SHS Web of Conferences **232**, 11004 (2026) https://doi.org/10.1051/shsconf/202623211004
Congrès Mondial de Linguistique Française - CMLF 2026

Identification automatique de la valeur de *on* dans un corpus d'apprenants du français L2

Iris Eshkol-Taravella[1*], *Sülün* Aykurt-Buchwalter[1], *Keren* Dague[2], *Juliette* Henry[2] & *Maïwenn* Plevenage[2]

[1]MoDyCo UMR 7114 Université Paris Nanterre, France
[2]Université Paris Nanterre, France

* Corresponding author : ieshkolt@parisnanterre.fr

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

Abstract. Le pronom indéfini *on* en français se caractérise par la polysémie et l'ambiguïté : son référé doit être interprété selon le contexte. Cette interprétation est parfois difficile, a fortiori lorsque le discours est produit par un locuteur non-natif. Cette étude, fondée sur l'apprentissage supervisé, explore la capacité d'un classifieur entraîné sur un corpus d'apprenants à identifier automatiquement la valeur du pronom *on*. Deux valeurs sont prises en compte : l'usage indéfini et l'usage où *on* fonctionne comme substitut du pronom *nous*. Deux approches d'apprentissage supervisé sont comparées : un modèle d'apprentissage de surface reposant sur une régression logistique, et un modèle d'apprentissage profond basé sur CamemBERT, affiné (*fine-tuned*) sur le corpus. Les résultats obtenus sont encourageants et indiquent que ces approches permettent d'atteindre des performances satisfaisantes pour la tâche étudiée. L'analyse montre notamment que le classifieur parvient à distinguer les contextes où *on* se substitue à *nous*, correspondant généralement à un registre familier, de ceux où il renvoie à une entité indéfinie. Le module automatique développé pourrait être utilisé pour analyser de plus grands corpus, ou pour des évaluations pédagogiques.

1 Introduction

Le pronom indéfini *on* en français se caractérise par la polysémie et l'ambiguïté, en ce sens que son référé doit être inféré dans le contexte du discours. *On* est d'abord considéré comme « une troisième personne indéfinie dont la référence, à l'intérieur de la sphère de la non-personne, désigne un être humain indéterminé dans son identité et son nombre » [1] (p. 103). Pour [2] (p.23), « ne référant à personne spécifiquement, il peut tout aussi bien désigner "tout le monde y compris moi", "n'importe qui", "personne", etc. ». Il s'agit là de la valeur indéfinie de *on*. Cependant, dans certains contextes discursifs, *on* peut signifier *nous*. L'utilisation de *on* pour désigner *nous* est souvent attribuée à un registre familier et à une modalité orale [1]. Toutefois, l'emploi de *on* comme substitut de *nous* est également attesté dans des textes écrits [3]. Ainsi, le pronom *on* peut changer de sens selon son emploi dans le discours, et c'est au récepteur du discours d'en interpréter le sens. Cette opération mobilise des inférences basées, entre autres, sur le contexte [4]. L'interprétation de certaines occurrences de *on* peut pourtant être difficile [2]. D'ailleurs, des travaux dans le domaine de l'acquisition du français langue étrangère (L2) ont montré que l'emploi adapté de *on* constitue un défi pour les apprenants, et que les scripteurs en français L2, en comparaison avec les scripteurs natifs, ont recours à des emplois de *on* dont la valeur est plus difficile à interpréter [5].

Étant donné cette double difficulté liée à l'interprétation – même par un humain – de *on* dans le discours en français et a fortiori dans des textes rédigés par des scripteurs non-natifs, cette étude investigue la possibilité d'une caractérisation automatique de la valeur du pronom *on* dans des écrits d'apprenants du français L2. La disponibilité de nouveaux corpus d'apprenants, tels que TCFLE-8 [6], rend possible l'entraînement d'un classifieur dans un tel but. Si ce corpus a déjà fait l'objet de travaux en TAL visant notamment l'annotation automatique des erreurs, notre étude adopte une approche différente, dans la mesure où la plupart du temps, l'emploi de *on* ne constitue pas une erreur mais potentiellement une maladresse liée au registre de langue : ainsi, seule une interprétation en

contexte peut permettre de conclure à un usage adapté ou non. Or, pour conclure à une utilisation inadaptée de *on*, il faudrait d'abord que l'identification automatique de la valeur de *on* (en particulier la discrimination entre sa valeur indéfinie et sa valeur comme substitut de *nous*) dans un contexte donné soit possible. La méthodologie adoptée pour l'étude est fondée sur l'apprentissage supervisé, une méthode fréquemment utilisée dans le domaine du TAL pour les tâches de classification automatique.

Notre recherche a donc pour objectif d'établir dans quelle mesure un classifieur entraîné à partir d'un corpus d'apprenants qui comprend de nombreuses erreurs et maladresses peut identifier automatiquement la valeur du pronom *on* (usage indéfini ou usage comme substitut de *nous*) en discours dans des textes rédigés par des apprenants de L2. Après avoir mis en évidence les enjeux liés à l'interprétation du pronom *on* en français, nous présenterons le corpus TCFLE-8 et la chaîne de traitement mise en place pour cette recherche, afin d'évaluer sa performance.

