*Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles*
*Avignon, France, 27 juin au 1er juillet 2022*

# Impact du français inclusif sur les outils du TAL

Cyril Grouin
Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numériques (LISN)
507 rue du Belvédère, 91400 Orsay, France

RÉSUMÉ

Le français inclusif est une variété du français standard mise en avant pour témoigner d'une conscience de genre et d'identité. Plusieurs procédés existent pour lutter contre l'utilisation générique du masculin (coordination de formes féminines et masculines, féminisation des fonctions, écriture inclusive, et neutralisation). Dans cette étude, nous nous intéressons aux performances des outils sur quelques tâches du TAL (étiquetage, lemmatisation, repérage d'entités nommées) appliqués sur des productions langagières de ce type. Les taux d'erreur sur l'étiquetage en parties du discours (TreeTagger et spaCy) augmentent de 3 à 7 points sur les portions rédigées en français inclusif par rapport au français standard, sans lemmatisation possible pour le TreeTagger. Sur le repérage d'entités nommées, les modèles sont sensibles aux contextes en français inclusif et font des prédictions erronées, avec une précision en baisse.

ABSTRACT

**Impact of French Inclusive Language on NLP Tools**

French Inclusive language (Gender-Neutral language) is a variety of standard French that is used to highlight an awareness of gender and identity. Several processes exist to substitute the generic use of the masculine form (coordination of feminine and masculine forms, feminization of functions, inclusive writing, and neutralization). In this study, we focus on the performance of a few NLP tools (labeling, lemmatization, name entity recognition) applied to language productions in French Inclusive language. The error rate of TreeTagger and spaCy on Part-of-Speech tagging increases from 3 to 7 points on spans written in French inclusive language with respect to standard French language, and no lemmatization was possible using the TreeTagger. In named entity recognition, models are sensitive to contexts written in French Inclusive and produce erroneous predictions, implying a lower precision.

MOTS-CLÉS : Français inclusif, Traitement Automatique des Langues, Taux d'erreur.

KEYWORDS: French Inclusive Language, Natural Language Processing, Error Rate.

## 1 Introduction

Le français inclusif est une variété du français standard, mise en avant pour témoigner d'une conscience de genre et d'identité (Alpheratz, 2018, 2019) au regard de l'utilisation générique du masculin, actuellement employé pour rassembler un collectif composé de femmes et d'hommes (« *les lecteurs* »), ou dans les tournures impersonnelles (« *il fait beau* »). En raison du caractère performatif du langage (Austin, 1962), l'utilisation du masculin générique développe un habitus de pensée que les personnes promouvant le français inclusif cherchent à combattre.

Volume 1 : conférence principale, pages 126–135.

Bien qu'attesté depuis plus d'un siècle dans les communications politiques sous la forme de coordination de gentilés au féminin et au masculin [1], le français inclusif s'est récemment illustré avec des polémiques et des prises de position autour de l'« écriture inclusive », mise en avant par le Haut Conseil à l'Égalité entre les femmes et les hommes qui a produit un guide pratique (HCE, 2015). Ainsi, l'Académie française a adopté à l'unanimité une déclaration lors de la séance du 26 octobre 2017, considérant l'écriture inclusive comme une « *langue désunie* » qui crée « *une confusion qui confine à l'illisibilité* » et concluant que « *la langue française se trouve désormais en péril mortel* » [2]. Dans une lettre ouverte cosignée par Hélène Carrère d'Encausse, Secrétaire perpétuel de l'Académie française, et Marc Lambron, Directeur en exercice de l'Académie française, datée du 7 mai 2021 [3], les cosignataires dénoncent à la fois « *une injonction brutale, arbitraire et non concertée* » ainsi que le « *principe [d']une corrélation entre le genre des vocables et le sexe de leur référent* ». Suite à ces déclarations, une proposition de Loi « visant à sauvegarder la langue française et à réaffirmer la place fondamentale de l'Académie française » a été enregistrée le 1er juin 2021 auprès de l'Assemblée nationale [4], dans la mesure où les circulaires gouvernementales (datées du 21 novembre 2017 par Édouard Philippe, Premier Ministre, et du 6 mai 2021 par Jean-Michel Blanquer, Ministre de l'Éducation nationale, de la Jeunesse et des Sports), « *en l'absence d'une vraie volonté, ne sont pas appliquées* ». Cette Loi interdit formellement tout usage de l'écriture inclusive par les administrations. Toute personne morale contrevenant à cette interdiction est passible d'une amende de 7500€.

