Combinaison d'étiqueteurs morphosyntaxiques, de lexiques flexionnels et de marqueurs de glose pour détecter les néologismes en français du Burkina

Célestin Zoumbara[1], Mathieu Roche[2,3], Sascha Diwersy[4], Youssouf Ouédraogo[1], Pierre Martin[2,5]

[1] Université Joseph Ki-Zerbo, Burkina Faso – celestin.zoumbara@cirad.fr; youed89@yahoo.fr

[2] CIRAD, F-34398 Montpellier, France – mathieu.roche@cirad.fr

[3] TETIS, Université de Montpellier, AgroParisTech, CIRAD, CNRS, Inrae, France – mathieu.roche@teledetection.fr

[4] Université Paul-Valéry Montpellier 3, France – sascha.diwersy@univ-montp3.fr

[5] AIDA, Université Montpellier, CIRAD, France – pierre.martin@cirad.fr

Abstract

The French language of Burkina is marked by national multilingualism. In order to extract automatically neologisms from textual data, we have developed the Extranéo approach. This approach integrates Automatic Natural Language Processing (NLP) tools to detect formal neologisms and their contexts of use. Extranéo combines a morphosyntactic labeler and an inflectional lexicon to identify candidate neologisms, which are then validated manually. The use of gloss patterns finally allows access to the contexts of validated neologisms. In this paper, four morphosyntactic labelers, five inflectional lexicons, and gloss markers are evaluated in order to detect the most efficient ones. The results show that on newspaper articles, TreeTagger obtains the highest F-measure, i.e. 0.86, for labelling. To identify candidate neologisms, the DELA and Morphalou lexicons obtain the highest F-measure, i.e. 0,52. Highlighting the context of the validated neologisms reveals that the defining context has the highest accuracy, i.e. 0.49, ahead of the naming (0.22) and equivalence (0.24) contexts.

Keywords: Extranéo – NLP – neologisms – Burkina - gloss

Résumé

Le français du Burkina est marqué par le multilinguisme national. Dans l'objectif d'extraire automatiquement des néologismes sur des données textuelles, nous avons développé l'approche Extranéo. Celle-ci intègre des outils de Traitement automatique du langage naturel (TALN) pour détecter les néologismes formels et leurs contextes d'utilisation. Extranéo combine un étiqueteur morphosyntaxique et un lexique flexionnel pour identifier les néologismes candidats, validés ensuite manuellement. L'usage de patrons de glose permet enfin d'accéder aux contextes des néologismes validés. Dans cet article, quatre étiqueteurs morphosyntaxiques, cinq lexiques flexionnels et des marqueurs de glose sont évalués afin de détecter les plus performants. Les résultats montrent que sur les articles de journaux, l'étiqueteur TreeTagger obtient la plus forte F-mesure, 0.86 en matière d'étiquetage. Sur l'identification des néologismes candidats, les lexiques DELA et Morphalou obtiennent la F-mesure la plus élevée, 0.52. La mise en relief des contextes des néologismes validés révèle que le contexte définitoire détient la précision la plus élevée, à savoir 0.49, devant les contextes de dénomination (0.22) et d'équivalence (0.24).

Mots clés : Extranéo – TALN – néologismes – Burkina - glose

1. Introduction

À l'instar des autres pays de l'Afrique francophone caractérisés par le multilinguisme, le contexte sociolinguistique du Burkina est marqué par le français, utilisé comme langue officielle, et cinquante-neuf langues nationales (Kedrebéogo et al., 1988). Cette cohabitation entraîne une variation diatopique du français (Coseriu, 1964) se rapportant à l'espace

géographique du Burkina tout comme l'écrit Frey (2004) au sujet du Burundi, du Gabon, de la Côte-d'Ivoire, etc. Le français qui en résulte est marqué de diatopismes (Poirier, 1995) appelés burkinabismes au Burkina (Ouédraogo, 2019). Les travaux de l'Équipe de l'Inventaire des particularités lexicales du français en Afrique noire (IFA, 2004) et de Prignitz (1996) montrent que dans le français du Burkina, les diatopismes se manifestent sur le plan lexical sous forme de néologismes. Le néologisme est un mot nouveau (Sablayrolles, 2002). Lorsque la nouveauté porte sur la forme du mot, le néologisme est dit formel. Il est dit sémantique lorsqu'un mot existant déjà dans la langue est revêtu d'un nouveau sens.

Dans les écrits du Burkina, les néologismes sont manifestes dans la littérature (Ouédraogo, 2000), la presse écrite (Raschi, 2009 ; Koama, 2015 ; Ouédraogo, 2019), etc. Les études qui ont visé à les extraire ont exclusivement été conduites de façon manuelle. Ce mode opératoire n'a pas permis de prendre en compte un vaste corpus. Qu'elles aient porté sur des œuvres littéraires ou sur la presse écrite, ces études ont été conduites sur un corpus restreint. Raschi (2009) a limité son étude aux publications de septembre 2006 à mars 2007 de *Lefaso.net* et l'étude de Ouédraogo (2000) a porté sur les œuvres *Rougbêinga* de Norbert Zongo (262 pages) et *Papa oublie-moi* de Jean-Pierre Guingané (79 pages). Ces travaux ont permis de décrire le français du Burkina à travers les néologismes utilisés dans les corpus étudiés, mais n'ont pas permis de conclure de façon générale sur le recours aux néologismes, de déterminer leurs contextes et de s'intéresser à leur cycle de vie pour cerner la vitalité de la langue. Ce cycle de vie s'exprime en termes d'émergence, de diffusion et d'une éventuelle lexicalisation ou adoption (Cartier, 2019). Pour réaliser cela, l'enjeu est de disposer d'une plateforme de détection automatique de néologismes à partir d'un vaste corpus de données textuelles burkinabè s'étalant sur plusieurs années. Dans la perspective de combler ce vide, nous développons une approche appelée Extranéo (EXTRaction Automatique de NÉOlogismes), qui s'inspire de la méthode de Néoveille (Cartier, 2019). L'approche Extranéo se base sur l'utilisation d'outils du Traitement automatique du langage naturel (TALN) pour extraire les néologismes formels dans la presse écrite burkinabè en ligne. Elle vise à constituer une base de données lexicales caractéristiques du français du Burkina. Cette base permettra d'appréhender la vitalité de ce français et le cycle de vie des néologismes. Extranéo combine un étiqueteur morphosyntaxique, un lexique flexionnel et des marqueurs de glose. Tout cela est coordonné par des programmes informatiques développés à l'aide du langage de programmation Perl (Wall et al., 2001).

