي” (Rajhi) font référence à Farhat Rajhi, ministre de l’Intérieur du 27 janvier au 28 mars. Très populaire, il est surnommé Monsieur propre pour une série de mesures visant à assainir le ministère de l’Intérieur, il est vivement critiqué en mai pour une vidéo dans laquelle il développe des théories conspirationnistes comme la menace d’un coup d’Etat par l’armée en cas d’arrivée des islamistes au pouvoir. La majorité des commentaires sont des témoignages de soutien.

ahna nsad9ou si farhat w koulna m3ah
➢ Nous, nous croyons Monsieur Farhat et nous sommes tous avec lui.
سنختار الراجحي لرئاسة الحكومة

➢ Nous allons choisir Rajhi pour être le chef du gouvernement.
nous tous avec toi farhat w rabi m3ak
Nous sommes tous avec toi Farhat et que dieu soit avec toi.

- Juin

A partir du moins de juin les formes présentent dans les spécificités positives ont un caractère beaucoup moins politique. On trouve ainsi “ons”, “Ons”, “Bravo”, “bonne”, “continuation” en référence à Ons Jabeur, joueuse de tennis qui a remporté le tournoi Rolland Garros junior. On trouve également les formes “bac”, “bacheliers”, “bonne”, “bon”, “chance” en référence au début des épreuves du baccalauréat. Il y a également la fête des pères : “papa”, “bon”, “bonne”, “fête”, “fete”.

Il y a tout de même des références à Marzouki : “marzouki”, “المرزوقي”(Marzouki), homme politique tunisien nommé Président après l’élection de l’assemblée constituante en 2011 suite à son alliance avec Ennahdha. Les commentaires sont postés sous une vidéo d’une interview. Sur onze commentaires, deux sont hostiles à Marzouki et rédigés en alphabet latin. Les autres, en majorité rédigés en alphabet arabe sont des témoignages de soutien.

- Juillet

La plupart des commentaires du mois de Juillet évoquent le ramadan. On retrouve ainsi les formes “romdhankom” (votre ramadan), “romdhan” (ramadan), “mabrouk” (béni). On retrouve également les formes “emoticonHeart” et “emoticoneSmile” ainsi que les formes “Bonjour,” “SBE7” et “el5ir” (bonjour).

Les formes “BCE”, “sebsi” et “beji” font référence à Beji Caïed Essebsi, alors Premier ministre. Certains commentaires sont hostiles “BCE et BEN ALI 3omala2 israel et amerique c tout” (“BCE et Ben Ali sont les agents d’Israël c’est tout”), “ya si beji rahou echa3b karhik saye ifhim” (“Monsieur Beji, le peuple vous déteste, ça y est, comprenez”), “sebsi degage”.

- Août

Les spécificités positives réfèrent surtout à des formules de politesse caractéristiques du mois de ramadan ou de l'aïd el fitr (ou fête de la rutpure du jeûne, célébrée le premier jour du mois qui suit le mois de ramadan) : "3idkom" (votre fête/aïd) et mabrouk (béni).

On retrouve également en quantité les formes "allah" et "yar7mou" (que dieu ait son âme) formule postée sous une publication de la page annonçant le décès d'un humoriste tunisien très populaire en Tunisie.

- Septembre

Les formes les plus spécifiques du mois de septembre sont "bonjour", "Bonjour" et leurs équivalent en tunisien. On retrouve également emoticoneSmile et emoticoneHeart.

Beaucoup de commentaires évoquent la rentrée scolaire d'où la présence des formes "scolaire" et "mou3almin" (enseignants) : "bonne rentrée scolaire", "excellente rentrée scolaire pour tous", "bonne rentrée scolaire".

Le forme "" (filme) فلم apparaît dans le même message posté à plusieurs reprises et accusant la radio de fabriquer des films notamment par rapport aux informations publiées sur des faits divers impliquant des islamistes ou des salafistes que les commentateurs jugent mensongères et ridicules.

- Octobre

Là encore les formes les plus fréquentes sont "bonjour" et ses équivalents.

On retrouve plus loin les formes "nessma" et "نسمة" (Nessma, chaîne de télévision privée) qui avait été beaucoup critiquée pour avoir diffusé une soirée débat sur la laïcité ainsi que le film Persépolis doublé en tunisien dialectal (on y voit notamment "Dieu", représenté sous les traits d'un vieil homme discuter de la foi et de l'athéisme avec une petite fille).

3.11.3. Les commentaires Facebook entre spam-attack et [banalités ?] du quotidien

L'outil segments répétés de Lexico3 permet de repérer les occurrences formées de deux formes ou plus répétées à l'identique tout au long du corpus. L'observation des segments répétés fait d'abord ressortir une caractéristique essentielle de l'écriture sur Facebook (que l'on retrouve plus globalement dans les sites communautaires et les réseaux sociaux) à savoir l'usage très fréquent

d'émoticônes, images symboliques représentant une émotion. La forme "emoticoneHeart" est la plus répétées. Le dessin "cœur" qui indique que l'utilisateur aime quelque chose ou quelqu'un est souvent utilisé à répétition pour indiquer l'intensité.

On retrouve également dans les segments répétés des séquences répétées à l'identique dans différents messages. Il ne s'agit pas ici de séquences répétées dans un même commentaire par une même personne mais de séquences copiées/collées par de nombreux internautes à plusieurs reprises sous différentes publications de la page. La séquence répétée porte un message contestataire comme par exemple "» يا إعلام يا حقير أخرج شوف الجماهير" médias minables sortez et regardez le peuple »).

Cette pratique, qui est devenue très rependue en Tunisie après la révolution, s'appelle le *spam-attack* ou attaque de spam. Il s'agit d'une sorte de manifestation virtuelle où des internautes se donnent rendez-vous sur la page Facebook officielle de la personnalité publique ou de la marque à attaquer afin d'y poster le même message un très grand nombre de fois de sorte que plus aucun autre contenu n'est visible. Une telle attaque permet d'attirer l'attention du grand public et des médias sur une polémique ou une cause d'abord en utilisant la page attaquée qui par définition regroupe un grand nombre d'abonnés, ensuite parce que les administrateurs de la page se retrouvent souvent obligés de désactiver la page en question aussi bien pour faire cesser les attaques que pour la "nettoyer" des commentaires invasifs, suscitant de fait l'intérêt du publique et des médias quant aux causes de la désactivation.

Cependant, toutes les séquences répétées n'indiquent pas une attaque de spam. On retrouve par exemple dans les segments répétés de nombreuses formules de salutations comme "bonne journée, "seb7 el" ou "sbeh el", "el ward", "el 5ir", "ennour" ("seb7 el ward", qui pourrait être traduit par "que ta matinée soit faite de roses" et "sbe7 ennour", qui pourrait être traduit par "que ta matinée soit faite de lumière, sont des formules équivalentes à "sbe7 el 5ir", "bonjour", toutes utilisées uniquement dans la matinée). On retrouve également les séquences "tout le monde", "a tous" et "el kol" ("tout le monde"). La séquence "allah yar7mou" ("que dieu ait son âme") apparaît également un grand nombre de fois dans le corpus pour commenter très majoritairement le décès d'un humoriste très populaire en Tunisie mais aussi, dans une moindre mesure, pour évoquer les martyrs de la révolution. Il est intéressant de constater que les mois où la page a subi des attaques de spam sont aussi les mois où la courbe du vocabulaire enregistre une baisse.

Les séquences les plus répétées sont donc issues d'expressions usuelles du quotidien sans sujet particulier. Elles sont d'autant plus visibles que les graphies ne sont pas aussi variées que pour les autres termes. Elles témoignent par ailleurs d'une certaine importance des salutations parmi les usagers tunisiens de Facebook. En parcourant le HTML brut de notre corpus, on se rend en effet compte du très grand nombre de messages dans lesquels les internautes souhaitent le bonjour à la page ou aux autres internautes avant d'entamer toute conversation. Ces séquences peuvent aussi indiquer qu'un très grand nombre d'internautes sont sur Facebook dès le matin puisque ces formules sont spécifiquement employées au cours de la matinée.

En dehors des formules de politesse ou des messages de spam, les formes que l'on retrouve le plus dans les segments répétés relèvent surtout de l'actualité politique : "ben ali" suivie plus loin de " بن علي" ("Ben Ali"), "la tunisie" et "la Tunisie", "la justice", "le peuple", "la liberté", "la démocratie", "la révolution".

3.11.4. Quelques exemples de cooccurrences

Certaines formes peuvent être associées de façon récurrente mais dans des structures changeantes ce qui ne permet pas de les détecter en tant que segments répétés. Pour les repérer nous utilisons le calcul des cooccurrences. Le corpus étant découpé en commentaires, le calcul des cooccurrences permet de repérer pour chaque forme pôles les formes qui sont le plus fréquemment employées dans les mêmes sections. Ici nous observerons les cooccurrents de certains termes en rapport avec des notions politiques. Pour une même notion, nous distinguerons la forme écrite en alphabet arabe et en alphabet latin, et pour les formes en alphabet latin nous distinguerons les formes en tunisien et celles en français. Nous constituerons donc des groupes de formes par système d'écriture/langue et nous regrouperons les cooccurrents par thème sous forme de tableau. Il est à noter cependant que les formes cooccurrentes qui peuvent être détectée par ce calcul sont celles qui sont orthographiées à l'identique, la variété des graphies pour un même mot peut faire perdre des informations précieuses de ce point de vue.

- Ben Ali et Trabelsi

Dans le corpus il y a, outre les différentes graphies, deux formes pour le patronyme Ben Ali : la forme complète et les initiales (ZABA). Pour le nom "Trabelsi" il est employé avec ou sans l'article défini, au pluriel ou au singulier.

Ben Ali	بن علي	Zaba	Trabelsi	طرابلسي
Noms propres				
trabelsi karoui Ghannouchi Karkar Mourou	الطرابلسي (Trabelsi) جراد (Jrad) ليلى (Leila) صخر (Sakhr)		laila belhassen 3imed (Imed) ben ali fehri nasra	ليلى (Leila) بلحسن (Belhassen) عماد (Imed) بن علي (Ben Ali) محجوب (Mahjoub) الماطري (Matri) شيبوب (Chiboub)
Richesses				
		chba3 (rassasié) financements ressources	milk (propriété) leflous (l'argent) clan actionnaire	
Justice				
		trahie 9atala (a tué) arnaqué	yet7asoub (jugés) arrêté	إقصاء (exlusion)
Termes injurieux				
kleb (chiens) salaud mafia	(le déchu المخلوع)	kacha5tou (sa gueule) kalb (chien) joubana (lâches) voleur		
Politique				
etajamo3 (RCD) MTI Ennahdha 1991 militants islamistes la démocratie eti7ad (UGTT) edawla (l'Etat) le peuple attentats	التجمع (RCD) الدستور (la (constitution الجيش (l'armée) الثورة (la révolution)			دستور (constitution) البرلماني (parlementaire) النظام (le régime) المواطن (le citoyen) مجلس (assemblée) البلاد (le pays) ثورة (révolution)

Tous ces groupes de formes sont en spécificité positive au mois de janvier et, dans une moindre mesure, au mois de mars pour les groupes de formes liés à "Ben Ali". Il semblerait donc que dans

notre corpus la conversation s'est très vite détournée des figures de l'ancien pouvoir. Ou tout du moins, l'apparition de ces formes a diminué de façon telle qu'elles semblent anecdotiques ou particulièrement absentes dans les autres mois.