La plupart des procédés linguistiques disponibles pour réaliser des productions en français inclusif conduisent à la création de formes nouvelles, telles que « *lecteur·ice* » par la combinaison des formes masculine et féminine des désinences ou « *locutaire* » par le procédé de neutralisation (voir section 3.1). Cette créativité morphologique génère des mots nouveaux qui peuvent constituer des obstacles ou engendrer des erreurs pour les outils actuels du traitement automatique des langues. À notre connaissance, il n'existe pas d'étude spécifique au TAL appliqué aux productions en français inclusif. L'objectif de cette étude exploratoire vise à vérifier quelles sont les performances de quelques outils actuellement disponibles en traitement automatique des langues sur deux tâches classiques (étiquetage en parties du discours et lemmatisation, repérage d'entités nommées) réalisées sur des productions langagières en français inclusif, et quel est l'impact du français inclusif sur ces outils.

## 2 État de l'art

L'usage du français inclusif a été jusqu'à présent uniquement étudié sous les aspects sociologiques et linguistiques (Abbou, 2017; Díaz & Heap, 2020; Manesse, 2021). En traitement automatique des langues (TAL), le traitement des mots hors vocabulaire constitue un domaine de recherche actif (Spriet *et al.*, 1996; Maurel, 2004), en particulier pour la compréhension (Bensoussan & Laufer, 1984), le traitement de la parole (Stouten *et al.*, 2010), la traduction automatique (Cartoni, 2008), le traitement des langues de spécialité (Rabary *et al.*, 2015) ou des sorties bruitées par des fautes d'orthographe (Baranes & Sagot, 2014) ou des erreurs de reconnaissance optique de caractères (Oprean *et al.*, 2014).

1. Le quotidien *Le Drapeau* du 5 octobre 1910 reproduit la prise de parole de M. Dontenville qui commence par l'adresse « *Françaises, Français* » lors de la commémoration annuelle de la défense du Colonel Denfert-Rochereau lors du siège de Belfort (https://gallica.bnf.fr/ark:/12148/bpt6k40795430).

2. https://www.academie-francaise.fr/actualites/declaration-de-lacademie-francaise-sur-lecriture-dite-inclusive

3. https://www.academie-francaise.fr/actualites/lettre-ouverte-sur-lecriture-inclusive

4. https://www.assemblee-nationale.fr/dyn/15/textes/l15b4206_proposition-loi

De manière similaire, les langues de spécialité (juridique, médecine, etc.), les registres de langue (académique, familier, soutenu, etc.), et les usages particuliers des locuteurs d'une langue (courriers électroniques, réseaux sociaux, messages textuels par téléphone, etc.) engendrent des pratiques linguistiques particulières aux niveaux orthographique, sémantique et syntaxique, qui peuvent s'apparenter à des mots inconnus de la langue générale ou des mots hors vocabulaire, et qui nécessitent des ressources et outils dédiés pour leur analyse par des outils informatiques. Il a été démontré que l'analyse de ces productions langagières au moyen d'outils de traitement automatique des langues est de meilleure qualité si les ressources développées (lexiques, modèles statistiques) sont représentatives des usages qu'elles doivent traiter (L'Homme, 2008). Par ailleurs, ces mêmes productions langagières peuvent servir à la production de ressources dédiées (Bourigault & Aussenac-Gilles, 2003). Ainsi, face à des ressources à large couverture telles que WordNet (Miller *et al.*, 1990) et EuroWordNet (Vossen, 1997), des ressources adaptées ont été développées pour permettre l'analyse automatique du contenu des langues de spécialité, telles que l'UMLS (Unified Medical Language System) (Lindberg *et al.*, 1993) et l'extension Medical WordNet (Smith & Fellbaum, 2004) pour le domaine biomédical en anglais, en parallèle de ressources spécifiques pour une tâche donnée telles que WordNet Affect (Strapparava & Valitutti, 2004) pour l'analyse des émotions. Par ailleurs, nombreux sont les travaux apportant des méthodes et ressources linguistiques autour des langues de spécialité comme le biomédical (Laparra *et al.*, 2021), les usages spécifiques aux réseaux sociaux (Farzindar & Roche, 2013; Benamara *et al.*, 2018), ou encore les registres de langue (Mekki *et al.*, 2021).