Pour mettre en œuvre cette approche, l'étape préalable consiste à évaluer les outils existants susceptibles d'être utilisés dans le processus. Cela constitue une des contributions de cet article. Il présente une évaluation d'étiqueteurs, de lexiques et de marqueurs de glose sur différentes données textuelles. L'objectif principal est de repérer l'étiqueteur ou les étiqueteurs et le (s) lexique (s) les plus performants pour détecter les néologismes formels et pour mettre en relief les contextes de ces néologismes à partir d'un corpus journalistique. Trois sections constituent l'ossature de cet article. La première section présente un état de l'art sur les diatopismes du français d'Afrique noire francophone et sur les méthodes de détection automatique des néologismes formels. La deuxième section décrit Extranéo ainsi que les différents outils mobilisés pour l'évaluation. La troisième section présente les résultats et la discussion.

2. État de l'art

Sur la base des diatopismes du français d'Afrique noire francophone, cette section décrit la typologie des néologismes formels dans le français du Burkina. Elle présente également un état de l'art des méthodes de détection automatique des néologismes formels.

2.1. Le français d'Afrique noire francophone

2.2.1. Typologie des diatopismes

Les diatopismes du français d'Afrique noire francophone ont fait l'objet d'un inventaire élaboré par l'Équipe de l'Inventaire des particularités lexicales du français en Afrique noire (IFA) coordonnée par Racelle-Latin D. Cet inventaire, dont la dernière édition date de 2004, a concerné douze pays dont le Burkina. Sur la base de ces travaux, Ouédraogo (2008) identifie trois catégories de diatopismes. Les diatopismes lexématiques renvoient à des constructions nouvelles, à des calques ou à des emprunts (*abacos* pour « à bas le costume » qui est un modèle de chemise pour homme). Les diatopismes sémantiques renvoient à des sens nouveaux donnés à des mots existants, soit par transfert, soit par extension sémantique ou par métaphorisation (*chinoiserie* pour qualifier « tout produit qui a l'air luxueux mais qui n'est pas de bonne qualité »). Les diatopismes grammaticaux se manifestent à travers les changements de classe grammaticale (*façon,* utilisé comme un adjectif qualificatif et non un nom commun) ou les constructions syntaxiques (*fréquenter et frotter*, verbes transitifs en français de référence, utilisés de façon intransitive).

À partir de cette catégorisation, les régionalismes sont différenciés des particularismes. Les régionalismes sont des faits linguistiques propres à la région d'Afrique noire considérée, qui présentent la double spécificité de n'être usités que dans cet espace géographique tout en possédant un équivalent en français de référence (*toubab, aujourd'hui nuit* qui sont respectivement des équivalents de « Blanc » et « cette nuit »). Les particularismes désignent des faits linguistiques créés sans équivalents en français de référence et qui peuvent s'étendre dans tout l'espace francophone d'Afrique noire (*attiéké, banco, parenté à plaisanterie,* etc.). Le recours à cette seconde catégorie permet de partager une réalité, tout en évitant de recourir à de longues paraphrases, généralement approximatives.

2.2.2. Typologie des néologismes formels dans le français du Burkina

Concernant la typologie des néologismes formels contenus dans le français du Burkina, Prignitz (1993), distingue trois principaux procédés de création. Le premier considère les formations internes locales via un changement de catégorie grammaticale (*moyen,* qui est un nom, est employé comme un verbe), une abréviation devenant la forme usuelle (*à plus tard* devenant *à plus, paludisme* devenant *palu*), un redoublement (*chaud-chaud*, formé de *chaud + chaud*), une dérivation (*circonciseur*, formé de *circoncis- + -eur*) ou une composition (*aller prendre jeter,* formé de *aller + prendre + jeter*). Le second concerne les formations hybrides, scindées par Kéita (2013) en hybridation par composition (*mossicratie*, formé de *mossi- + -cratie*) et en hybridation par dérivation (*bwamufier*, formé de *bwamu- + -fier*). Le dernier procédé considère les emprunts aux langues non africaines telles l'anglais (*wanted*) ou le portugais (*tapade, lougan*), aux langues africaines non locales telles l'arabe (*doua*) et le wolof (*zaki*), aux langues africaines locales comme le dioula (*yougou-yougou*), le moré (*naaba*) et le foulfouldé (*gniiwa*), et enfin aux calques linguistiques (*demander la route*).

Les études portant sur les néologismes dans le français du Burkina ont été conduites manuellement. Toutefois, il existe des plateformes permettant la détection automatique de néologismes formels, principalement en s'appuyant sur des corpus journalistiques.