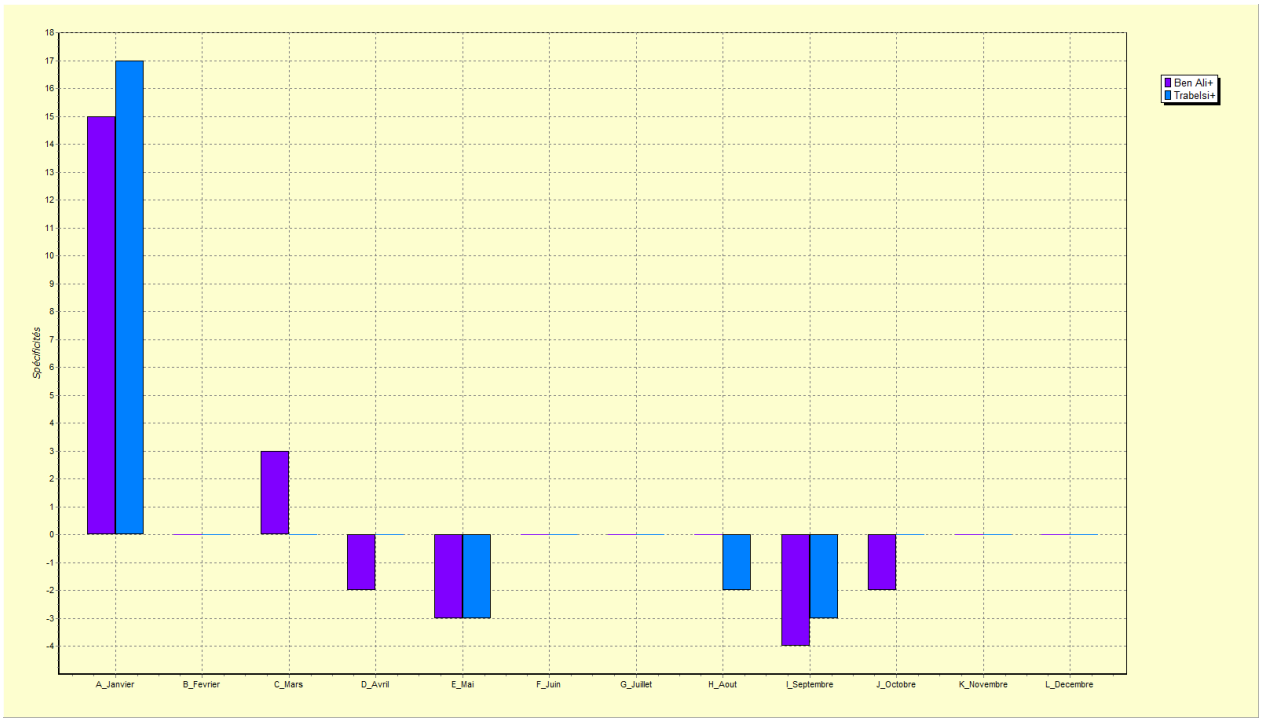


Figure 5 Graphique de ventilation des formes « Ben Ali-Trabelsi » par mois

A l'intérieur du mois de janvier, on constate également un écart significatif dans la ventilation des formes Ben Ali et Trabelsi ainsi que par rapport au reste du corpus.

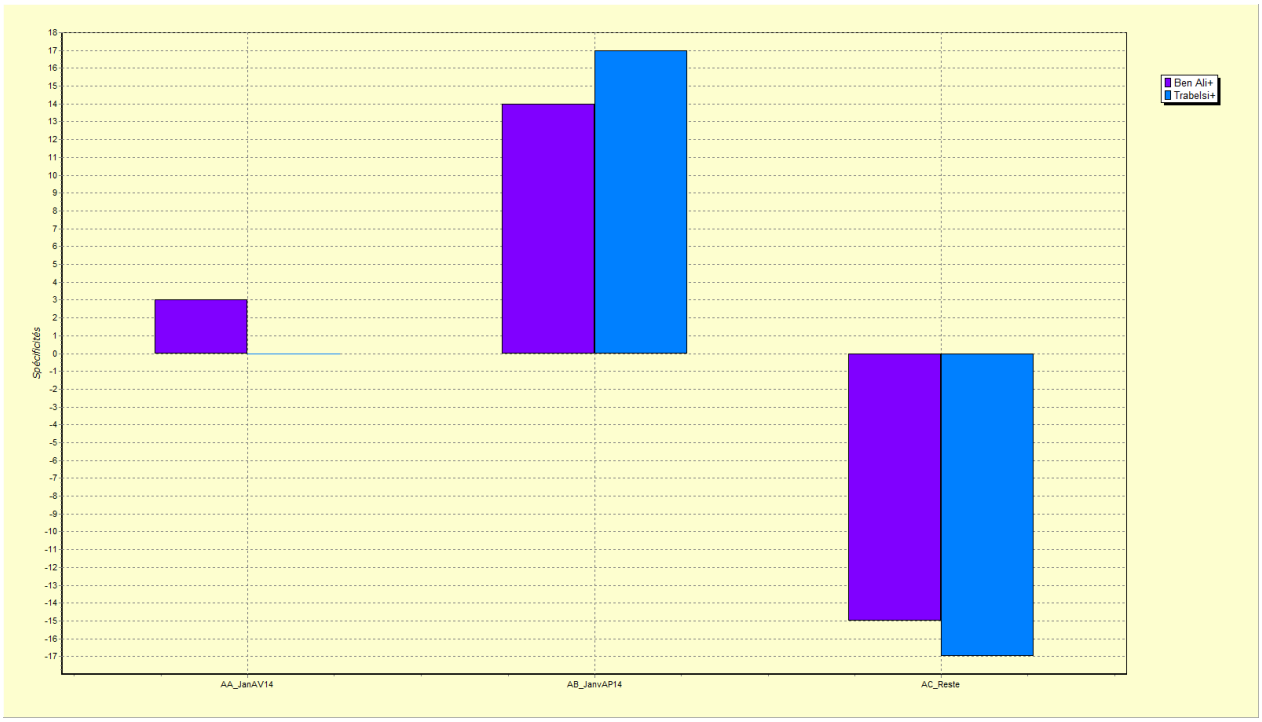


Figure 6 Graphique de ventilation des groupes de formes « Ben Ali-Trabelsi » pour janvier

Concernant les formes “Ben Ali” et “Zaba”, il est intéressant de voir que seules les formes “Ben Ali” sont présentes avant le 14 janvier alors qu’après le 14 janvier, on voit apparaître les formes “ Zaba” et “Ezzine”.

Partie : AA_JanAV14, Nombre de contextes : 13

عباد ماتت في سبيل مطالب الحرية § yeziw bla t7in ya mozaique § Merci ben ali merci § ! !
yoghayerou ma bi enfousshoum " " § Bravo ben ali § ti bjeh rabii flous edawla jbed alihom
hmou 3ala rouahkom ! ! ! ! ! § President Ben Ali it is time to go as you can not run
a3 ta9ra wnes tarja3 te5dem brabi § Tawa ben ali wala behi mafamech wa7id f tounis machitmouch
nebdaw fi mar7la jdida weli ma ye3jbouch ben ali y9olli chkoun y7ebou w rana ne7kiow
let estis9a fil youtube § Ya7ya el zin . Zin ya batal § Chey iesef . . . . 9alek discours
discourt du pr#esident pour dire : vive ben ali ! ! Et ce dans plusieurs ville de tunisie
r#esident a #emus tous le monde ! § Vive BENALI VIVE LA TUNISIE weli moch 3ajbou yemchi
bech twalli , el 7assel 7lewett § Leila Ben ALI , Belhassan , ou Sakhr Pr#esident 2014
! h twalli , el 7assel 7lewett § Leila Ben ALI , Belhassan , ou Sakhr Pr#esident 2014
a what ever will be will be , the end of Ben Ali ca sera sera . you will see said Sabi
ens § TT LE MONDE PHOTO DE PROFIL DEGAGE BEN ALI SVP SVP SVP , na9rh#ek mosaique , va
TT LE MONDE PHOTO DE PROFIL DEGAGE BEN ALI SVP SVP SVP , na9rh#ek mosaique , va en §

Partie : AB_JanvAP14, Nombre de contextes : 92

annouchi n ' a tu#e personne . . w RCD w ZABA houma illi 9atlou ya msatka . . § bienvenuuu
§) : bienvenu § Ya mar7b#e § شي كلمته المشهورة ثقتي في الله اولا و في بن علي ثانيا
fouh saf#e7 kbir min 9bal w k#en ami l ' zaba ama ki marka7louch w ma3tachi une part du
f#eha mouch hrab kil kalb , , kifou kif ZABA ki hrab ! ! ! § ma 7achetnech bih men gher
مفهمتوش § ! ! ! f#eha mouch hrab kil kalb , , kifou kif ZABA ki hrab اش تعرفو عليه حتى شي؟
, e#e#e7 ! ! #eli yo9til tw#ensa kifou kif zaba ! ! w howa 9tal barcha m3ah w controu
iiiiiiiiiir mel bled w ya5let 3la sidou ben ali fi jeddah 5ater eli 9oulou ghannouch
l~et ta3rf 5atir awil 3b~ed 9alou l~e ki ben ali dhhor 3la 79i9tou m~em pas b3d 2 ans
ns mili chad l7okm , w tawa ki say~e tar ben ali wal~ew lkol ya3rfou eli howa dictateur
akom min loughit 5wenjiya 5atirha loghet ben ali w ben ali maw 5ayib § ghriba fennes
oughit 5wenjiya 5atirha loghet ben ali w ben ali maw 5ayib § ghriba fennes elli takrah
orc#es , ne dut la vie vie sauve qu ' #a Ben Ali d#ecourageant Bourguiba de le condamner
s le projet d ' assassiner le pr#esident Ben Ali , a #et#e probablement prise en mars
, Jean - Pierre Tuquoi intitul#e Notre ami Ben Ali , sorti en 2002 #a la D#ecouverte
pages 212 - 213) . § m^eme chose si ben ali revient il aura cet accueil par les) "
: § ! ! ! ab yfarga3 d#enia , , 3l#ech ? ? ? 5atir Zine l ' kalb ma3tachi wzara b#ehia
fhim rak ! ! si rachid h#edha k#en sa7ib zaba ki marka7louch w 3tah une part du gateau

Pour tous les groupes de formes hormis “Zaba” il y a de nombreux noms propres parmi les cooccurrents. Ces noms font soit référence à des membres de la famille Ben Ali-Trabelsi ou des membres de familles apparentées (Leila, Imed, Belhassen, Mahjoub, Chiboub, Matri) ; soit à des hommes d’affaires proches de l’ancien pouvoir (Karoui, Fehri, Nasra) ; soit à des personnalités politiques tunisiennes (Ghannouchi, Mourou, Karkar, Jrad).

En ce qui concerne les noms et prénoms des membres de la famille Ben Ali-Trabelsi et des familles apparentées, on les retrouve dans des messages diffusant des informations concernant leur arrestation ou leur mise en examen, ou dans des messages s’inquiétant de ne pas les voir arrêtés ou mis en examen. Les informations diffusées sont contradictoires et semblent témoigner d’une grande confusion. Certains messages expriment leur défiance par rapport aux informations véhiculées par les médias

سليم شيبوب وينو ?? عماد الطرابلسي كي مات وينو ?? واخواته وينهم ?? والطرابلسية إلي تشدو في مطار وينهم ?? كي شدهم علاش معدوهمش في تلفزة??? فيقو يا توانسة !!!! راهم يلعبولنة في لعبة !!!

- Slim Chiboub, il est où ? Imed Trabelsi, s'il est mort, il est où ? Leurs frères et sœurs, où sont-ils ? Et les Trabelsi qui ont été arrêté à l'aéroport, où sont-ils ? S'ils ont été arrêtés, pourquoi on ne les a pas passés à la télé ? Tunisiens réveillez-vous ! Ils nous jouent des tours !

Les noms propres qui apparaissent dans les cooccurrents des formes "Ben Ali" sont liés à des figures politiques ou publiques. Les références aux islamistes évoquent surtout les dangers de leur retour en rappelant notamment les violences commises par le passé.

Le Rocher , 2002) que " Frapp#e de stupeur , fou de rage , Bourguiba exigeait la t^ete des hauts) dirigeants de la secte islamiste . Il s ' appr^etait ainsi #a entrer dans le jeu diabolique des strat~eges islamistes en leur offrant des martyrs . ' Si Dieu veut que je devienne le martyr des mosqu#ees , qu ' il en soit ainsi . Mais je vous dis que ma mort ne sera pas vaine et que de mon sang , na^itront des fleurs islamiques ' communiqua Ghannouchi (cit#e par Le Monde du 2 septembre 87) . " Condamn#e aux travaux forc#es , ne dut la vie vie sauve qu ' #a Ben Ali d#ecourageant Bourguiba de le condamner #a mort afin de ne pas en faire un martyr . Erreur : - 8 novembre 1987 . Tentative de coup d ' #etat pr#epar#ee par Ghannouchi et son bras droit Salah Karkar . En mai 1991 , Abdallah Kallel qui fut ministre de l ' Int#erieur lors de la tentative avort#e de coup d ' #etat , r#ev~ele que " la d#ecision de passer #a l ' action violente y compris le projet d ' assassiner le pr#esident Ben Ali , a #et#e probablement prise en mars lors du dernier congr~es d ' Ennhada " (rapport#e par Lib#eration du 23 mai 1991) . Le premier , 1988 a vendre la m~eche sera l ' ancien num#ero deux de Ennhada , Abdelfattah Mourou dans Jeune Afrique# du 12 juin 1991 : " Rached Ghannouchi a toujours refus#e de dialoguer ; il a choisi le recours #a la violence " . Ce m^eme Mourou d#emissionna du mouvement " non violent " du Figaro #a la suite de l ' action terroriste du 17 f#evrier 1991 contre un local du parti RCD #a Bab Souika . Mieux , dans un livre d entretien avec les journalistes Nicolas Beau et Jean - Pierre Tuquoi intitul#e Notre ami Ben Ali , sorti en ' a la D#ecouverte , Salah Karkar (bras droit de Ghannouchi) avoue carr#ement : " Les# 2002 sympathisants du MTI au sein de l ' arm#ee pr#eparaient un coup d ' Etat , pr#evu pour le 8 novembre (. . cette d#ecision a #et#e adopt#ee par le bureau politique du mouvement islamiste (. . .) Nous n ' avions (.