# 3 Matériel et méthodes

## 3.1 Procédés du français inclusif

Alpheratz (2019) rappelle les différents procédés linguistiques d'atténuation ou de remplacement du masculin générique existants :

- la coordination des formes féminines et masculines dans un même énoncé : *bonjour à toutes et à tous ; les françaises et les français*. Ce procédé est particulièrement prisé des politiques et attesté depuis plusieurs années, tel que le discours « *Françaises, Français, aidez-moi !* » prononcé par le Général de Gaulle le 27 juin 1958, ou l'adresse « *travailleuses, travailleurs* » utilisée par Arlette Laguiller depuis ses premiers discours de 1974 ;
- la féminisation des noms de fonction, comme recommandé par le Haut Conseil à l'Égalité femmes-hommes [5] : *autrice, maîtresse de conférence* ;
- l'utilisation de termes épicènes : *personne, élève, individu, journaliste, membre, politique* ;
- la combinaison de désinences morphologiques masculine et féminine dans une même unité lexicale au moyen d'un point médian entre chaque flexion (cette forme particulière d'écriture est qualifiée d'« écriture inclusive ») : *les chercheur·euses, les doctorant·e·s, les locuteur·rices* ;
- la neutralisation par la créativité morphologique [6] :
  - de nouvelles unités lexicales $(i)$ correspondants à une construction de type « mot-valise » (parmi les formes les plus utilisées, citons les pronoms personnels *iel* et *iels* (par combinaison des formes *il, elle* et *ils, elles*) [7], les pronoms démonstratifs *cellui* et *celleux* (par combinaison des formes *celle, celui* et *celles, ceux* ; la forme *ceuxelles* a également été

5. https://www.haut-conseil-egalite.gouv.fr/
6. Les exemples suivis du symbole alpha en exposant ($^{\alpha}$) sont issus des travaux d'Alpheratz (2019).
7. Les formes alternatives suivantes ont également été proposées avec un moindre succès : *yel, ael, el, ille, ol, ul*.

proposée), le nom *frœur*$^{\alpha}$ pour englober *frères* et *sœurs*), $(ii)$ en puisant de nouvelles racines dans le substrat gréco-latin (*adelphité*$^{\alpha}$ en remplacement de *fraternité* qui réfère à la fratrie, forcément masculine), ou $(iii)$ en conférant à la graphie « æ » une valeur neutre (*mæ*$^{\alpha}$ vs. *mon, ma* ; *lesquælsα* vs. *lesquels, lesquelles*), etc. ;

— de nouvelles formes fléchies fondées sur des désinences existantes du français standard (*les écrivans*$^{\alpha}$*, les lectaires*$^{\alpha}$), ou par de nouvelles désinences, telle que la ligature « æ » pour noter les participes passés sans modifier la prononciation (*je suis blessæ*$^{\alpha}$) ;

— en réactivant des formes du français médiéval (le pronom *el* ou *al* (Marchello-Nizia, 1989) dans les tournures impersonnelles : *al fait beau*) et d'anciennes règles (accord de proximité : *les garçons et les filles sont belles*, titre de la vidéo de Riban & Gerin (2017)).

Au-delà des difficultés d'oralisation que pose l'écriture inclusive [8], l'absence de touche sur le clavier pour générer facilement le point médian conduit les utilisateurs à préférer des variantes proches telles que *chercheur.euse, lecteur.ice* (utilisation du point), *chercheur-euse, lecteur-ice* (utilisation du trait d'union), *chercheur'euse, lecteur'ice* (apostrophe, plus rare), et *chercheureuse, lecteurice* (absence de séparateur) ; ce dernier cas se rencontre plus fréquemment sur les mots courts tels que *toustes*.

## 3.2 Corpus

Le corpus se compose de courts extraits de discours politiques publiés sur le site *Vie publique* [9] et de contenus publiés sur des sites gouvernementaux français tels que celui des données ouvertes publiques ou le *Journal officiel* [10], dans l'objectif d'une redistribution de ces ressources. Le corpus final se compose de 21 fichiers. La taille réduite du corpus s'explique par la difficulté à identifier des productions en français inclusif qu'il est possible de redistribuer, et du coût humain pour constituer la référence sur les annotations en parties du discours.

**Version en français inclusif (VFI).** Nous identifions dans les discours et données sources les extraits en français inclusif au moyen de quelques exemples représentatifs des procédés linguistiques. Nous avons ainsi listé empiriquement les locutions les plus fréquentes et cherché ces locutions pour constituer notre corpus (voir les locutions en gras dans le tableau 1). Pour le procédé de neutralisation, en raison de l'impossibilité d'identifier de tels exemples dans les discours ou dans les données ouvertes disponibles, nous avons inventé des exemples en suivant les règles de neutralisation décrites par Alpheratz (2019). Le corpus final comprend un total de 1313 mots et de 56 portions relevant du français inclusif.