2.2. Méthodes de détection automatique des néologismes

Cartier et Sablayrolles (2009) signalent que l'usage de l'informatique dans le processus de détection des néologismes « offre des possibilités de traitement bien plus grandes et systématiques ». Deux principales approches se dégagent, d'après Cartier (2016). La première approche utilise la méthode dite « dictionnaire de référence ou d'exclusion » (Cartier et al., 2018). Cette méthode consiste à utiliser un corpus d'exclusion ou de référence qui peut être une ressource lexicographique, c'est-à-dire un répertoire comportant les mots attestés dans une période antérieure (Falk et al., 2014), pour repérer dans un corpus tous les mots inconnus, avant de mettre en œuvre divers filtres afin d'identifier des néologismes candidats. La seconde approche repose sur l'usage d'un étiqueteur morphosyntaxique entraîné au moyen d'une méthode d'apprentissage automatique supervisée, basée sur un corpus annoté manuellement. Cette approche est mobilisée par TreeTagger (Schmid, 1994) pour être adaptée à différentes langues. Elle ne nécessite pas l'utilisation d'une ressource lexicographique, mais est limitée pour reconnaître les mots composés.

La plateforme Neoveille (Cartier, 2016) combine ces deux approches. Tout d'abord, l'étiqueteur TreeTagger (Schmid, 1994) détecte les mots inconnus. Les mots commençant par une majuscule sont ensuite supprimés et le correcteur orthographique Hunspell est utilisé pour éliminer les erreurs typographiques et les fautes. Enfin, intervient l'utilisation des listes d'exclusion, qui sont continuellement enrichies grâce aux résultats de la validation manuelle. Une des limites de cette approche, selon Cartier (2016), est qu'il est très difficile de disposer d'une ressource lexicographique à jour quelle que soit la langue considérée. Cartier et al. (2018) présentent les résultats d'une détection réalisée par Neoveille sur un corpus constitué d'articles de journaux de cinq pays dont deux pays africains, à savoir l'Algérie et le Sénégal. Ce travail a permis de montrer la prédominance du français métropolitain qui a concentré 83% des occurrences des néologismes détectés. Il n'a pas permis de caractériser le français d'Afrique noire francophone.

L'approche Extranéo s'inspire de la méthode de Néoveille (Cartier, 2019). Elle est fondée sur l'utilisation d'un étiqueteur et d'un lexique flexionnel utilisé comme corpus d'exclusion. Ce qui distingue les deux approches est que dans l'approche de Néoveille, le corpus d'exclusion est enrichi de manière permanente et des règles sont utilisées pour supprimer les mots commençant par une majuscule, les erreurs typographiques et les fautes afin de filtrer les néologismes candidats. L'apport d'Extranéo par rapport à Néoveille est qu'elle permet d'extraire non seulement les néologismes, mais aussi leurs contextes en ayant recours à la glose. Celle-ci renvoie à des commentaires en situation parenthétique encadrant l'insertion du néologisme, souvent introduits par des marqueurs tels que *appelé*, *c'est-à-dire*, *ou,* etc. (Mela et al., 2011). Le contexte d'un mot, dans la perspective distributionnelle de Harris (1968), renvoie à l'ensemble des mots qui entretiennent avec lui une relation de dépendance sémantique ou syntaxique (Morlane-Hondère, 2013).

3. Approche d'identification de néologismes

Dans cette section, nous décrivons tout d'abord les corpus utilisés (section 3.1.) pour évaluer l'approche Extranéo, décrite en sections 3.2. et 3.3. Nos propositions reposent sur une combinaison de différents outils qui est présentée dans sa globalité (section 3.2.) puis de manière détaillée (section 3.3.).

3.1. Corpus

3.1.1. Constitution du corpus

Deux corpus ont été constitués pour réaliser l'évaluation de nos propositions. Le premier corpus (Tableau 1), utilisé pour évaluer les étiqueteurs et les lexiques, comporte quatre

différents types de textes : des articles de journaux, des commentaires d'internautes sur des articles de journaux, des messages via le réseau social Facebook et des textes littéraires. Cette volonté de diversifier les genres textuels constitutifs du corpus vise à montrer la généricité de l'approche. Ce premier corpus a été constitué manuellement dans l'objectif d'évaluer le comportement des outils sur les néologismes. Il est composé de 132 énoncés comportant 1183 mots, dont 144 néologismes. Nous sommes partis des emplois contextuels des néologismes dans chaque type de texte et avons sélectionné l'énoncé ou le paragraphe dans lequel le néologisme est employé. Dans ce premier corpus, les burkinabismes sur le plan syntaxique concernent la conversion des noms propres en noms communs ou antonomase (Lafage, 1977), la substitution et l'effacement des prépositions (Gandon, 1994), l'omniprésence du déictique – là (Raschi, 2009), l'utilisation de verbes transitifs sous forme intransitive (Ouédraogo, 2008) et le défigement des expressions figées (Ouédraogo, 2019). Dans les commentaires, par exemple, nous relevons « Les zeph sont entrain » au lieu de « zeph et ses camarades sont entrain ». Dans les messages Facebook, nous relevons « Face à ces foutages de gueule » au lieu de « Devant ces foutages de gueule ». Dans les articles, nous relevons « La Ouagattitude est là pour nous rappeler » au lieu de « La Ouagattitude nous rappelle » et « Mystère et boule de Zorgho ! » au lieu de « Mystère et boule de gomme ! » Sur le plan lexical, les burkinabismes sont manifestes dans le corpus à travers le redoublement, l'hybridation et l'emprunt (Prignitz, 1993). Nous relevons, ainsi, « comment-comment » et « nassaréen » dans les articles.

Le second corpus, utilisé pour évaluer et guider l'extraction des contextes des néologismes, a été téléchargé de façon automatique sur les sites des journaux burkinabè en ligne *L'Observateur paalga* (https://www.lobservateur.bf), *Sidwaya* (https://www.sidwaya.info) et *Lefaso.net* (https://www.lefaso.net). Ce second corpus rassemble des articles produits en 2011 et en 2012. Il s'agit d'un total de plus de 18 000 articles correspondant à environ 15 millions de mots.