Quant aux références aux hommes d'affaires ou autres personnalités politiques, ce sont le plus souvent des accusations par rapport aux liens entre ces personnes et l'ancien régime. Le terme "Ben Ali" dans ces messages devient une accusation en soit.

vous les hommes d'affaires étrangers au métier de la communication vous vous etes trouvé parachuté dans ce domaine à l'epoque de Ben Ali , en aucun cas vous ne pouvez construire un

média digne !!! a voir la honte qu'a fait nabil karoui !! le virage de 180 degré !!! couilles moles !!!!!!

Les termes faisant référence aux richesses accumulées par la famille Trabelsi sont caractéristiques des groupes de formes “Zaba” et “Trabelsi” alors que de tels termes sont absents des cooccurrents des autres groupes de formes. Les termes relatifs aux “crimes” de Ben Ali-Trabelsi et à la justice suivent également cette répartition. Les termes injurieux sont cooccurrents de “Ben Ali” et de “Zaba” avec deux emplois différents selon le groupe de formes : dans le premier cas les injures ne désignent pas Ben Ali lui-même mais ses partisans ou ses anciens alliés (“les chiens de Ben Ali”, “les chiens du RCD”, etc.) ; dans l’autre, c’est directement “Zaba” qui est visé.

ya ta7anna wa9et elli les jeunes imoutou fi kassrine entom ya ihoud te7teflou bel réveillon !!!!!!!! ya mosaique enti wa9teha t3adi fel musique donc brabi sakrou famkom ya kleb ben ali
➢ Mosaïques, lèche-cul, quand les jeunes mourraient à Kasserine, vous, espèces de juifs vous fêtiez le réveillon ! Mosaïque tu passais de la musique donc s’il vous plaît fermez vos gueules, chiens de Ben Ali.
Zaba le menteur Zaba le voleur

Tous les groupes de formes hormis “Zaba” et “Trabelsi” ont dans leurs cooccurrents des termes relatifs à des questions politiques. Dans ces messages les auteurs dénoncent surtout les alliés de l’ancien régime, on les noms de Ben Ali et de Trabelsi comme accusation voire comme insulte.

j'aime bien sami fehri et j'adore ses feuilletons fé ton boulot et boucle la fils de ben ali

- Révolution - Thawra - ثورة

Nous étudions ici les cooccurrents pour trois groupes de formes constitués autour de la forme en français, de la forme en tunisien et de la forme en arabe.

Révolution	**Thawra**	**ثورة**
Révolution		

peuple jeunes martyrs servir vive pays solidarité liberté	cha3b (peuple) chabeb (jeunes) el7oria (la liberté) modhtahdin (opprimés)	الشعب (peuple)
Contre-révolution		
voler (au sens de confisquer la révolution) manipulent	mounef9in (hypocrites)	ميليشيات (milices) الفوضى (désordre) التجمع (RCD) الفساد (la corruption)
Termes politiques		
gouvernement	7oukouma (gouvernement)	الحكومة (le gouvernement) الدولة (l'Etat) الأمن (la sécurité) الجيش (l'armée)

Les occurrences relatives à la révolution sont spécifiques du mois de février. Le groupe de formes "révolution" présente un léger pic au mois de mars et le groupe de formes "thawra" un léger pic au mois d'août.

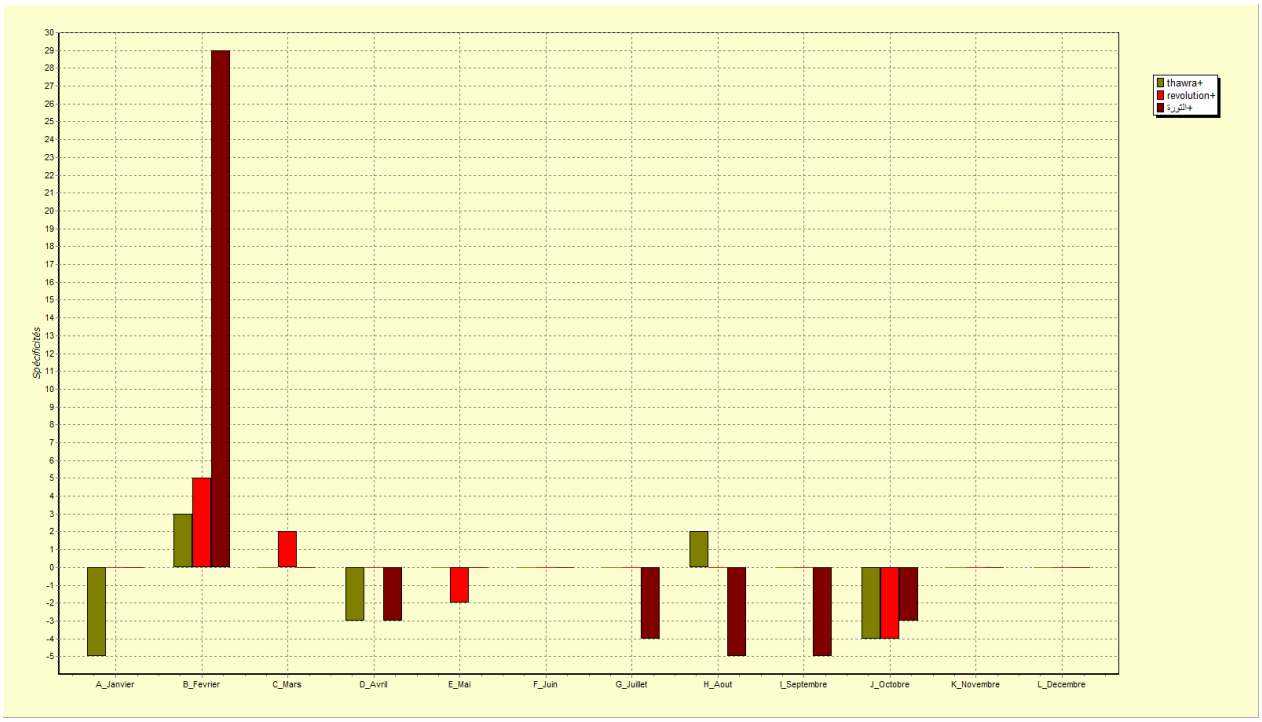


Figure 7 Graphique de ventilation des groupes de formes "révolution"

On note une certaine symétrie entre les formes en alphabet latin et les formes en alphabet arabe. Les thématiques que l'on retrouve dans les commentaires sont les mêmes mais la proportion de cooccurrents rattachés à ces thématiques varie selon le système d'écriture.

Pour les formes en alphabet latin, il y a une majorité de termes révolutionnaires et patriotiques. On y trouve notamment des références au peuple, aux jeunes, aux martyrs, aux opprimés, etc. ainsi qu'à des valeurs telles que "la liberté", "la solidarité".

svp respectez les sentiments des familles des martyrs de la révolution et du peuple tunisien en général, les chansons doivent être toutes à propos de la tunisie, la révolution et les martyrs, mouch mzia, et vous devez respecter aussi les auditeurs qui vous appelent et ne pas leur raccrocher au nez, et mahich mzia zada
tous unis contre ben ali et compagnie ces minables veulent nous terroriser ils pensent qu'on est pas capable de vivre la démocratie et qu'on doit dépendre de la mafia du rcd et des trabelsi mais le peuple qui arrive à renverser le dictateur (inchallah) réussira à surmonter tout ses problèmes et imposera la demo dans tout les pays arabes vive la Tunisie et vive le peuple

Les formes en alphabet arabe renvoient à des messages plus politisés. L'idée d'une contre-révolution en marche est notamment très présente. Les appels à l'action pour défendre la révolution sont plus nombreux.

<commentaire=JanvierAP_809>قبل الفرحة يلزم التظاهر والمطالبة بحل التجمع حزب بن علي , يا ناس راهي مؤامرة كبيرة تحاك ضدنا ومسرحيات لإجهاض **الثورة** . . يا ناس فكرو بسوية عقل . . منع جولان يرافقه سحب قوات **الجيش** والشرطة وترك المواطنين العزل يواجهون مصيرهم أمام العصابات المدربة والمسلحة من **ميليشيات** بن علي وفي كل العالم معروف أن منع الجولان يصاحبه تعزيزات مضاعفة لعناصر **الجيش** ؟؟؟ أليس غريبا هذا ! ! ! اللهم إلا في حالة أنهم يريدون ترهيب الشعب لعدم التظاهر مجددا والمطالبة بحل التجمع والإنشغال بتأمين حياتهم وممتلكاتهم . إعطاء أرقام نجدة للمستغيثين لا ترد عليهم عند الإتصال ؟؟؟ الإعلام عن إعتقال أفراد من العائلة الحاكمة تم نعرف من خلال الجزيرة أن هذا في دبي والآخر في فرنسا والثالث في الصين والرابع في الواق واق ؟؟؟ أليس غريبا هذا ! ! ! اللهم إلا أنهم يريدون خداع الناس لمساعدتهم في الفرار خارج البلد . الإعلام عن إعتقال السرياطي وبعض العصابات المسلحة ولا نرى سوى هراوات وسكاكين ؟؟؟ ومواطنين يقع تسريحهم لاحقا . أليس غريبا هذا ! ! ! عبد الله القلال مجرم الحرب يشرف على منح السلطات لهذا وذاك بكل وقاحة ورقعة صحيحة والمنطق يقول أن مكانه في إحدى الزنزانات ينتظر الإعدام ؟؟؟ أليس غريبا هذا ! ! ! نقيب في الأمن يصرح لقناة الجزيرة أن الأمور غامضة وهناك رائحة مؤامرة في الأفق ولن تهدأ الأوضاع قريبا ثم قطع المكالمة قائلا إنني أصرح بكلام خطير ؟؟؟ ألأيس غريبا هذا ! ! ! §
➢ Avant les célébrations, il faut manifester et exiger la dissolution du RCD, le parti de Ben Ali. Hé les gens, il y a une grande conspiration qui se prépare contre nous pour faire faire

> avorter la révolution. Hé les gens, réfléchissez avec un peu de logique. L'interdiction de circuler est accompagnée du retrait des forces armées et de police laissant les citoyens affronter seuls leur sort face aux gangs entraînés et armés des milices de Ben Ali (...)

- Démocratie - Dimo9ratia - ديمقراطية

Nous avons constitué ici trois groupes de formes. Le premier est en arabe, le deuxième est en français et le troisième est écrit en lettres latines mais suit une prononciation arabisée (dimocratie/dimo9ratia).

Démocratie	Dimo9ratia	ديمقراطية
Références politiques		
vive pays peuple cytoyen (citoyen)	vive cha3b (peuple)	تونس (Tunisie) لعباد (les gens) ثورة (révolution)
Valeurs humaines/civiques		
liberté RESPECT habet	el7oriyya (la liberté) i7tiram (respect) esselmi (pacifique)	
Mots grammaticaux		
ne pas ne sont pas	mouch (pas)	لم (négation du présent ,ne...pas) لن (négation du futur ,ne...pas) لا (négation du passé ,ne...pas)
Défiance		
	kedhba (mensonge) sila7 (armes)	الكذب (le mensonge) إشاعات (rumeurs) بندر(jouer du bendir, "équivalent à "lécher les bottes")
Références politiques		

		الحكومة (le gouvernement) برلماني (parlementaire)

Les références à la démocratie se retrouvent surtout dans les mois de janvier, de mars et d'avril, qui correspondent à des périodes où de nombreux débats ont animés la page Facebook.