**Version en français standard (VFS).** Nous produisons une version du corpus en remplaçant chaque formulation du français inclusif [11] par sa correspondance en français standard (ainsi, « *Merci à toutes et tous d'être nombreux aujourd'hui* » devient « *Merci à tous d'être nombreux aujourd'hui* »). Cette version conduit à une diminution de 2,26% du nombre total de tokens dans le corpus.

8. La phrase « *Les doctorant·e·s sont venu·e·s nombreux·ses.* » serait prononcée par décomposition des éléments de flexion (« doctorant point e point s ») à l'image du domaine *.fr* (« point f r ») dans les adresses électroniques, conduisant à une perte de fluidité du discours. La prononciation de « *nombreux·euses* » avec une consonne sourde sifflante suivie d'une consonne sonore sifflante est impossible en français et engendre généralement une assimilation [nɔ̃bʀøgzøz] mais reste complexe.
9. https://www.vie-publique.fr/
10. https://www.data.gouv.fr/fr/ — https://www.legifrance.gouv.fr/
11. Nous conservons les coordinations de gentilés dans la mesure où ils peuvent constituer un type d'entité nommée.

| Procédé | Extraits |
|---|---|
| Coordination de termes (10 fichiers) (34 exemples) | **Martiniquaises, Martiniquais,** me voici donc à la Martinique où personne n'est jamais venu sans aimer à la fois cette terre et ce peuple. Voici donc **les Martiniquaises et les Martiniquais** rassemblés pour accueillir le President de la République qu'ils ont élu eux-mêmes démocratiquement. (Valéry Giscard-d'Estaing, 13 décembre 1974) |
| | Je remercie les enseignants, les professeurs, **celles et ceux** qui consacrent le meilleur de leur vie à la formation de la jeunesse de France. (François Mitterrand, 23 mai 1987) |
| | Merci à **toutes et tous** d'être nombreux aujourd'hui à Paris pour réaffirmer que le travail doit venir avant la Bourse. (Marc Blondel, 21 novembre 1998) |
| | Vous incarnez, en réalité, **toutes et tous**, l'Europe que nous souhaitons, c'est-à-dire l'Europe de l'union, l'Europe du progrès, l'Europe de l'ambition. (Jacques Chirac, 28 avril 2005) |
| Combinaison flexionnelle (5 fichiers) (10 exemples) | Nous voulons élargir à toute la société les possibilités d'accès aux formes les plus élaborées du savoir scientifique et permettre à **tout(e) étudiant(e)** d'aller au bout de ses possibilités, avec le souci permanent de la validation des parcours et des acquis. (Jean-Luc Mélenchon, 24 février 2012) |
| | Les **transexuel/es** doivent pouvoir changer d'état-civil et/ou de numéro de Sécu, sans passer par un parcours psychiatrique, par une opération chirurgicale ou une stérilisation et les **intersexué/es** ne doivent pas être **mutilé/es** à la naissance pour les faire correspondre à un sexe ou un autre. [...] La prostitution touche des dizaines de milliers de personnes, dont plus de 20 000 **étudiant/es**. (Philippe Poutou, 11 avril 2012) |
| | Une trace est la succession des inscriptions d'**un·e étudiant·e** à l'université en partant de son Bac jusqu'à sa dernière inscription. Une étape (de diplôme) est ce à quoi l'**étudiant·e** s'inscrit. Dans la version publique, présentée ici on masque l'effectif des cohortes de moins de 10 **étudiant·e·s**. (Visualisation des traces des étudiant·e·s de UP13, 6 mai 2017) |
| Féminisation de fonctions (3 fichiers) (6 exemples) | Par arrêté de la ministre de la défense en date du 26 septembre 2005, Mme Charlier (Dominique, Thérèse, Marthe), épouse Dagras, est nommée au grade de **lieutenante-colonelle**, en qualité d'**officière** recrutée au titre de l'article 29 du statut général des militaires, pour occuper un emploi de spécialiste des affaires politiques pour une durée de deux ans à compter du 1er octobre 2005. (Arrêté) |
| | Giorgia Marras, jeune **autrice** italienne [...] Giorgia Marras, illustratrice et **auteure** de bande dessinée, est née à Gênes en Italie, en 1988. (BD2020) |
| Neutralisation (3 fichiers) (6 exemples) | **Touz** les **députæs** et **sénataires** ont abouti à un accord de principe sur le pass vaccinal |
| | Les premiers romans des **écrivans** sont rarement **appréciæs** des critiques littéraires |
| | Demain, **al** fera beau sur toute la France avec des températures de saison |