Pour évaluer les étiqueteurs et les lexiques, le premier corpus a été étiqueté manuellement puis automatiquement. La description de cette démarche est donnée en section 3.1.2.

Tableau 1- Description du corpus pour l'évaluation des étiqueteurs et des lexiques

Type de texte	**Source**	**Période**	**Nombre d'énoncés**	**Nombre de mots**	**Nombre de néologismes**
Article de journal	- *Journaldujeudi.bf* - *Lefaso.net* - *Fasozine.com* - *Sidwaya.bf*	2015 à 2018	38	342	43
Commentaire	- *Lefaso.net*	2018	34	283	37
Message Facebook	- *Communauté Dafing-Bobo-Peulh* - *La voix des professeurs* - *Pages Facebook individuelles*	2018	47	309	49
Texte littéraire	- *Le Parachutage* - *La défaite du yargha* - *Le triomphe de l'amour* - *La traversée nocturne* - *Lucia ou le bout du tunnel*	1962-2015	13	249	15
TOTAL			132	1183	144

3.1.2. Étiquetage manuel et automatique du corpus de référence

Le premier corpus a d'abord été étiqueté manuellement. Nous nous sommes appuyés sur le guide d'annotation du référentiel *Universal Dependencies*[1] (UD) pour effectuer l'étiquetage manuel de ce corpus. Pour chaque type de texte, nous avons d'abord segmenté le corpus en plaçant chaque mot sur une ligne avant de lui attribuer une étiquette selon les critères morphologique et syntaxique de UD en matière d'étiquetage. Suite à cette phase manuelle, le corpus a été étiqueté automatiquement par les différents étiqueteurs, dont les étiquetages ont alors été comparés. Sur la base de ces résultats, le second corpus a été étiqueté automatiquement à l'aide du TreeTagger. Chaque mot de ce corpus a été doté de quatre propriétés : *word* renvoie au mot-forme, *frlemma* renvoie au lemme, c'est-à-dire la forme canonique du mot, *frttpos* est l'étiquette attribuée par TreeTagger et *frudpos* est la correspondance de l'étiquette de TreeTagger dans le référentiel UD.

3.2. Processus global

Le processus descriptif du fonctionnement d'Extranéo est présenté en Figure 1. Il comporte quatre principales étapes : l'étiquetage morphosyntaxique du corpus (étape 1), l'identification des néologismes candidats (étape 2), l'annotation manuelle (étape 3) et la mise en relief des contextes (étape 4). À l'étape 1, nous utilisons un étiqueteur morphosyntaxique pour attribuer une catégorie grammaticale à chaque mot du corpus. À l'étape 2, les mots, dont les catégories grammaticales sont caractéristiques des néologismes, à savoir les adjectifs qualificatifs, les adverbes, les noms communs et les verbes (Cartier et al., 2018), sont extraits et recherchés dans le lexique pour identifier les néologismes candidats. Les néologismes candidats sont validés par les utilisateurs à l'étape 3 en distinguant les « non néologismes » des « néologismes validés ». Des patrons de glose sont enfin projetés à l'étape 4 sur les données textuelles pour mettre en relief les contextes des « néologismes validés ».

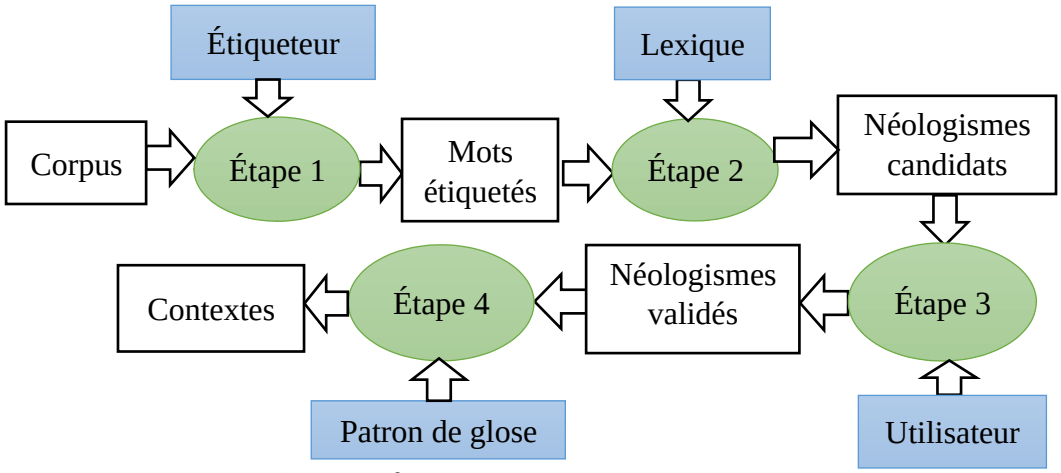

Figure 1 - Processus d'Extranéo. Les ovales représentent les traitements, les rectangles bleus renvoient aux outils utilisés et les rectangles blancs renvoient aux entrées/sorties.

3.3. Description des étapes du processus nécessitant un traitement automatique

Le processus comporte trois étapes qui reposent sur un traitement automatique (étapes 1, 2 et 4 - cf. Figure 1) dont l'exécution nécessite l'utilisation et la combinaison d'outils et de ressources qui sont présentées en sections 3.3.1., 3.3.2. et 3.3.3. Les résultats de l'évaluation de ces approches sont présentés en section 4.