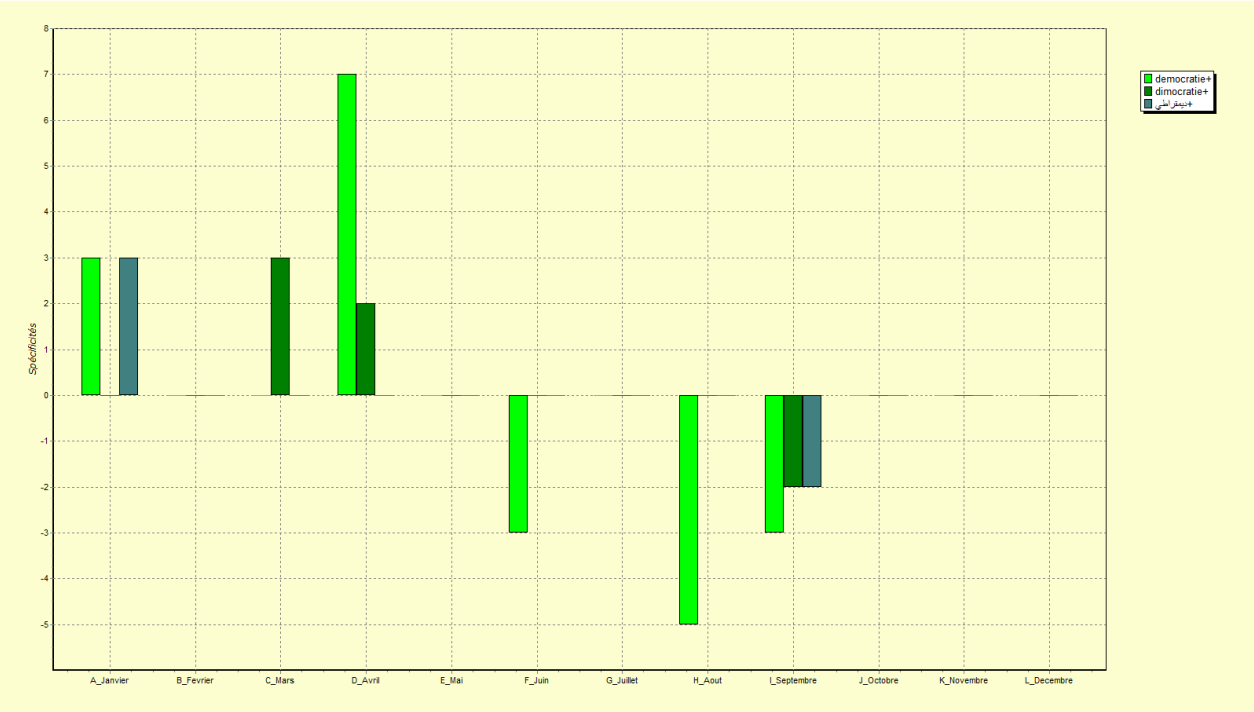


Figure 8 Ventilation des groupes de formes "démocratie" par mois

Pour ce qui est du mois de janvier, un fort pic est enregistré dans la deuxième partie du mois.

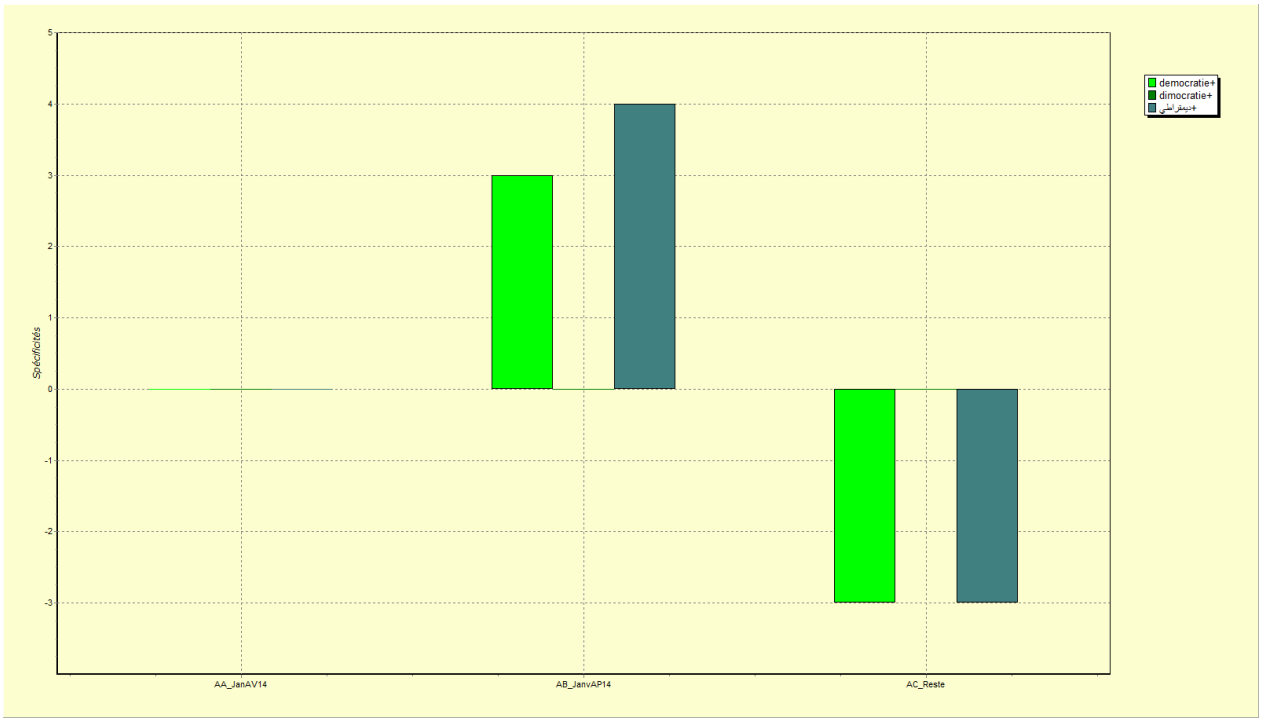


Figure 9 Ventilation des groupes de formes "démocratie" pour le mois de janvier

On retrouve ici quelques termes similaires aux cooccurrents des groupes de formes relatifs à la révolution : peuple, pays, vive, liberté, etc. De nombreux messages associent la démocratie, la liberté et le respect.

ca fé mal au coeur ces commentaires d'insultes!!!!!!!!! c'est ca la democratie, la liberté!!!! vous montrez bien que vous ne meritez pas !!!!

On remarque la présence de formes négatives dans les cooccurrents des trois groupes de formes. Les structures négatives sont le plus souvent utilisées dans une tentative de donner une définition au concept de “démocratie”.

eeeehhhhh oui la dimocratie mouch forcement nkounou 3adna nafs erray.... amma illi yajma3na houwa DRAAAAAAAAPEAU BLADNA
➢ Eh oui la démocratie ce n’est pas forcément qu’on a le même point de vue… mais ce qui nous réunit c’est le drapeau de notre pays.

Le terme “haybet” qui apparaissait aussi dans les cooccurrents des groupes de forme “révolution” fait référence “au prestige de l’Etat”, une expression souvent utilisée par Béji Caïed Essebssi, alors

Premier ministre. Les commentateurs associent cette expression à l'ancien régime et lui opposent "haybet echa3b", le prestige du peuple.

Bravo Mosaique, voilà un reportage qui reflète bien le citoyen tunisien : echa3b ettounsi féye9 wmatet3adech 3lih, mouch na7ina ben ali bech yjina wé7ed e5er y7eb yraja3 haybet edawla! ya 7asra, tawa haybet echa3b! et on doit servir haybet echa3b! (dignité, transaparence, sécurité, économie, démocratie..)le sécurité fait bien parti de servir le peuple, et le gouvernement doit respecter les acquis de la révolution et rendre compte au peuple! plus de place pour l'ancienne école

De nombreux appels sont lancés pour faire tomber le gouvernement, dissoudre le Parlement et la chambre des conseillers et instaurer un régime parlementaire au lieu du régime présidentiel. Il faut préciser ici que l'idée d'instaurer un régime parlementaire a été portée par le parti Ennahdha dès la reprise de ses activités politiques sur le sol tunisien quand les autres partis politiques étaient totalement contre ou exprimaient un avis mitigé. Les occurrences faisant référence à un régime parlementaire pourraient donc être indicatrices de l'appartenance politique de l'émetteur du message dans lequel elles apparaissent.

<commentaire=JanvierAP_84>إلى دعاة الديمقراطية المزيّفين . شن هجمات على الإسلاميين لا يعبر إلّا عن ديكتاتورية الحاقدين و عن ضرب الديمقراطية التي يحاولون إستحمالها لمصالحهم فقط . ليكن في علم الدعاة المزيّفين إن الديمقراطية تمثل حكم الأغلبية و تمثيل نسبي لكامل شرائح المجتمع . إذا إقصاء الإسلاميين الذين لهم دعم شعبي لا بأس به يقارب العشرين بالمئة هو تعدّي واضح و غير مقبول بتاتا على الديمقراطية و حق الشعب في إختيار من يمثله . ثانيا يجب تفريق الإسلاميين التونسيين عن إسلاميين المشرق العربي . فالتونسي أكثر وعي و أقرب للوسطية من بقية العرب و الأعاجم . ثم إن أكاذيب التجمعين التي ملأت أذاننا منذ الإستقلال يجب التحرر منها فلا يكفي قطع الحاضر عن الماضي في الوقت أن ألاعيب و أكاذيب التجمع و النظام السابق مازالت تتحكم في أراءنا و في إختياراتنا . ثالثا لا يحق لأحد إعتبار أرائه و مواقفه هي الصحيحة و إن مبادئه السياسية هي الأفضل لتونس و إن من يخالفه هو على خطإ و جهل، و أنا أولهم . الذين يشككون في نوايا الإسلاميين يجب أن يعرفوا أن الإسلاميين نفسهم يشككون في نوايا من يسعى إلى دحض حقهم السياسي و الفكري . فالإختلاف حق و الإقصاء جريمة في حق الأحرار و الشهداء . فتونس لكل التونسيين و ليست حكرا على من إغتاظ من دور الإسلام في تونس المسلمة . الأحرى الأحرى دعم الديمقراطية و حق الإختلاف و ليس إقصاء المنافسين و عودة إلى الديكتاتورية القديمة . والسلام §

- (...) derrière le gouvernement actuel, il y a un autre gouvernement qui s'appelle les fidèles de l'ancien régime (...) Tunisiens il faut un régime parlementaire (...) un régimes présidentiel n'est plus garant de la démocratie (...)

- Laïcité et sécularisation

On retrouve dans le corpus quatre groupes de formes qui expriment les concepts de laïcité et de sécularisation : laïcité/اللائكية et 3ilmaniyya/العلمانية (sécularisation).

Cooccurrents des groupes de formes pour le concept de laïcité			
Laïcité	**اللائكية**	**3ilmaniyya**	**العلمانية**
Références aux islamistes ou à Ennahdha			
Ennahdha islamiste islamistes islamique	(islamique) إسلامي	inahda (Ennahdha) ennahdha islem (islamique)	(Ennahdha) النهضة (les nahdhaouis) النهضاويين (les islamistes) الإسلاميين (le cheikh) الشيخ
Religion			
leslem (l'Islam) religion islem (Islam)	(musulmans) مسلمين (la religion) الدين	moslim (musulman) rabbi (Dieu) islem (Islam) chari3a (charia) kofr (apostasie)	(musulman) مسلم (l'islam) الإسلام
Exemples de pays			
	(Irak) عراق (Tunisie) تونس		(modèle) نموذج (l'exemple) المثال (saoudien) السعودي (français) الفرنسي (turc) التركي (Turquie) تركيا
Notions politiques			
Etat démocratie	(patrie) وطن (l'extrémisme) التعصب	laicité	(l'Etat) الدولة (le peuple) الشعب (politique) السياسية
Termes négatifs, péjoratifs ou injurieux			
contre	(contre) ضد (discrimination) التفرقة	anti chlayki (savate, insulte) tafaha (futilité) bhim (bête) bassesse débiles jahl (ignorance)	(malveillance) حقد (mensonge) كذب

On constate d'abord que les deux formes les plus employées sont "laïcité" et "العلمانية" ("sécularisation"). Le groupe de formes "laïcité' est en spécificités positives aux mois de mars et d'avril tandis que les trois autres sont concentrés dans un mois ou dans l'autre selon le graphique de ventilation.