TABLE 1 – Extraits du corpus en français inclusif. Les éléments du français inclusif sont en gras

## 3.3 Expériences

**Etiquetage en parties du discours et lemmatisation.** Nous appliquons deux outils d'étiquetage en parties du discours et de lemmatisation parmi les plus utilisés en TAL : TreeTagger (Schmid, 1994) [12], et spaCy (Honnibal *et al.*, 2020) avec le modèle fr_core_news_sm. Bien qu'il existe un nombre important d'outils, nous ne retenons que ces deux outils en raison de leur popularité actuelle ou passée, et de l'intervalle de temps qui les sépare relativement conséquent à l'échelle de l'informatique. Bien que récent, l'outil spaCy peut se révéler moins performant qu'un outil plus ancien.

12. Nous imposons notre tokénisation au TreeTagger afin de rassembler les différentes flexions de l'écriture inclusive dans un même token : *[permettre] [à] [tout(e)] [étudiant(e)] [d'] [aller] — [les] [intersexué/es] [ne] [doivent] [pas] [être] [mutilé/es] [à] [la] [naissance] — [mais] [un·e] [lecteur·ice] [avisé-e] [notera] [que] — [l'] [étudiant·e] [s'] [inscrit].*

**Repérage d'entités nommées.** Nous nous intéressons aux principales classes d'entités nommées (Grishman & Sundheim, 1996). De taille limitée, le corpus comprend 39 entités (17 lieux, 7 dates, 6 personnes, 4 villes, 3 organisations, 1 adresse et 1 code postal). Une seule entité [13] se trouve dans une portion en français inclusif. Nous entraînons et appliquons quatre modèles statistiques délexicalisés (aucune forme de surface n'est mémorisée lors de l'apprentissage ; les modèles varient selon le nombre de documents et de classes) au moyen de l'outil Wapiti (Lavergne *et al.*, 2010) qui implémente les champs aléatoires conditionnels (CRF) de chaîne linéaire.

# 4 Résultats et discussion

**Etiquetage et lemmatisation.** Le tableau 2 présente les taux d'erreur d'étiquetage en parties du discours et de lemmatisation, dans les 56 portions du corpus rédigées en français inclusif (177 tokens) et leur équivalent en français standard (141 tokens), une portion correspondant à un syntagme nominal. Nous renseignons également des performances des deux outils testés selon le type de procédé linguistique utilisé. Notre objectif d'évaluation portant uniquement sur les passages en français inclusif, les autres portions du corpus ne sont pas évaluées.

| Tâche | Etiquetage en POS | | Lemmatisation | |
|---|---|---|---|---|
| Corpus | VFI | VFS | VFI | VFS |
| TreeTagger, *tous procédés confondus* | 15,3% (27/177) | 12,8% (18/141) | 16,4% (29/177) | 1,4% (2/141) |
| ———, *procédé de coordination* | 10,9% (16/147) | 13,5% (15/111) | 0,7% (1/147) | 0,9% (1/111) |
| ———, *procédé de flexions* | 47,1% (8/17) | 11,8% (2/17) | 100,0% (17/17) | 5,9% (1/17) |
| ———, *procédé de féminisation* | 0,0% (0/7) | 14,3% (1/7) | 71,4% (5/7) | 0,0% (0/7) |
| ———, *procédé de neutralisation* | 50,0% (3/6) | 0,0% (0/6) | 100,0% (6/6) | 0,0% (0/6) |
| spaCy, *tous procédés confondus* | 15,8% (28/177) | 11,4% (16/141) | 11,3% (20/177) | 5,7% (8/141) |
| ———, *procédé de coordination* | 6,1% (9/147) | 7,2% (8/111) | 2,7% (4/147) | 4,5% (5/111) |
| ———, *procédé de flexions* | 70,6% (12/17) | 23,5% (4/17) | 64,7% (11/17) | 17,7% (3/17) |
| ———, *procédé de féminisation* | 42,9% (3/7) | 14,3 (1/7) | 85,7% (6/7) | 0,0% (0/7) |
| ———, *procédé de neutralisation* | 50,0% (3/6) | 50,0% (3/6) | 0,0% (0/6) | 0,0% (0/6) |

TABLE 2 – Taux d'erreur d'étiquetage en parties du discours (POS) et de lemmatisation sur les versions en français inclusif (VFI) et en français standard (VFS) sur les portions en français inclusif, par outil et pour chaque procédé étudié