3.3.1. Étape 1 - L'étiquetage morphosyntaxique

1 https://universaldependencies.org/guidelines.html (consulté le 27/10/2019 de 15h - 16h)

D'après Paroubek et Rajman (2000), l'étiquetage grammatical, également appelé étiquetage morphosyntaxique, consiste à affecter, à chaque mot d'un corpus, un symbole représentant sa catégorie grammaticale (nom, verbe...) et, éventuellement, les informations morphologiques associées (masculin, singulier, etc.). L'utilisation d'un étiqueteur morphosyntaxique permet d'automatiser cette opération (Toutanova et al., 2003). Pour construire Extranéo, nous avons comparé quatre étiqueteurs (Tableau 2), que nous avons retenus au regard de leur disponibilité et du fait qu'ils ont été entraînés pour le français. Il s'agit de SEM (Segmenteur-Étiqueteur Markovien, Constant et al., 2011), de Stanford POS Tagger (Toutanova et al., 2003), de TreeTagger (Schmid, 1994) et de WinBrill (Brill, 1992). Ces étiqueteurs ont été évalués (cf. section 4.1.) sur leur aptitude à étiqueter correctement les mots renvoyant aux catégories grammaticales caractéristiques des néologismes.

Tableau 2 - Description des étiqueteurs morphosyntaxiques évalués

Étiqueteurs	**Méthode d'étiquetage**	**Corpus d'apprentissage Étiquettes**	**Particularité**
SEM	Champs markoviens conditionnels	French Tree Bank ou FTB (Abeillé et al., 2003) 29 étiquettes	Segmentation incorporée
Stanford POS Tagger	CMM à maximisation d'entropie	FTB ; 17 étiquettes	Bidirectionnel ; Utilisation des étiquettes UD
TreeTagger	Arbres de décision	43 834 mots (Achim Stein, 2003) ; 33 étiquettes	Possibilité d'adaptation à d'autres langues
WinBrill	Règles déduites d'un apprentissage	450 000 occurrences tirées de FRANTEXT 50 étiquettes	Entrainement possible sur tout type de corpus étiqueté

3.3.2. Étape 2 - L'identification des néologismes candidats

Pour cette étape, nous utilisons un corpus d'exclusion constitué de lexiques flexionnels comportant les flexions de différents mots, appelées « formes ». Lorsqu'un mot est absent de ces lexiques, il est identifié comme un néologisme candidat. Cinq lexiques flexionnels (Tableau 3) ont été retenus au regard de leur disponibilité et évalués pour réaliser cette deuxième phase du processus. Il s'agit de la liste des noms communs de l'Association des bibliophiles universels ou ABU[2], du DELA (Gross, 1975), du Lefff (Sagot, 2010), du Lexique 3 (New et al., 2004) et de Morphalou (Romary et Salmon-alt, 2004). Une fois identifiés, les néologismes candidats sont vérifiés manuellement puis annotés (étape 3) en « néologismes validés » ou en « non néologismes » selon que la nouveauté lexicale est avérée ou pas. L'étape suivante consiste à mettre en exergue les contextes des néologismes validés.

Tableau 3 - Description des lexiques flexionnels évalués

Lexique	**Nb de Formes**	**Description**	**Version utilisée**	**Données de base**
ABU (Liste des mots communs)	289 612	Informations morphosyntaxiques Mots simples et composés	2002	Textes du domaine public
DELA	792 260	Informations morphosyntaxiques	1990	Tout texte en langue française

2 http://abu.cnam.fr/DICO/mots-communs.html

LEFFF	569 858	Informations morphosyntaxiques	2004	Listes de mots
Lexique 3	137 404	Informations morphosyntaxiques, phonologiques, etc.	2004	Textes littéraires et sous-titres de films
Morphalou	524 725	Informations morphosyntaxiques	2004	Trésor de la langue française

3.3.3. Étape 4 - La mise en relief du contexte des néologismes validés

Dans Extranéo, les marqueurs de glose permettent de mettre en lumière les contextes des néologismes validés car les marqueurs de glose « pointent l'explication de sens dans les textes et [...] explicitent la relation sémantique lexicale mise en jeu » (Mela, 2004). En d'autres termes, il existe un lien étroit entre les marqueurs de glose et les contextes des mots qu'ils introduisent. Pour cette étude, nous nous sommes appuyés sur la description de seize marqueurs de glose (Steuckardt et Niklas-Salminen, 2005) et sur les travaux de (Mela, 2004 ; Mela et Roche, 2006) pour attribuer chaque marqueur de glose à un type de contexte. Sur cette base, neuf marqueurs ont été sélectionnés et repartis en trois types de contextes. Il s'agit des contextes définitoire, de dénomination et d'équivalence. Le contexte définitoire « rassemble les informations les plus essentielles du sens des mots et des catégories d'objets qu'ils désignent » (Cartier, 2011). Quant au contexte de dénomination, il relève de la « néologie dénominative » qui renvoie à « la nécessité de donner un nom à un objet, à un concept nouveau » (Mortureux, 1984). Enfin, le contexte d'équivalence établit « une relation entre le signe et la chose en posant une équivalence sémantique entre deux unités qui désignent la même chose » (Steuckardt et Niklas-Salminen, 2005). Après avoir formulé un patron morphosyntaxique pour les neuf marqueurs de glose, ces différents contextes ont été identifiés au sein du corpus avec le logiciel de textométrie TXM (Heiden et al., 2010), qui intègre le moteur de requêtes CQP (Evert et Hardie, 2011).

Concernant le contexte définitoire, des patrons ont été définis pour les quatre marqueurs suivants : *signifiant, qui signifie, désignant* et *entendez par là*. Ces patrons ont été conçus à partir de la propriété *word* du corpus. Par exemple, celui du marqueur *entendez par là* se présente ainsi : [word="entendez"][word="par"][word="là"]. Concernant le contexte de dénomination, des patrons ont été définis pour les marqueurs *dénommé, appelé, qualifié de*. Ici, l'application du marqueur *appelé*, par exemple, tient compte du fait qu'il est généralement précédé d'un adverbe en *–ment* et qu'il peut s'accorder au féminin pluriel. Le patron se présente ainsi : [frudpos="ADV" & word=".*ment"][word="appelé.*?"]. Concernant le contexte d'équivalence, les patrons ont été conçus pour les marqueurs *ou* et *c'est-à-dire*.