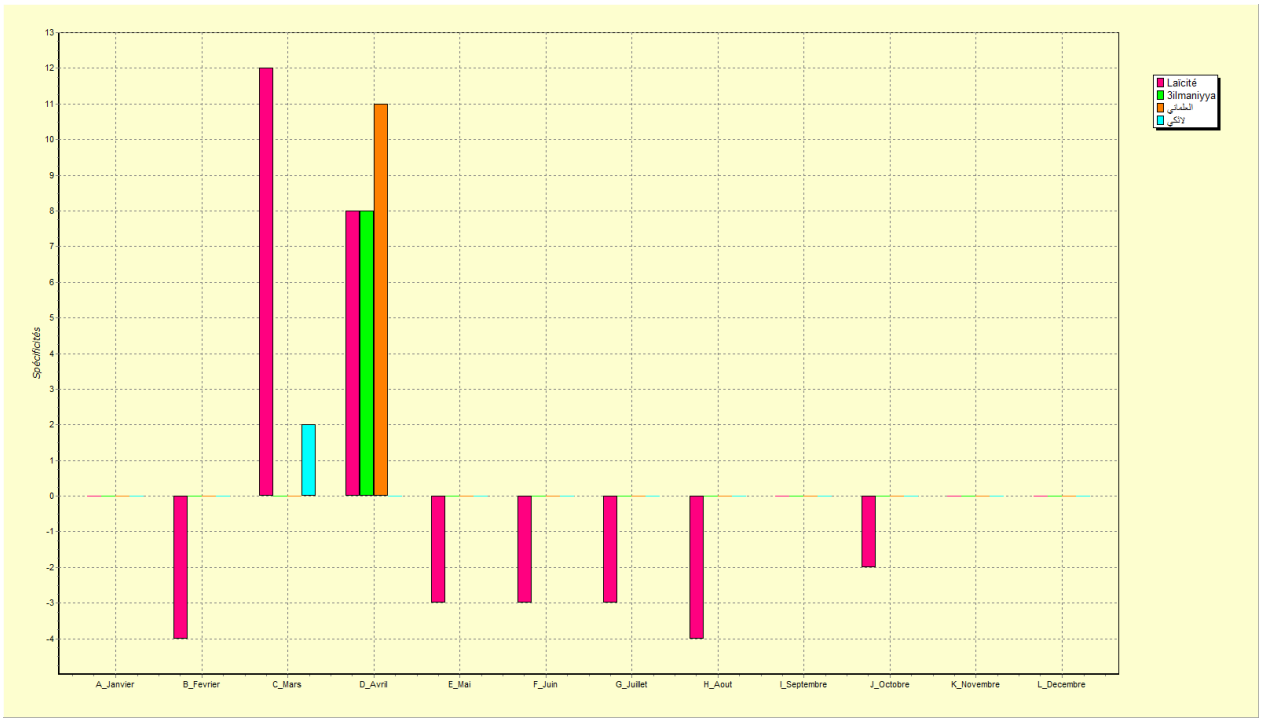


Figure 10 Ventilation des groupes de formes "laïcité"

Cependant en observant le détail des commentaires, les deux concepts et toutes les formes semblent employés sans distinction. L'avis exprimé est en très grande majorité fortement hostile. On retrouve dans les cooccurrences de nombreux termes exprimant l'opposition, négatifs ou même injurieux (contre, futilité, bassesse, etc.)

hedhy 7adhret rou7ha mli7 kif ma guelet, thomma 9ara2touhou thomma 9ara2touhou, looooooool, 9addechkom tafaha ya 3elmanyin
➢ Celle-là elle s'est bien préparé, quand elle a dit "et puis j'ai lu, et puis j'ai lu", looooooool, qu'est-ce que vous êtes futiles les laïcs.

D'autres revendiquent l'islam comme religion du peuple, de l'Etat, du pays et estiment qu'opter pour le modèle laïque équivaut à renier l'Islam.

borjouliya mele5 elli ysami fi rou7ou moslim mayenta5eb ken 7esb ytaba9 el chari3a elli hbat biha rabbi, donc elli mech mwefe9 ma3 enahdha famma a7zeb o5ra eslamiya welli mech bech yenta5eb 7ezb eslemi w yonsor din el islém bech yets'el 9odem rabbi 3léch ensort la3bed elli tnedi bel 3elmeniya w mansortech 7okm rabbi!! w wa9t'ha barraw 9oulou lé makech fehem "él lé'ikiya"rakom 9odem rabbi ma3and'kom 7atta 7ojja kif y9ollek chbik ensart we7ed 3elmeni ybadel 7okmi w 9anouni w chari3ti bel 9anoun men 3andourak kif t9oul rani 3elmeni ma3net'ha rak moch moslim 3ala 5ater el moslim houwa elli yestaslém el 7okom rabbi w yardha belli ychar3oula plupart des gens ye7kiw 3an jahl beddin emte3hommele5er le7keya fiha kofr w janna w nar alors réflicissez bien

- Sérieusement, finalement celui qui se dit musulman ne peut voter que pour un parti qui applique la charia de Dieu, donc celui qui n'est pas d'accord avec Ennahdha, il y a d'autres paris islamistes et celui qui ne votera pas pour un parti islamiste qui va défendre la religion musulmane devra s'en justifier devant Dieu pourquoi il a défendu les gens qui appellent à instaurer la laïcité au lieu de défendre la loi de Dieu !! Et là, allez dire à Dieu "non tu n'as pas compris la laïcité". Rendez-vous compte que devant Dieu vous n'aurez aucune excuse quand il vous dira "pourquoi tu as défendu un laïc qui va changer mes règles et mes lois et ma légitimité par des lois à lui" (...)

Les rares commentaires qui n'expriment pas un rejet de la notion de laïcité/sécularisation ne la défendent pas non plus mais orientent leur propos sur les islamistes et notamment sur Ennahdha :

<commentaire=Avril_744>الاسلاميين يحبونا ما ناخذوش المثال السعودي و الطالباني . . . اما ناخذوا المثال التركي . . . . . و كي تجي تقوللهم مالا خلي تكون تونس دول#e علماني#e . . . يقلك أعوذ بالله أستغفر الله نحن شعب مسلم و لن يستسلم . . . . يظهرلي فاهم نساو اللي تركيا منصوص في دستورها اللي هي علماني#e . . . . . مالا انفصام في الاشخصي#e . . . . النهضاويين يحبو يطبقوا المثال اللي في تركيا العلماني#e و في نفس الوقت يرفضو العلماني#e و ينعتو العلمانيين بالالحاد و باعداء الدين . . . . حوز تفهم . . . . . §

- Les islamistes ne veulent pas que nous suivions le modèle saoudien ou taliban… Mais le modèle turc… Et quand tu leur dis, laissez la Tunisie être un Etat laïc… Ils te disent "que Dieu me pardonne, nous sommes un peuple musulman et nous ne capitulerons pas"… Il me semble qu'ils ont oublié que la constitution turque stipule que c'est un Etat laïc… Quelle schizophrénie… Les nahdhaouis veulent appliquer le modèle turc laïc et en même

temps refusent la laïcité et traitent les laïcs d'athées ou de polythéistes… Allez comprendre...

Donnez moi un pays de l'histoire contemporaine dont le pouvoir est détenu par les extrémistes islamistes et qui est une vraie démocratie et je voterai enahdha !PS : Avant de parler de la Turquie, renseigner vous par rapport à son armée et sa position vis à vis de la laicité.

De nombreux commentateurs évoquent des modèles d'autres pays par rapport à la religion et à la question de la laïcité. Mais il s'agit le plus souvent d'une distanciation plus que de la recherche d'un exemple à suivre.

commentaire=Avril_818>Achour Anis> حتى اللي يناديو بالعلماني#e لا يحبو النموذج الفرنسي و لا التركي و لا السويدي يحبو نموذج للعلماني#e خاص بتونس كيف ما النهضاويين يحبو نموذج اسلامي خاص بتونس حسب اللي فهمتو من كلامك §
➢ Même ceux qui appellent à instaurer la laïcité ne sont ni pour le modèle français, ni turc, ni suédois, ils veulent un modèle laïc spécifique à la Tunisie, de même que les Nahdhaouis veulent un modèle islamiste spécifique à la Tunisie, d'après ce que j'ai compris de ton propos.

- Ennahdha

La variation des formes ici des est due surtout à l'utilisation ou non de l'article défini "el" ou "ال" devant le nom du parti. Nous constituons deux groupes de formes l'un pour l'alphabet arabe et l'autre pour l'alphabet latin.

Ennahdha	**النهضة**
Religion	
moslim (musulman) din (religion) lislam (l'islam)	(Allah) الله (la religion) الدين (le Cheikh) الشيخ

Laïcité	
turkie (Turquie) 3elmeni (laïc)	علمانية (laïcité) نموذج (modèle) التركي (turc) الطالبانية (taliban)
Elections	
voterai nenta5ab (je vote) 9arraret (j'ai décidé) pouvoir idiologie (idéologie) double i3lém acharnement	الإنتخابية (les élections) الديمقراطي (démocratique) للنقد (critique) السياسية (la politique) السلطة (le pouvoir)
Libertés individuelles	
alcool foulard mamnou3 (interdit) régresser	تهدد (menace)

Les deux groupes de formes sont spécifiques du mois d'Avril, mois notamment marqué par un débat sur la laïcité. Les formes en alphabet latin sont aussi spécifiques du mois d'octobre qui correspond au mois des premières élections après la révolution. On remarque notamment que les cooccurrents des deux groupes de formes sont assez similaires, il y a une cohérence dans les thèmes abordés. Par ailleurs, on voit se dessiner une certaine rhétorique dans les messages favorables à Ennahdha tandis que les messages défavorables semblent plutôt disparates dans leur argumentation.

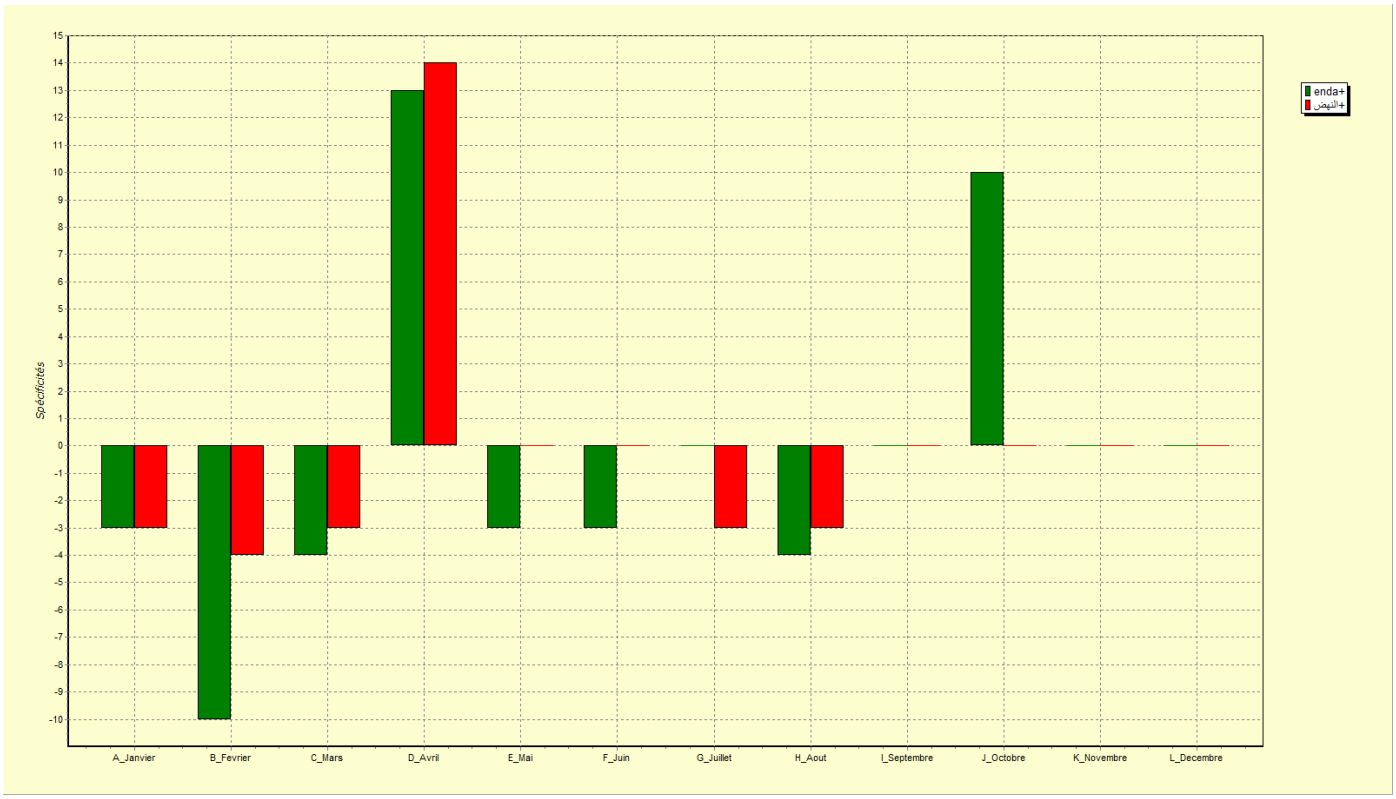


Figure 11 Ventilation des groupes de formes "Ennahdha"

Beaucoup de commentaires favorables à Ennahdha évoquent l'adhésion aux idées du parti par conviction. Il s'agit de véhiculer l'idée que c'est un choix réléchi et assumé avec déjà des intentions de vote exprimées dès le mois d'Avril.

c vrais ennahtha ebdina nefehmou fiha w moch kima kenou ya7kiw a3liha , esma3et el moncef ben salem , nordine leb7iri , lotfi zitoun rached el ghanouchi w franchement wellit ne3taz en3ich fi ebléd fiha nés kima héthouma mostawé 3lmi w tha9afi , w saberr moch normal , sem7ouni ejme3at ennahtha kont ghalet fikomm wenchallah rabi eysama7ni et mnt , ennahtha hiya el 7al fi tounes w rabbi yonserha
On commence à comprendre Ennahdha et ce n'est pas comme ce qu'ils disaient, écoute Moncef Bensalem, Nourdine Bhiri, Lotfi Zitoun, Rached Ghannouchi, et franchement maintenant je suis fier de vivre dans un pays dans lequel il y a des gens comme eux avec un niveau académique et culturel (...) Pardonnez-moi les partisans d'Ennahdha, j'étais dans l'erreur à votre sujet, j'espère

que dieu me pardonnera, et maintenant, Ennahdha est la solution pour la Tunisie et que dieu soit avec eux.