Parce que les outils n'ont pas été entraînés sur ce type de production langagière, le taux d'erreur est plus élevé sur le français inclusif qu'en français standard, aussi bien pour l'étiquetage en parties du discours que pour la lemmatisation. Nous observons que le TreeTagger produit à peine moins d'erreurs de parties du discours que spaCy (15,3% vs. 15,8%), mais ce dernier se révèle plus performant sur la lemmatisation des formulations en français inclusif (11,3%) que le TreeTagger (16,4%). Aucun lemme n'a été produit par le TreeTagger pour les procédés de féminisation, neutralisation, et combinaison de flexion. À l'inverse, spaCy a reproduit à l'identique les formes de surface pour les procédés de féminisation et de combinaison de flexion, et a pu inférer des lemmes pour la neutralisation (*touz, députæ, sénataire, écrivan, appréciæ*). Le procédé de coordination ne pose pas de problème aux outils testés. Nous observons que les deux outils obtiennent un taux d'erreur proche et relativement élevé sur le français standard, ce qui doit nous conduire à relativiser les taux d'erreur du français inclusif.

13. Une organisation nommée en écriture inclusive : *Association des ancien-ne-s etudiant-e-s vice-president-e-s d'universite*

Bien que les différents procédés linguistiques ne soient pas équitablement répartis dans notre corpus, nous observons des erreurs plus nombreuses sur les combinaisons flexionnelles et sur la neutralisation. La combinaison flexionnelle a eu un impact plus important sur l'étiquetage en parties-du-discours pour le TreeTagger (la séquence « *tout(e) étudiant(e)* » a été étiquetée NOM ADJ au lieu de PRO NOM), alors que spaCy a rencontré plus de difficultés à lemmatiser les expressions relevant de ce procédé (« *lecteur·ice* » a été lemmatisé en « *lecteur·ic* »). Le procédé de neutralisation a également généré plusieurs erreurs d'étiquetage en parties du discours pour les deux outils (les étiquettes sont correctes dans un cas sur deux grâce au contexte), et de lemmatisation pour le TreeTagger uniquement (toutes les formes sont lemmatisées par le label « unknown »).

**Repérage d'entités nommées.** L'évaluation repose sur les prédictions d'entités nommées réalisées sur les deux versions du corpus, utilisées comme corpus de test. Les exemples en français inclusif, absents de l'apprentissage, perturbent les modèles qui réalisent des prédictions de classes erronées sur la VFI. Pour chaque modèle, la différence VFI / VFS est uniquement significative sur la précision (sur le meilleur modèle, la précision globale passe de 0,85 en VFI à 0,92 en VFS, en raison de la baisse du nombre de prédictions erronées). La seule entité présente dans une portion en français inclusif n'a été annotée que par un modèle, et de manière partielle, mais avec un choix de classe pertinent sur la portion annotée [14]. Nous précisons que notre évaluation porte cependant sur l'ensemble des entités nommées du corpus, qu'elles figurent dans une portion rédigée en français inclusif ou non.

**Enrichissement du corpus.** À l'occasion des prises de parole lors des élections présidentielles de 2022, nous remarquons que le français inclusif est uniquement utilisé par des personnes politiques de gauche et du centre, et que les candidats sont plus nombreux à l'utiliser à l'oral qu'à l'écrit. Alors que les coordinations de formes féminines et masculines sont nombreuses dans les phrases prononcées par Emmanuel Macron, y compris sur de la parole spontanée (« *les Françaises et les Français* », « *les électrices et les électeurs* », « *celles et ceux* », etc.), la propagande électorale de ce candidat reste fidèle au français standard (« *En me faisant confiance, vous voterez pour les travailleurs de ce pays* », « *Les salariés du privé gagneront plus* », « *vous voterez pour les retraités* », « *je présiderai pour nous tous* », etc.). Seule la profession de foi de Philippe Poutou intègre différents précédés du français inclusif. Cette observation suggère une possibilité d'enrichissement du corpus par des transcriptions automatiques de parole lors d'échanges entre candidats ou avec des journalistes.

# 5 Conclusion

Dans cet article, nous comparons les performances d'outils d'étiquetage-lemmatisation et de repérage d'entités nommées sur un corpus annoté de courts extraits de discours politique en français inclusif (parmi quatre procédés linguistiques retenus) et leur correspondance en français standard [15]. Parce qu'ils n'ont pas été entraînés sur ce type de production langagière, les deux outils (TreeTagger et spaCy) font davantage d'erreurs d'étiquetage sur la version en français inclusif qu'en français standard.