4. Résultats et discussion

Nous présentons trois résultats dans cette section : la performance des étiqueteurs à attribuer les étiquettes selon le type de texte (cf. section 4.1.), la performance des lexiques à favoriser la détection des néologismes candidats (cf. section 4.2.) et la mise en relief des contextes des néologismes à partir des marqueurs de glose (cf. section 4.3.). Trois scores d'évaluation sont calculés pour permettre de comparer les différentes approches, à savoir la précision, le rappel et la F-mesure (Rijsbergen, 1979), qui est une moyenne harmonique entre la précision et le rappel obtenue au moyen de l'équation 1.

$$F-mesure=\frac{(2*pr\acute{e}cision*rappel)}{(pr\acute{e}cision+rappel)} \qquad \textit{(Equation 1)}$$

4.1. Performance des étiqueteurs selon le type de texte

La performance des étiqueteurs est présentée en Figure 2. Le score de précision correspond au nombre total de mots de la catégorie correctement étiquetés divisé par le nombre total de mots de cette catégorie détectés par l'étiqueteur. Le rappel est le rapport entre le nombre de mots de la catégorie correctement étiquetés et le nombre effectif total de mots du corpus correspondant à cette catégorie.

TreeTagger détient la plus forte valeur de F-mesure pour les commentaires (0.92), les articles de journaux (0.88) et les messages Facebook (0.85). Sa performance sur les articles de journaux est liée au fait qu'il obtient plus de 0.80 de F-mesure sur les adverbes, les noms et les verbes. Certaines erreurs d'étiquetage de TreeTagger sont liées à la syntaxe du français du Burkina. Ainsi, la majuscule placée en tête de *Femme* (« Femme est argentivore » : omission du déterminant) et *Guiro* (un « Guiro » désignant un milliard : antonomase) a conduit TreeTagger à les étiqueter comme des noms propres alors qu'il s'agit de noms communs. La performance de l'étiqueteur étant liée à la syntaxe de l'énoncé, les bonnes valeurs de F-mesure soulignent l'aptitude de TreeTagger pour traiter des textes en français du Burkina.

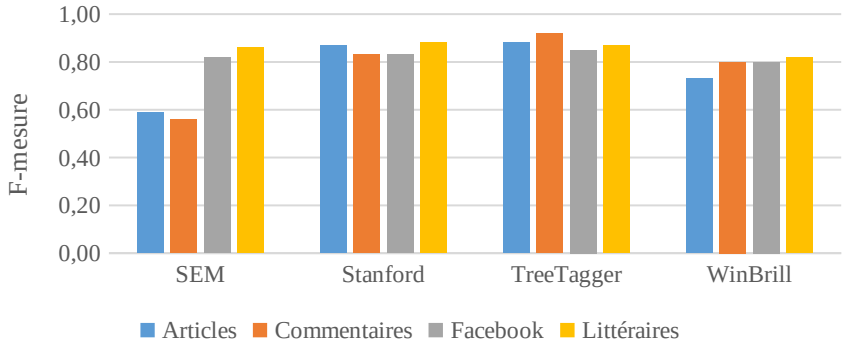

Figure 2 - Performance des étiqueteurs par type de texte

4.2. Identification des néologismes candidats

Pour juger de la capacité des lexiques à détecter les néologismes validés, la précision consiste à diviser le nombre de néologismes correctement détectés par le nombre de néologismes détectés ; le rappel, le nombre de néologismes correctement identifiés par le nombre total de néologismes. La Figure 3 présente la performance des lexiques à identifier les néologismes. Sur les articles de journaux, les lexiques DELA et Morphalou obtiennent la plus forte valeur de rappel et de précision, respectivement 0.84 et 0.52. Ces deux lexiques ont la plus forte F-mesure sur les articles de journaux, à savoir 0.64. Les néologismes, non identifiés par ces lexiques, ont été l'objet d'erreurs d'étiquetage par TreeTagger car ils étaient dotés de catégories différentes de celles caractéristiques des néologismes, à savoir l'étiquette ABR (abréviation) pour *OGM* « dépigmentée » et NAM (nom propre) pour *Cntistes* « membres du CNT », *Chérif* « prince », *Mba* « grand-père, en moré » et *Guiro* « un milliard ».

Par ailleurs, l'utilisation des lexiques, pour identifier les néologismes, conduit à recenser un nombre élevé de néologismes candidats, 230 pour 144 néologismes validés. Ce traitement requiert un temps d'annotation manuelle important pour les dissocier.