On remarque aussi que les commentaires favorables au parti Ennahdha ou à des personnalités politiques d'Ennahdha sont souvent introduits par la phrase "Je suis pas pour Ennahdha mais", ce qui appuie l'idée d'une opinion objective sur le parti.

Je suis pas pour nahdha mais j'apprecie cette femme, mafamech 3lech n5afou m nahdha tawa cha3b tounsi fe9 w ken syesset nahdha mech bech te3jbou bech y3awed yahbet l chere3, lezem nchoufouhom ech bech ya3mlou w ba3d no7kmou 3lehom, mech mazelou mabdeou fi chay w a7na deja bdina nos fausses idées
➢ Je suis pas pour Ennahdha mais j'apprécie cette femme, il n'y a pas de raison d'avoir peur d'Ennahdha. Maintenant le peuple tunisien s'est réveillé et si la politique d'Ennahdha ne lui convient pas il redescendra dans la rue, on doit d'abord voir ce qu'ils vont faire et après on jugera. Ils n'ont rien commencé encore et nous on commence déjà avec nos fausses idées.

On note en parallèle un certain discours de victimisation accusant les "adversaires" d'Ennahdha de les critiquer gratuitement et de "s'acharner" sur eux, pour certains, ceci constitue la motivation première de leur soutien au parti.

yé5i innsé illi mitrach7in m3a ennahdha ilkolhom bhéyém ou ennsé lo5rin homa barka illi féhmin looooooooool
➢ Mais est-ce que tous les candidats aux élections d'Ennahdha sont bêtes et il n'y a que les autres seulement qui comprennent looooooooool
Koi ke vous fassiez je voterai pour nahdha ce n'est pas parce que Nahdha est meilleure que les autres mais parce que tous les partis ne font que critiquer Nahdha et donner des coups en dessous de la ceinture.

Les médias sont aussi accusés d'être partisans des adversaires d'Ennahdha.

ena milli 5raj bin 3li 9arrert nkoun contre ejjma3a elli bech ya3mlelhom el i3lèm ettounsi di3aya, w bech nenta5ab elli bech ychawhoulou som3tou!!! 5ater elli 3ach 3omrou yta77en l bin 3li, 3omrou ma yetsalla7 w ywalli mou7ayed!!!hètha w en plus nal9a elli program ennahdha bèhi w ma39oul baaaarcha, wettachwih wel kethb elli ta3mlou fih tawa, yzid y5allini nkabbech fiha akther 5ater houma akther jma3a nathlou w akther jma3a ndhaf!!!bèlekchi mèchi fi bèlkom cha3b tounis bhim bech ysadda9 ettachwih mta3kom wel icha3at????
Soyons clair, Il faut tout d'abord voir la totalité du débat. d'après ce que j'ai entendu, monsieur ghannouchi a repondu à toutes les questions d'une façon claire et précise. Je ne peux pas comprendre cet acharnement sur le parti Ennahdha et son chef. La Tunisie a besoin de nous tous, islamistes, marxcistes, laiques et pour se faire il faut que chacun de nous purifie son ame et fait un effort pour accepter l'autre. c'est la démocratie qu'on cherche tous

Dans les messages partisans d'Ennahdha, les commentateurs emploient souvent le titre "Cheikh" en référence à Rached Ghannouchi.

<commentaire=Avril_819>أيا هاو مناظر#e جديد#e بين علماني#e و مسلم . . . العلماني#e كبشت في القانون الأساسي لحرك#e النهض#e و كي ما لقاتش فيه حجج ضد النهض#e ولأ#e تتزعبن على فتاوى علماء السلطان و تقدمهم على أنهم من الإسلام . . . لوول و شوفوا رد الشيخ راشد . . . بدون تعليق فرق كبير برشا بصراح#e و متأكّد إليّ موزاييك ما صورتش كل شيء (تعرفوا علاش بالطبيع#e) هههههه برشا فضائح على العلماني#e §
➢ (…) regardez la réponse du Cheikh Rached… sans commentaire, la différence est très grande honnêtement (…)

Les références à la religion sont aussi très fréquentes dans les commentaires favorables à Ennahdha. Que ce soit pour revendiquer l'identité musulmane de la Tunisie ou pour accuser les adversaires d'Ennahdha d'être contre l'Islam.

fil 79i9a il madame hedi méhich contre chay5 rachid il ghannouchi ila n3izou barcha ;lékinha contre al islem wil 9oren ,point à la ligne. wlaw kenit inahda 7izb sghir w mouch ma3rouf bi cha3bitou il kbira rahi ma jebitouch wala bdet tistajwib fih ; lékin jebitou 5atir ta3rif annou

inahda bech tarba7 bi idni elléh 3ad bdew min tawa y7arbou féha ;9ol moutou bi gaydikom liannou acha3b moslim wa lan yastaslim lil 3ilménya
➢ En réalité cette dame n'est pas contre le Cheikh Rached Ghannouchi que j'adore ; mais elle est contre l'Islam et le Coran, point à la ligne (...) Je dis, crevez de dépi parce que le peuple est musulman et il ne capitulera pas devant la laïcité.

Les opposants au parti islamiste évoquent surtout leurs craintes de voir Ennahdha arriver au pouvoir. Certains l'accusent d'user d'un double discours pour arriver à ses fins.

G T présente au debat vraiment madame neyla a excellé ya3tiha sa77a !! ghanouchi na pas repondu à aucune question presque toujours le double discours , il se contredit à chaque fois !! il est pour la democratie alors que il trouve que manifester c hram , contredire el 7akém hram , puis il fait oui c le peuple qui va decider (contradiction totale)!! pour fuir les questions audacieuses de Mme Neyla il s'est mis à jouer la victime a7na 3amlouna w a9sawna et machin ..demagogie pur !!!! ce que nahdha essaye de faire une fois au pouvoir c la dictature irreversible le modele iranien !! pour les fans d'ennahdha arretz d'insulter les gens qui ne sont pas daccord avec vous réflechissez bien faites attention notre revolution était pour la dignité et la liberté matsalmouch tounes lélli yeswa welli mayéswéch !!

D'autres estiment que la question de la religion est un faux débat étant donné qu'ils estiment que "tous les Tunisiens sont musulmans" mais qu'en cas de victoire d'Ennahdha les répercussions sur l'économie, et notamment sur le tourisme, pourraient être néfastes.

Vraiment je trouve que les Tunisiens sont encore très limités, regardez les asiatiques comment ils se développent et prospèrent avec le travail et le travail, ils sont capables de tout fabriquer et de tout monter. Nous on est un peuple assisté, en plus avec Ghannouchi on va revenir au temps d'Aljazira Alarabya, et on ne fait que parler de réligion, un faux problème car tous les Tunisiens sont musulmans, je n'ai jamais vu un tunisien kafer ensuite Annahdha utilise la religion pour influencer les plus faibles, est ce que ghannouchi va résoudre les problèmes economiques de la Tunisie avec la religion, ya nass fikou ala rwahkom, il n'y aura pas de développement sans s'ouvrir sur le monde et encourager les investissements étrangers et sans developper le tourisme surtout qu'on n'a ni petrole ni gaz, Rahou Gannouchi a des idées limitées, il n'est pas ouvert, il

veut interdir à la femme de travailler pour résoudre le prob du chomage, il va interdit l'alcool et la il tue le tourisme, il a des interêts avec l'Iran et Alkaada, il a des financements louches, il a un double langage, aujourd'hui il a appelé Ghaddafi??.....etc. Enfin, soyez réalistes et non idéologues, le mélange entre la religion et la politique n'a jamais été la solution pour résoudre les problèmes? mais si vous insistez vous allez le payer cher, surtout que les intégristes quant ils prennent le pouvoir il ne le lâchent plus!!! moi je dis tjrs qu'on a ce qu'on mérite, et bien si on mérite de régresser on doit régresser!!!

Quelques commentaires revendiquent leur droit de critiquer Ennahdha en rappelant qu'il ne faut pas assimiler le parti à la religion :

<commentaire=Avril_586>للتذكير حزب النهض#e ليس الاسلامو الغنوشي ليس الرسول ! و هو معرض للنقد كيفي كيفاي حزب آخر . موش نحينا دكتاتور تجيبولنا دكتاتور آخر و تزيدو تألهوه و تعصموه مالخطأ ! فيقو بربي و حسو بتونس شوي#e و خفو عليها مالجهل متاعكم راهي تعبت برشا § emoticoneFrown

- Pour rappel, le parti Ennahdha n'est pas l'Islam, Ghannouchi n'est pas le prohète ! Il peut être critiqué

Enfin, d'autres commentateurs rappellent également qu'il faut dissocier entre Ennahdha et Islam mais sont aussi critiques envers les autres partis politiques :

<commentaire=Avril_817>اهم شي انه الي يصوت للنهض#e لا يصوت قال شنو#e خاطر يسمون انفسهم حزب اسلامي بل الإقتناع ببرامجهم و ايضا توضيح تاريخهم و مواقفهم . لا يجب ان ننسى ان الدين الإسلامي دين ديمقراطي يأمر بالشورى و انه لا يوجد اي انسان منزه عن الخطأ في الإسلام حتى اكبر شيخ و امام قابل للنقد و ممكن يغلط . يعني النهض#e كغيرها من الأحزاب قابل#e للنقد و الإختلاف . و ما لازمناش ننساو انه مثل كل الأحزاب كلهم يسعاو للسلط#e يعني فقط البرامج الإنتخابي#e الإقتصادي#e و الإجتماعي#e و السياس#e الخرجي#e . . . و الإنجازات على ارض الواقع بعد ذلك هي اساس الإختيار §

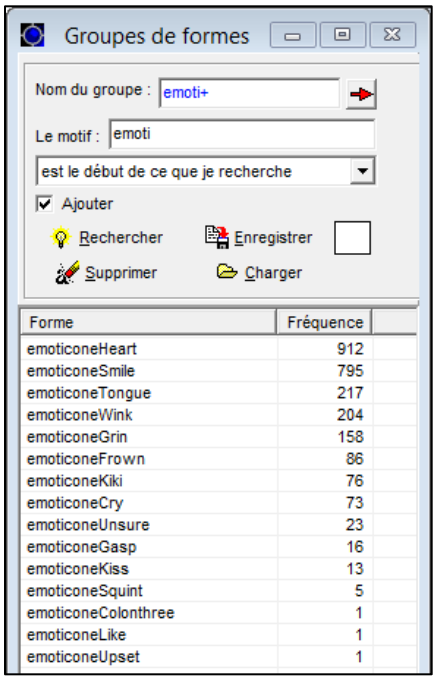
> ➢ Le plus important c'est que celui qui vote pour Ennahdha ne vote pas parce que soit disant ils se disent un parti islamique mais par conviction (…) on ne doit pas oublier que comme tous les partis, ils sont tous en quête de pouvoir (…)

Les références aux modèles d'autres pays et notamment du modèle turc rentrent dans le cadre du débat sur la laïcité (cf. 4.8.1.4).

3.11.5. Les émoticônes : expression non verbales des émotions

Les spécificités positives montrent que plus on avance dans l'année et plus les sujets des conversations semblent s'éloigner de la politique. A titre d'exemple nous choisissons d'étudier la distribution des occurrences des formes "emoticone" au fil des mois.