14. Dans la référence, l'ensemble de la portion « *Association des ancien-ne-s etudiant-e-s vice-president-e-s d'universite* » est annoté avec une étiquette Organisation. Notre modèle a seulement considéré la portion « *etudiant-e-s vice-president-e-s d'universite* » et a identifié une entité de type Fonction.

15. Le corpus est disponible dans ses deux versions et porteur d'annotations (portions en français inclusif et entités nommées) à l'adresse : `https://github.com/grouin/corpus-francais-inclusif.git` (ce corpus sera régulièrement complété par de nouveaux extraits par rapport à la version utilisée dans les expériences détaillées dans cet article).

Seul spaCy parvient à inférer des lemmes au moyen de règles sur les exemples de neutralisation, par suppression de la marque du pluriel. Il apparaît donc utile de réentraîner ces systèmes sur des productions langagières en français inclusif. Sur le repérage d'entités nommées, les modèles sont sensibles au contexte rédigé en français inclusif et prédisent de mauvaises classes d'entités dans ces contexte, conduisant à une baisse de précision, sans impact sur le rappel. Les entités en français inclusif restent complexes pour les modèles testés.

Même si les conclusions auxquelles nous aboutissons semblaient prévisibles, nous estimions pertinent de les vérifier de manière expérimentale.

Dans de futurs travaux, nous prévoyons d'enrichir le corpus par de nouveaux exemples de manière à prendre en compte l'évolution diachronique des usages actuellement non normés du français inclusif. Dans cette perspective, un outil de détection des productions en français inclusif constituerait une aide appréciable pour enrichir le corpus actuel, dont nous reconnaissons sa taille limitée et la sur-représentation du procédé de coordination, cette sur-représentation en corpus étant néanmoins représentative de son utilisation abondante par rapport aux autres procédés. Nous envisageons également d'appliquer des plongements de mots pour étudier les voisins des portions rédigées en français inclusif. D'autres tâches, telles que la traduction automatique, restent à être explorées. L'utilisation d'approches neuronales, plus récentes et apprises sur des données contenant potentiellement du français inclusif, devrait améliorer les résultats, mais les productions langagières en français inclusifs semblent pour l'instant principalement présentes à l'oral. En marge de ce travail, notons également que la définition du lemme (forme au masculin singulier) pose question du point de vue de l'idéologie véhiculée par les personnes promouvant l'usage du français inclusif.

## Remerciements

Ce travail a été réalisé dans le cadre du projet GEM (Gender Equality Monitor) sous le financement ANR-19-CE38-0012.

## Références

ABBOU J. (2017). (Typo)graphies anarchistes. Où le genre révèle l'espace politique de la langue. *Mots. Les langages du politique*. DOI : 10.4000/mots.22637.

ALPHERATZ (2018). Français inclusif : conceptualisation et analyse linguistique. In *SHS Web Conf, Congrès Mondial de Linguistique Française*, volume 46. DOI : 10.1051/shsconf/20184613003.

ALPHERATZ (2019). Français inclusif : du discours à la langue ? *Le discours et la langue*, **1**(111). HAL : hal-02323626v2.

AUSTIN J. L. (1962). *How to Do Things with Words*. Oxford : Clarendon Press. The William James Lectures delivered in Harvard University in 1955.

BARANES M. & SAGOT B. (2014). Normalisation de textes par analogie : le cas des mots inconnus. In *Actes de TALN*, Marseille, France. HAL : hal-01019998.

BENAMARA F., INKPEN D. & TABOADA M. (2018). Introduction to the special issue on language in social media : Exploiting discourse and other contextual information. *Computational Linguistics*, **44**(4), 663–681.

BENSOUSSAN M. & LAUFER B. (1984). Lexical guessing in context in efl reading comprehension. *Document numérique*, **7**(1), 15–32. DOI : 10.1111/j.1467-9817.1984.tb00252.x.

BOURIGAULT D. & AUSSENAC-GILLES N. (2003). Construction d'ontologies à partir de textes. In *Actes de TALN*, Batz-sur-Mer.

CARTONI B. (2008). *De l'incomplétude lexicale en traduction automatique : vers une approche morphosémantique multilingue*. Thèse de doctorat, Université de Genève.

DÍAZ Y. & HEAP D. (2020). Variation dans les accords du français inclusif. In *Actes de l'Association canadienne de linguistique*.