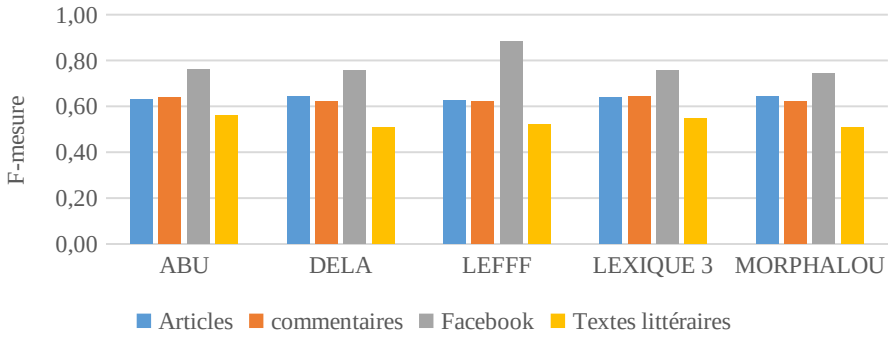

Figure 3 - Performance des lexiques par type de texte

4.3. Mise en relief des contextes

La représentativité des contextes des néologismes est présentée dans le Tableau 4. Le score de précision correspond au rapport entre le nombre de contextes intégrant un néologisme et le nombre de contextes retournés par les patrons. Quant au rappel, il correspond à la proportion entre le nombre de contextes intégrant un néologisme et le nombre de néologismes réels. Les résultats montrent que c'est le contexte définitoire qui détient la plus forte valeur de précision. L'exécution des patrons conçus pour extraire ce contexte (cf. section 3.3.3.) donne une précision de 0.49, alors que celle-ci est de 0.22 pour le contexte de dénomination et de 0.24 pour le contexte d'équivalence. La représentativité élevée du contexte définitoire traduit une volonté de faire comprendre le message véhiculé par les néologismes, car la glose offre l'avantage de proposer une définition initiale du néologisme comme dans l'exemple « la "napaga", entendez par là l'épouse du président du Faso ». L'usage du contexte de dénomination montre que les néologismes sont utilisés pour désigner des réalités nouvelles, comme c'est le cas dans « pain enfoui de brochettes communément appelé "pain-bro" ». Le contexte d'équivalence, mis en exergue dans « chenille de karité ou " chitoumou" », permet d'établir une relation synonymique pouvant favoriser la compréhension du message.

Tableau 4 - Représentativité des contextes des néologismes

	Précision	**Rappel**	**F-Mesure**
Définitoire	0.49	0.19	0.28
Dénomination	0.22	0.59	0.32
Équivalence	0.24	0.22	0.23

Conclusion

L'évaluation que nous avons réalisée montre que l'approche Extranéo peut être mise en œuvre pour détecter les néologismes formels dans des articles journalistiques burkinabè et en extraire les contextes. Sans un entraînement spécifique au français du Burkina, TreeTagger obtient une F-mesure de 0.86, un résultat d'étiquetage satisfaisant qui permet de l'utiliser en l'état pour des articles de journaux. Les cinq lexiques évalués permettent de détecter la quasi-totalité des néologismes du corpus d'évaluation. L'identification des néologismes validés requiert, par contre, une sollicitation importante d'experts au regard du nombre de néologismes candidats. Une solution pour réduire le nombre de candidats consisterait à combiner plusieurs lexiques. Enfin, l'étude des contextes met en évidence l'emploi privilégié du contexte définitoire ; ce qui traduit une volonté d'assurer la compréhension du message véhiculé par les néologismes auprès des destinataires. Grâce à cette approche combinant

étiqueteur, lexique et recours aux gloses, il sera désormais possible d'appréhender la vitalité du français du Burkina via l'analyse des créations lexicales et de leur cycle de vie.

Références

Association des Bibliophiles Universels. (2001). ABU, Dictionnaire des mots communs. In *La Bibliothèque Universelle*, http://abu.cnam.fr/DICO/mots-communs.html,. CNAM. Consulté le 1er janvier 2020.

Brill E. (1992). A simple rule-based part of speech tagger. In *Proceedings of the third conference on Applied natural language processing (ANLC '92)*, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 152-155.

Cartier E. (2011). Utilisation des contextes dans le cadre dictionnairique : état des lieux, typologie des contextes, exemple des contextes définitoires. In *Actes des Huitièmes Journées scientifiques du Réseau de chercheurs Lexicologie, terminologie, traduction*, Lisbonne, 15-17 octobre 2009, pp. 619-632.

Cartier E. (2016). Neoveille, système de repérage et de suivi des néologismes en sept langues. *Neologica*, 10 : 101-131.

Cartier E. (2019). Néoveille, plateforme de repérage et de suivi de néologismes en corpus dynamique. *Néologica*, 13 : 23-54.

Cartier E., Sablayrolles J.-F. (2009). Néologismes, dictionnaires et informatique. In *Cahiers de Lexicologie*, Centre National de la Recherche Scientifique, pp. 175-192.

Cartier E., Sablayrolles J.-F., Boutmgharine N., Humbley J., Bertocci M., Jacquet-pfau C., Kübler N. et Tallarico G. (2018). Détection automatique, description linguistique et suivi des néologismes en corpus : point d'étape sur les tendances du français contemporain. In *Actes du Congrès Mondial de Linguistique Française*, Mons (Belgique), 9-13 juillet 2018, 20 p.

Constant M., Tellier I., Duchier D., Dupont Y., Sigogne A. et Billot S. (2011). Intégrer des connaissances linguistiques dans un CRF : application à l'apprentissage d'un segmenteur-étiqueteur du français. In *Actes de TALN'11*, Montpellier, France, 13 p..

Coseriu E. (1964). Pour une sémantique diachronique structurale. *Centre de philologie et de littérature romanes*, Université de Strasbourg, pp. 139-186.

Evert S., Hardie A. (2011). Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. *In Proceedings of the Corpus Linguistics 2011 conference*, Birmingham, UK. http://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2011/Paper-153.pdf.

Falk I., Bernhard D. et Gérard C. (2014). From Non Word to New Word: Automatically Identifying Neologisms in French Newspapers. In *Proceedings of the International Conference on Language Resources and Evaluation*, Reykjavik, Iceland.

Frey C. (2004). Particularismes lexicaux et variétés de français en Afrique francophone : autour des frontières. In Moreau, *Langues de frontières et frontières de langues, GLOTTOPOL* n°4, Revue de sociolinguistique en ligne : 136-149.