Si on observe les segments répétés et le dictionnaire, on se rend compte que les formes les plus fréquentes représentent des émotions positives. On peut notamment observer ces fréquences via les groupes de formes en lançant une recherche sur toutes les formes qui commencent par "emoti".



Forme	Fréquence
emoticoneHeart	912
emoticoneSmile	795
emoticoneTongue	217
emoticoneWink	204
emoticoneGrin	158
emoticoneFrown	86
emoticoneKiki	76
emoticoneCry	73
emoticoneUnsure	23
emoticoneGasp	16
emoticoneKiss	13
emoticoneSquint	5
emoticoneColonthree	1
emoticoneLike	1
emoticoneUpset	1

- Ventilation des occurrences des formes "emoticone" au fil des mois

Si on traite toutes les émoticônes comme un groupe de formes et qu'on observe la ventilation des occurrences, on voit très distinctement deux périodes dans l'année : de janvier à mai où les

émoticônes sont en spécificités négatives et de juin à octobre où les émoticônes sont en spécificités positives. Les mois de novembre et de décembre présentent des spécificités négatives mais ces deux parties sont à traiter avec précaution en raison de leurs écarts en nombre de commentaires avec le reste du corpus. Dans ces deux ensembles, les mois qui se distinguent le plus sont janvier et avril pour les spécificités négatives et août et septembre pour les spécificités positives.

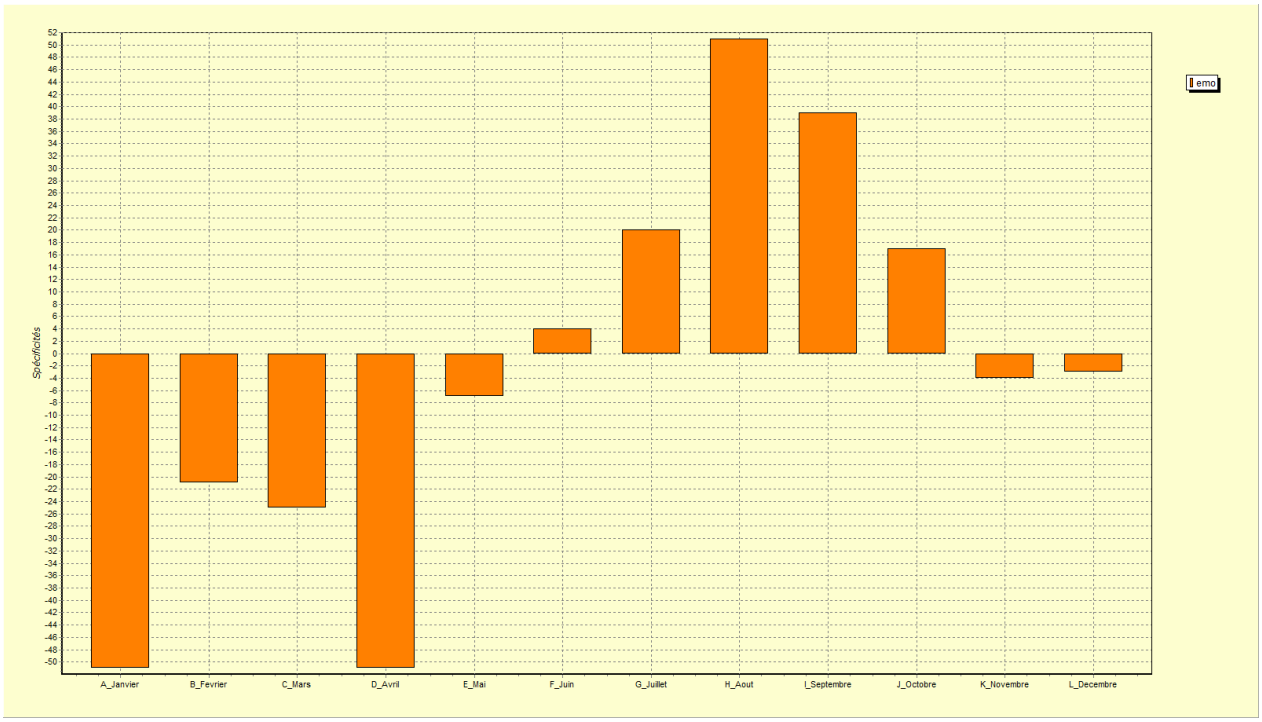


Figure 12 Ventilation des émoticônes par mois

Si on regroupe d'un côté les émoticônes associées à des émotions négatives (upset, frown, cry, unsure, gasp, squint) et de l'autre celles associées à des émotions positives (heart, smile, tongue, wink, grin, kiki, kiss, colonthree, like), la ventilation des occurrences montre que les émoticônes négatives sont en spécificité positive uniquement au mois d'août tandis que les émoticônes positives sont en spécificité positive aux mois de juillet, août, septembre et octobre. Pour le mois d'août, la barre représentant les émoticônes négatives dépasse de plusieurs points celle représentant les émoticônes positives. Pour les mois de l'année où les émoticônes, aussi bien négatives que positives, sont en spécificités négatives, on constate un très grand décalage entre les barres représentant les émoticônes négatives, qui se rapprochent du zéro, et celles représentant les émoticônes positives qui s'en éloignent beaucoup.

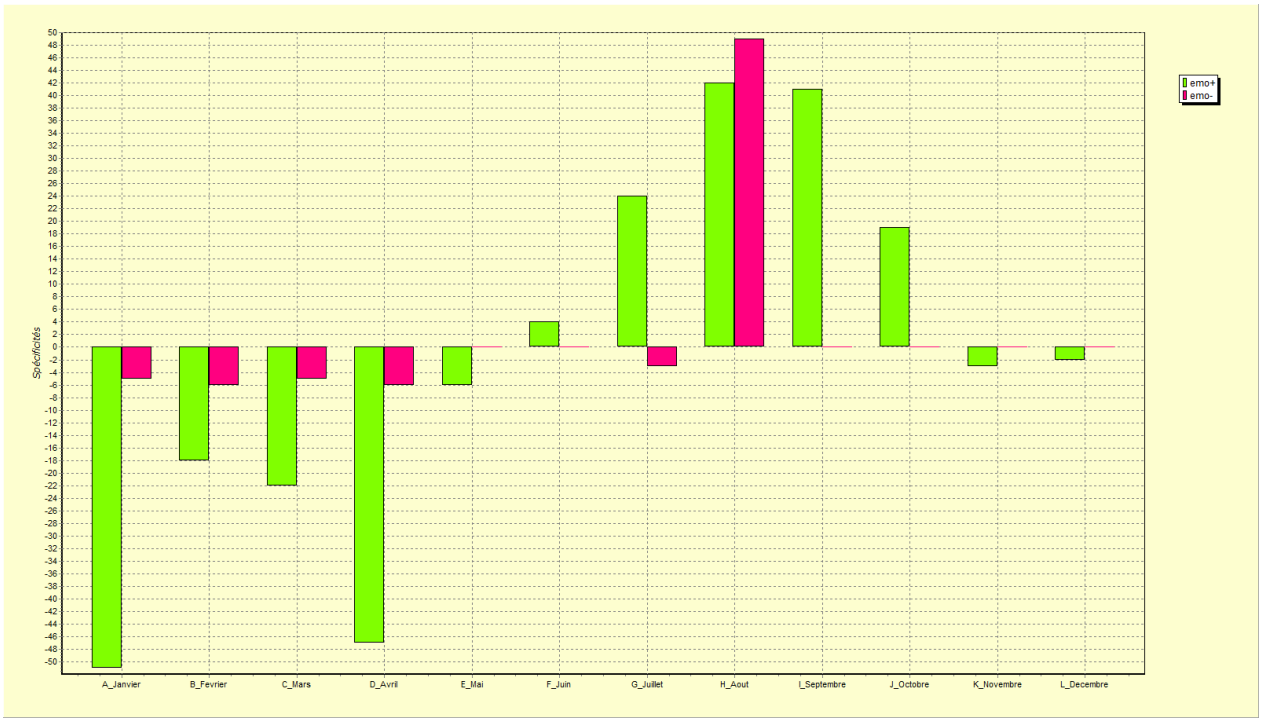


Figure 13 Ventilation des émotions positives Vs émotions négatives

Lorsque l'on observe ces émoticônes dans le détail, on voit que l'émoticône qui a le plus fort pic est l'émoticône "Cry" en spécificité positive au mois d'août. Le mois d'août présente également le plus d'émoticônes en spécificités positives en comparaison aux autres mois. L'émoticône "Heart" présente également un pic en spécificité positive, bien qu'inférieur à celui de l'émoticône "Cry", au mois de Juillet, ainsi que des pics en spécificité négative aux mois de janvier, février, mars, avril et mai, les valeurs les plus élevées étant enregistrées pour les mois de janvier et d'avril. L'émoticône "Smile" est en spécificité positive à partir du mois de juin, jusqu'au mois d'octobre avec un pic aux mois d'août et de septembre.

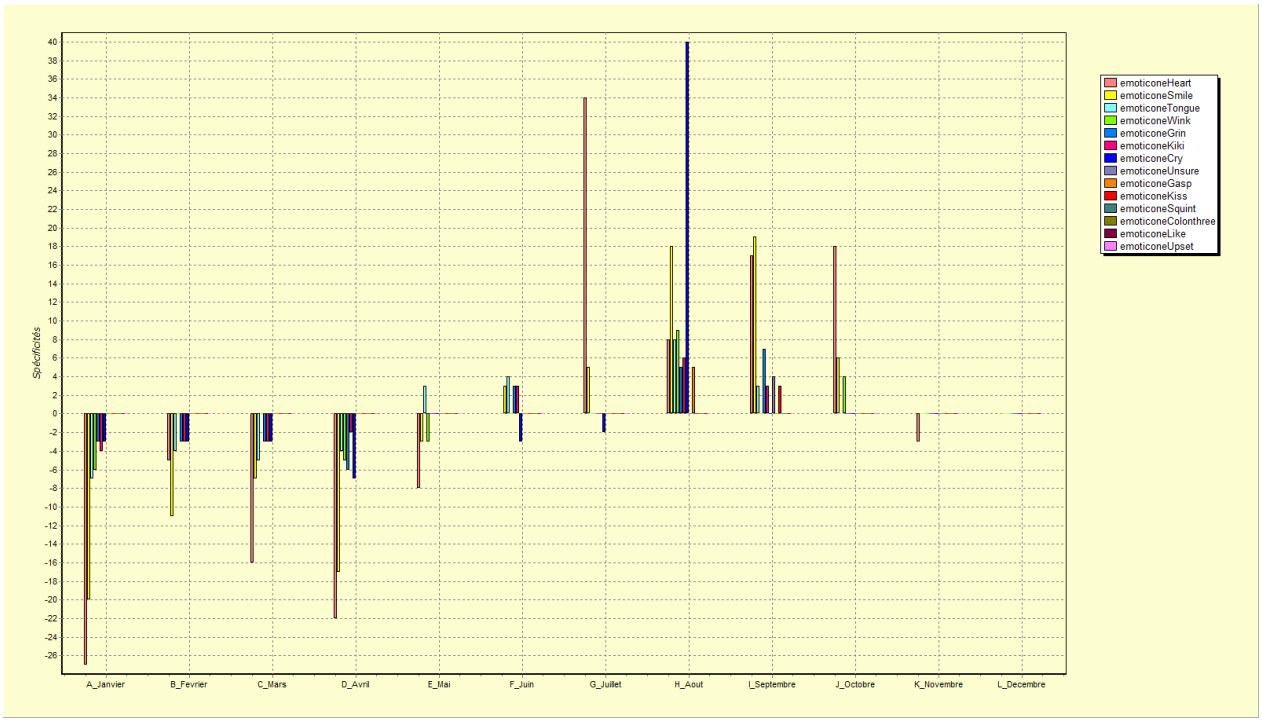


Figure 14 Ventilation des émoticônes

Cette ventilation est plutôt cohérente avec les événements marquants de chaque mois et les spécificités relevées pour les différentes parties. “emoticoneCry” est ainsi surtout employé en août ce qui correspond au décès d’un humoriste très populaire en Tunisie. Il est intéressant de constater que l’expression des émotions positives correspond à la deuxième période de l’année où les commentaires sont moins politiques. Alors que pour la première période de l’année les émotions positives présentent un pic dans les spécificités négatives.