FARZINDAR A. & ROCHE M. (2013). *TAL et réseaux sociaux*, volume 54(3). ATALA. HAL : lirmm-01184554.

GRISHMAN R. & SUNDHEIM B. (1996). Message Understanding Conference- 6 : A brief history. In *Proc of COLING*, Copenhague, Danemark.

HCE (2015). *Pour une communication publique sans stéréotype de sexe*. Haut Conseil à l'Égalité entre les femmes et les hommes. http://bit.ly/2fejwZ7.

HONNIBAL M., MONTANI I., VAN LANDEGHEM S. & BOYD A. (2020). spaCy : Industrial-strength Natural Language Processing in Python. DOI : 10.5281/zenodo.1212303.

LAPARRA E., MASCIO A., VELUPILLAI S. & MILLER T. (2021). A review of recent work in transfer learning and domain adaptation for natural language processing of electronic health records. *Yearb Med Inform*, **30**, 239–244. DOI : 10.1055/s-0041-1726522.

LAVERGNE T., CAPPÉ O. & YVON F. (2010). Practical very large scale CRFs. In *Proc of ACL*, p. 504–513, Uppsala, Sweden.

L'HOMME M.-C. (2008). Ressources lexicales, terminologiques et ontologiques : une analyse comparative dans le domaine de l'informatique. *Revue française de linguistique appliquée*, **13**, 97–118. DOI : 10.3917/rfla.131.0097.

LINDBERG D. A., HUMPHREYS B. & MCCAY A. (1993). The Unified Medical Language System. *Methods Inf Med*, **4**(32), 281–91. DOI : 10.1055/s-0038-1634945.

MANESSE D. (2021). Les grands écarts de l'écriture inclusive. Entre l'amour de la langue et l'amour de moi, moi, moi. *Cités*, **86**(2), 71–86. DOI : 10.3917/cite.086.0071.

MARCHELLO-NIZIA C. (1989). Le neutre et l'impersonnel. *Linx*, (21), 173–179. Actes du colloque tenu à Paris X-Nanterre les 14-15-16 décembre 1988, DOI : 10.3406/linx.1989.1139.

MAUREL D. (2004). Les mots inconnus sont-ils des noms propres ? In *Actes des JADT*, Louvain-la-Neuve, Belgique.

MEKKI J., BATTISTELLI D., BÉCHET N. & LECORVÉ G. (2021). TREMoLo : un corpus multi-étiquettes de tweets en français pour la caractérisation des registres de langue. In *Actes de TALN*, p. 237–245, Lille, France. HAL : hal-03265873.

MILLER G. A., BECKWITH R., FELLBAUM C., GROSS D. & MILLER K. J. (1990). Introduction to WordNet : An on-line lexical database. *International Journal of Lexicography*, **3**(4), 235–244. DOI : 10.1093/ijl/3.4.235.

OPREAN C., MOKBEL C., LIKFORMAN-SULEM L. & POPESCU A. (2014). Reconnaissance de mots manuscrits hors-vocabulaire en utilisant des ressources web. *Document numérique*, **17**(3), 77–96. DOI : 10.3166/dn.17.3.77-96, HAL : cea-01822860v1.

RABARY C. T., LAVERGNE T. & NÉVÉOL A. (2015). Etiquetage morpho-syntaxique en domaine de spécialité : le domaine médical. In *Actes de TALN*, Caen, France.

RIBAN C. & GERIN M. (2017). Les garçons et les filles sont belles. https://www.youtube.com/watch?v=8X45yYIF1Gw.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

SMITH B. & FELLBAUM C. (2004). Medical WordNet : A new methodology for the construction and validation of information resources for consumer health. In *Proc of Coling*, p. 31–38, Geneva, Switzerland.

SPRIET T., BÉCHET F., EL-BÈZE M., DE LOUPY C. & KHOURI L. (1996). Traitement automatique des mots inconnus. In *Actes de TALN*, Marseille, France.

STOUTEN F., ILLINA I. & FOHR D. (2010). Regroupement des occurrences des mots hors-vocabulaire répétés en vue de leur modélisation pour la transcription d'émissions radio. In *Actes des JEP*, Mons, Belgique.

STRAPPARAVA C. & VALITUTTI A. (2004). WordNet affect : an affective extension of WordNet. In *Proc of LREC*, Lisbon, Portugal.

VOSSEN P. (1997). EuroWordNet : a multilingual database for information retrieval. In *Proceedings of the DELOS workshop on Cross-language Information Retrieval*, Zürich, Switzerland.