Gandon F.-M. (1994). Appropriation et syntaxe du français écrit dans la presse de Ouagadougou (Burkina Faso) : préposition, rection, pronoms. *Langue française*, 104 : 70-88

Gross M. (1975). *Méthodes en syntaxe*. Hermann, Paris, France.

Harris Z. (1968). *Mathematical Structures of Language*. New-York, John Wiley & Sons.

Heiden S., Magué J.-P., Pincemin B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie-conception et développement. In *Processings of 10th International Conference on the Statistical Analysis of Textual Data*, 2 : 1021-1032.

IFA. (2004). *Inventaire des particularités lexicales du français en Afrique noire*. 3e édition, Paris, EDICEF/AUF.

Kedrebéogo G., Yago Z. et Hien T. (1988). Burkina Faso. Carte linguistique, IRSSH, CNRST, Ouagadougou. *Barreteau*, 1998 : 7.

Kéita A. (2013). Hybridation et productivité lexicale en français parle au Burkina. *Revue électronique des Sciences du langage*, 19 : 88-101.

Koama C. (2015). L'amalgamation lexicale comme procédé satirique dans le Journal du jeudi. *Neologica*, 9 : 153-168

Lafage S. (1977). *Facteurs de différenciation entre le français central et le français d'Afrique*. CILF, Paris.

Mela A. (2004). Linguistes et « talistes » peuvent coopérer : repérage et analyse des gloses. *Revue française de linguistique appliquée*, IX : 63-82.

Mela A. et Roche M. (2006). Des gloses de mots aux types de textes : un bilan différencié. In *Actes du colloque « Corpus en Lettres et Sciences sociales : des documents numériques à l'interprétation »*, Albi, juillet 2006. Texto ! [en ligne], XI, n°2.

Mela A., Roche M. et Bekhtaoui M. E.-A. (2011). Mixer les moyens pour extraire les gloses. *EGC : Extraction et Gestion des Connaissances*, Jan 2011, Brest, France, pp. 95-106

Morlane-Hondère F. (2013). Une approche linguistique de l'évaluation des ressources extraites par analyse distributionnelle automatique. Linguistique. Université Toulouse le Mirail – Toulouse II, 290 p.

Mortureux M.-F. (1984). La dénomination, approche socio-linguistique. *Langages*, 19e année, 76 : 95-112.

New B., Pallier C., Brysbaert M., Ferrand L. (2004). A New French Lexcical Database. *Behavior Research Methods, Instruments & Computers*, 36 (3) : 516-524.

Ouédraogo A. (2019). Du défigement des expressions figées du français dans la presse burkinabè : une autre forme de burkinabismes. In Ouédraogo, Y., *L'écart linguistique*, Sankofa et Guirli Editions Ouagadougou 2019, pp. 59-84.

Ouédraogo Y. (2000). Le français basilectal dans la littérature burkinabè. In *Actes de la XVIIIe de la Biennale de la langue française*, Paris.

Ouédraogo Y. (2008). Les particularismes du français d'Afrique noire : entre écart et enrichissement. *Language, culture and littérature*, pp. 82-95, Ghana, D. D. Kuupole.

Paroubek P. et Rajman M. (2000). Étiquetage morphosyntaxique. In Pierel, J.-M. (ed.), *Ingénierie des Langues*, Hermes Science, Paris, pp. 131-150,

Poirier C. (1995). Les variantes topolectales du lexique français : propositions de classement à partir d'exemples québécois. In Francard, M. et Latin, D. (eds), *Le régionalisme lexical*, Louvain-la-Neuve, Duculot-de-Boeck et AUPELF-UREF, pp. 13-56.

Prignitz G. (1993). Place de l'argot dans la variation linguistique en Afrique : le cas du français à Ouagadougou. In *Le français au Burkina Faso*, sous la direction de Caïtucoli, C. (éd.), *Cahiers de linguistique sociale*, Université de Rouen, Collection Bilans et perspectives, pp. 117-128.

Prignitz G. (1996). Aspects lexicaux, morphosyntaxiques et stylistiques du français parlé au Burkina Faso. Paris III, Sciences du langages, thèse de doctorat, Université de la Sorbonne nouvelle, 505 p.

Raschi N. (2009). La langue française dans la presse du Burkina Faso. *Alternative Francophone* 1 (2) : 136-154.

Rijsbergen V. (1979). *Information Retrieval* (second ed.). London: Butterworths.

Romary L. et Salmon-alt S. (2004). Standards going concrete : from LMF to Morphalou. *COLING 2004 Enhancing and using electronic dictionaries*, GENEVA, pp. 22-28.

Sablayrolles J.-F. (2002). Fondements théoriques des difficultés pratiques du traitement des néologismes. *Revue française de linguistique appliquée*, VII (1) : 97-111.

Sagot B. (2010). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In *7th international conference on Language Resources and Evaluation* (LREC 2010), May 2010, Valletta, Malta.

Schmid H. (1994). Probabilistic Part-Of-Speech Tagging Using Decision Trees. In *Proceedings of the Conference on New Methods in Language Processing*, Manchester, UK.

Stein A. (2003). Étiquetage morphologique et lemmatisation de textes d'ancien français. Kunstmann, Pierre et al. (éd.): *Ancien et moyen français sur le Web: Enjeux méthodologiques et analyse du discours*, Ottawa: Les Éditions David, 273-284.

Steuckardt A. et Niklas-Salminen A. (2005). Les marqueurs de glose. Langues et langage, Aix-en-Provence, Publications de l'Université de Provence.

Toutanova K., Klein D., Manning C., Singer Y. (2003). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of HLT-NAACL 2003*, pp. 252-259.

Wall L., Christiansen T., Orwant J. (2001). *Programmation en Perl*, 3e édition Paris : O'Reilly.