Récapitulatif

- La textométrie réunit un ensemble d’approches mathématiques dans le but d’analyser des données textuelles numérisées réunies sous forme de corpus. Elle place le texte au centre de ses procédures et accorde une place fondamentale au retour au texte.
- Pour procéder à une analyse textométrique il est nécessaire de segmenter le texte (selon des caractères délimiteurs) et de le découper (en parties).
- Les domaines d’applications des méthodes textométriques sont nombreux. L’essor des réseaux sociaux et des sites communautaires sur internet offre de nouveaux lieux d’exploration pour les méthodes textométriques notamment pour l’analyse de la subjectivité.
- Sur notre corpus, les spécificités par mois renvoient à des événements marquants de l’actualité tunisienne de 2011. On y distingue, de plus, deux périodes dans l’année : l’une (de janvier à mai) est très politisée, l’autre (de juin à octobre) est non-politisée.
- La ventilation des émoticônes dans le corpus suit cette répartition politique/non politique
- Les cooccurrences montrent parfois un certain recoupement entre choix de système d’écriture et choix de langue, parfois non.

Conclusion et perspectives

La langue tunisienne à l'écrit, comme langue d'une culture numérique émergente particulièrement vivante en Tunisie, présente un intérêt à la fois linguistique, scientifique et sociétal.

Linguistique et scientifique, parce que cette écriture est le fruit d'un brassage culturel et linguistique au croisement de l'arabe et du français et qui présente un véritable défit quant à sa normalisation. Ce travail a été l'occasion de se confronter concrètement aux difficultés qui peuvent émerger de l'étude d'un corpus de textes écrits dans une langue peu dotée, le tunisien. Les questions soulevées sont certes plus nombreuses que les réponses apportées mais cela permet de réfléchir à de nombreuses pistes d'amélioration.

La normalisation, d'abord, semble essentielle pour une analyse quantitative plus pertinente. En effet, la variabilité des graphies nous a obligé à souvent quitter trop vite l'analyse quantitative pour procéder à une analyse qualitative. Mais cette normalisation devra tenir compte de la complexité de la situation linguistique tunisienne. Il semblerait en effet que souvent le choix du système d'écriture, plus que celui de la langue, est significatif dans le regroupement des opinions exprimées dans les commentaires. Mais le multilinguisme est une réalité tunisienne qu'il n'est pas possible de gommer, langues et systèmes d'écritures pouvant se mélanger dans un même commentaire. Ce double niveau de variations (langue/système d'écriture) oblige donc à se poser des questions aussi bien pour la segmentation des textes que pour le découpage du corpus.

Ensuite, ce corpus permet d'attirer l'attention sur la fragilité des données sur le web. L'état de l'art du traitement automatique du tunisien étant pauvre et les travaux semblant peu enclins à s'émanciper du schéma diglossique arabe-tunisien, le tunisien continue à être une langue peu, voire pas dotée. Il n'existe pas de fonds tunisien pour le TAL car cette langue souffre encore de sa catégorisation comme "dialecte oral". Pourtant les données textuelles en tunisien sont nombreuses et accessibles sur le web. Seulement, elles sont fragiles, car n'étant pas préservées sous forme de ressources, elles sont constamment menacées de disparition, rien n'étant plus facile que de "supprimer" des données sur une page web.

Sociétal, enfin, parce que l'écriture tunisienne, notamment sur les réseaux sociaux, offre une clé de lecture incontournable pour mieux comprendre les sujets qui animent la société tunisienne. Ces textes en tunisien, et à travers eux le développement d'un véritable TAL du tunisien, permettraient

non seulement de préserver ces productions et les idées qu'elles véhiculent mais aussi d'apporter un éclairage de l'intérieur sur une actualité riche qui a marqué l'Histoire contemporaine et qui est trop souvent commentée de l'extérieur par des observateurs internationaux coupés de la réalité des premiers acteurs de ces événements, les citoyens tunisiens.

Bibliographie

AMOSSY, Ruth et BURGER, Marcel. Introduction : la polémique médiatisée.Semen. Revue de sémio-linguistique des textes et discours, 2011, no 31, p. 7-24.

BACCOUCHE, Taïeb. La langue arabe dans le monde arabe. L'Information Grammaticale, 1998, vol. 2, no 1, p. 49-54.

BEN ACHOUR, Yadh. Les implications politiques du problème linguistique au Maghreb. - Paris, La Pensée, n° 303, IRM, 1995.- p.p. 93-102.

BOUCHLAGHEM, Rihab et ELKHLIFI, Aymen. Tunisian dialect Wordnet creation and enrichment using web resources and other Wordnets. ANLP 2014, 2014, p. 104.

BOUJELBANE, Rahma, KHEMEKHEM, Mariem Ellouze, et BELGUITH, Lamia Hadrich. Mapping rules for building a tunisian dialect lexicon and generating corpora. In: Proceeding of International Joint Conference on Natural Language Processing (IJCNLP), Nagoya, Japan. 2013.

BOWKER, Lynne et PEARSON, Jennifer. Working with specialized language: a practical guide to using corpora. Routledge, 2002.

BOWKER, Lynne et PEARSON, Jennifer. Working with specialized language: a practical guide to using corpora. Routledge, 2002.

BRAS, J. P. Internet au Maroc et en Tunisie. M. MEZOUAGHI, Le Maghreb dans l'économie numérique, Paris, Maisonneuve & Larose, 2007, p. 161-180.

CHAÏB, Mohammed. Al-'arabiyya al-wusta (l'arabe intermédiaire).- Tunis, Revue Tunisienne des Sciences Sociales, n° 46, Publication du CERES, 1976.- p.p. 47-66

CHOUIKHA, Larbi. L'audiovisuel en Tunisie : une libéralisation fondue dans le moule étatique. L'Année du Maghreb, 2007, no II, p. 549-558.

CLA2T, Université Sorbonne Nouvelle.

DUCOS, Alexia, BONNET, Valérie, MARCHAND, Pascal, et al. Classification d'un corpus hétérogène : la page Facebook de soutien au « bijoutier de Nice » (septembre 2013).

EENSOO, Egle et VALETTE, Mathieu. Sur l'application de méthodes textométriques à la construction de critères de classification en analyse des sentiments. In : TALN 2012. GETALP-LIG, 2012. p. 367-374.

ELIMAM, Abdou. Choix de modèle de développement et glottopolitique.Langages, 1986, p. 75-85.

ELIMAM, Abdou. Du Punique au Maghribi: Trajectoires d'une langue sémito-méditerranéene'. Synergies Tunisie, 2009, no 1, p. 25-38.

ELIMAM, Abdou. Le maghribi, langue trois fois millénaire, éd. ANEP, Alger, 1997.

ELIMAM, Abdou. Le maghribi, vernaculaire majoritaire à l'épreuve de la minoration, 2012.

FERGUSON, Charles Albert. Diglossia. Word-Journal of the International Linguistic Association, 1959, vol. 15, no 2, p. 325-340.

GEISSER, Vincent et GOBE, Éric. Un si long règne… Le régime de Ben Ali vingt ans après. L'Année du Maghreb, 2008, no IV, p. 347-381.

GIBSON, Maik. Dialect levelling in Tunisian Arabic: towards a new spoken standard. Language Contact and Language Conflict Phenomena in Arabic, 2002, p. 24-40.

GRAJA, Marwa, JAOUA, Maher, et BELGUITH, Lamia Hadrich. Building Ontologies to Understand Spoken Tunisian Dialect. arXiv preprint arXiv:1109.0624, 2011.

HABASH, Nizar, DIAB, Mona T., et RAMBOW, Owen. Conventional Orthography for Dialectal Arabic. In: LREC. 2012. p. 711-718.

HABERT, Benoît, NAZARENKO, Adeline, et SALEM, André. Les linguistiques de corpus. Colin, 1997.

HADJ-SALAH A. (1978) : Linguistique arabe et linguistique générale. Thèse d'État.

HAMDI, Ahmed, BOUJELBANE, Rahma, HABASH, Nizar, et al. Un système de traduction de verbes entre arabe standard et arabe dialectal par analyse morphologique profonde. In : Traitement Automatique des Langues Naturelles. 2013. p. 396-406.

HASSOUN, Mohamed. Les nouveaux défis du TAL Exploration des médias sociaux pour l'analyse des sentiments : Cas de l'Arabish, 2014.

LAROUSSI, Foued. La diglossie arabe revisitée. Quelques réflexions à propos de la situation. Insaniyat/.إنسانيات Revue algérienne d'anthropologie et de sciences sociales, 2002, no 17-18, p. 129-153.

LECOMTE, Romain. Internet et la reconfiguration de l'espace public tunisien : le rôle de la diaspora. tic&société, 2009, vol. 3, no 1-2.

LECOMTE, Romain. Révolution tunisienne et Internet : le rôle des médias sociaux. L'Année du Maghreb, 2011, no VII, p. 389-418.

LEECH, Geoffrey. New resources, or just better old ones? The Holy Grail of representativeness.Language and Computers, 2006, vol. 59, no 1, p. 133-149.

MARCELLESI, Jean-Baptiste. De la crise de la linguistique à la linguistique de la crise : la sociolinguistique in Langages et sociétés. Pensée (La) Paris, 1980, no 209, p. 4-21.

MAYAFFRE, Damon. Les corpus réflexifs : entre architextualité et hypertextualité. Corpus, 2002, no 1.

MCENERY, Tony et WILSON, Andrew. Corpus linguistics: An introduction. Edinburgh University Press, 2001.

MEJRI, Salah, SAID, Mosbah, et SFAR, Inès. Pluringuisme et diglossie en Tunisie. Synergies Tunisie n, 2009, vol. 1, p. 53-74.

MEJRI, Salah. 1. Le français en/de Tunisie ? LE FRANÇAIS EN AFRIQUE, 2012. p. 219.

MONIÈRE, Denis et LABBÉ, Dominique. Un siècle et demi de discours gouvernemental au Canada Contribution de la lexicométrie à l'Histoire politique. In: 12th International Conference on Textual Data Statistical Analysis. 2014. p. 485-494.

PINCEMIN, Bénédicte. Sémantique interprétative et textométrie–Version abrégée. Corpus, 2011, no 10, p. 259-269.

POUESSEL, Stephanie. Les marges renaissantes : Amazigh, Juif, Noir. Ce que la révolution a changé dans ce « petit pays homogène par excellence » qu'est la Tunisie. L'Année du Maghreb, 2012, no VIII, p. 143-160.

RASTIER, François. Enjeux épistémologiques de la linguistique de corpus. La linguistique de corpus. Presses Universitaires de Grenoble, 2005.

SALEM A. et al. (2003), Lexico 3 – Outils de statistique textuelle. Manuel d'utilisation, Syled-

SALEM, André. Approches du temps lexical [Statistique textuelle et séries chronologiques]. Mots, 1988, vol. 17, no 1, p. 105-143.

SALEM, André. Les séries textuelles chronologiques. Histoire & mesure, 1991, vol. 6, no 1, p. 149-175.

SINCLAIR, John. Preliminary recommendations on corpus typology. EAGLES Document TCWG-CTYP/P (available from http://www. ilc. pi. cnr. it/EAGLES/corpustyp/corpustyp. html), 1996.

TOUATI, Zeineb. La Révolution tunisienne : interactions entre militantisme de terrain et mobilisation des réseaux sociaux. L'Année du Maghreb, 2012, no VIII, p. 121-141.

VALETTE, Mathieu et EENSOO, Egle. Approche textuelle pourle traitement automatique du discours évaluatif. Langue française, 2014, vol. 184, no 4, p. 109-124.

VANHOVE, Martine. La dialectologie du maltais et son histoire. In : Revue d'Ethnolinguistique. Diasystème et longue durée (Cahiers du Lacito). Catherine Paris éd. 1999. p. 171-191.

VANHOVE, Martine. La langue maltaise : un carrefour linguistique. Revue du monde musulman et de la Méditerranée, 1994, vol. 71, no 1, p. 167-183.

VANHOVE, Martine. Un marqueur polysémique en maltais : għad (/° ad/). In : Bulletin de la société de linguistique de Paris. 1997. p. 269-293.

ZRIBI, Ines, BOUJELBANE, Rahma, MASMOUDI, Abir, et al. A Conventional Orthography for Tunisian Arabic. In: Proceedings of the Language Resources and Evaluation Conference (LREC), Reykjavik, Iceland. 2014.

ZRIBI, Inès, GRAJA, Marwa, KHMEKHEM, Mariem Ellouze, et al. Orthographic transcription for spoken tunisian arabic. In: Computational Linguistics and Intelligent Text Processing. Springer Berlin Heidelberg, 2013. p. 153-163.

ZRIBI, Inès, KHEMAKHEM, M. Ellouze, et BELGUITH, Lamia Hadrich. Morphological Analysis of Tunisian Dialect. In : International Joint Conference on Natural Language Processing. 2013. p. 992-996.