2 Cadrage théorique : L'interprétation du pronom *on* en français

En français contemporain, *on* est généralement considéré comme un pronom indéfini, qui peut se référer, entre autres, à « tout le monde », « quelqu'un » ou encore « certaines personnes » [1], [2]. En plus de cette valeur indéfinie, *on* peut être employé comme substitut de *nous*. Cet usage est souvent associé à un registre de langue bien spécifique : pour certains, il caractérise « la langue orale familière » [1], même si l'utilisation de *on* comme substitut de *nous* est également attestée dans des corpus écrits [3]. Ainsi, une des principales caractéristiques du pronom *on* est que son sens peut changer selon son emploi dans le discours, et que sa valeur comme substitut de *nous* est étroitement liée au registre employé.

C'est donc par des opérations d'interprétation que le récepteur du discours peut identifier le référé du pronom *on*. Principalement, il s'agit pour le récepteur de réaliser une inférence, définie comme un « mécanisme cognitif par lequel le récepteur d'un message interprète (...) un sens qu'il tire des éléments qui ont été énoncés, soit en les combinant entre eux, soit en faisant appel à des données de l'entourage linguistique (...) » [4] (p. 37). Or, la valeur de *on* n'est pas toujours évidente à interpréter, même dans le contexte ; par exemple, si le lecteur tente des opérations de substitution, il apparaît que le référé de *on* peut tout aussi bien inclure le scripteur que l'exclure [2]. Il se peut que cette ambiguïté soit intentionnelle : le choix de *on* plutôt qu'un autre pronom pourrait par exemple permettre de déresponsabiliser l'énonciateur dans une prise de position [7].

Des travaux dans le domaine de l'acquisition des L2 ont montré que les apprenants du français L2 n'emploient pas le pronom *on* de la même façon que les locuteurs natifs, ce qui pourrait être attribué à difficultés rédactionnelles plus globales, liées à la maîtrise des genres écrits, au positionnement du scripteur ou à la gestion des chaînes anaphoriques. Il apparaît par ailleurs que l'utilisation de *on* est plus difficile à interpréter, en contexte, dans les écrits d'apprenants, que dans les textes rédigés par des natifs [5].

Une utilisation inadaptée de *on*, et particulièrement son emploi comme substitut de *nous* dans certains genres de textes écrits, pourrait contribuer à un registre de langue trop informel. L'emploi d'un registre de langue approprié est un enjeu important pour les apprenants, particulièrement aux niveaux avancés. La capacité à identifier différents registres de langue et de les utiliser à bon escient est associée à des compétences rédactionnelles et socio-culturelles attendues au niveau B2 du CECRL. Or, l'appropriation des formes adaptées aux différentes situations de communication est un processus long et complexe chez les locuteurs natifs d'une langue donnée ; a fortiori, chez un apprenant de L2, la maîtrise des registres de langue implique de nombreux défis [8]. Dans un contexte où

l'automatisation joue un rôle croissant dans l'enseignement/apprentissage des langues, il est pertinent de se pencher sur la possibilité d'identifier de manière automatisée les usages adaptés et inadaptés de certaines formes linguistiques selon le registre de langue attendu.

3 Le corpus TCFLE-8

3.1. Présentation du corpus

Cette étude porte sur TCFLE-8, un corpus de productions écrites en français L2 composé de 6 569 textes rédigés dans le cadre du Test de connaissance du français opéré par France Education International. Les scripteurs sont des candidats dont le niveau varie entre le niveau A1 du Cadre européen commun de référence pour les langues (CECRL) et le niveau C2. Leurs langues maternelles (L1) sont l'anglais, l'arabe, le chinois, l'espagnol, le japonais, le kabyle, le portugais et le russe [6]. Les textes sont rédigés à partir de trois types de consignes. Le premier type de tâche est un court message descriptif ou informatif ; le deuxième est un récit ou compte-rendu ; enfin, le troisième type de tâche est un texte plus long qui compare deux points de vue sur un sujet de société.

Ce corpus contient un grand nombre d'erreurs et d'autres phénomènes caractérisant l'interlangue des apprenants, plus particulièrement aux niveaux débutants et intermédiaires. Il s'agit de phénomènes variés, allant d'erreurs d'orthographe ou de morphosyntaxe à l'alternance codique, et qui peuvent nuire à l'interprétation des énoncés et à la compréhension des textes. À titre d'exemple, nous reprenons ci-dessous un texte rédigé dans le contexte de la tâche de type 1 et évalué au niveau A1 :

(1) Salut Je vouz propone rendez-vous cet week- end, Je veux si vous pouvez asister avec tu famillie, je le dis dejà a mon frère y ma mère Ce sortie sera especial, On aurait jouex et others. Nous irons avec nous petit fille.Vous pouvez igalment apporte des enfant si vous voulez. Je suis atentive pour sa reponse, Reponse pour moi portable.

Le corpus TCFLE-8 a fait l'objet d'études récentes dans le domaine du TAL, notamment en ce qui concerne l'annotation automatique des erreurs et la correction automatisée des écrits d'apprenants en fonction des descripteurs du CECRL [6]. Cependant, la forme sur laquelle nous proposons de travailler dans cette étude n'entre pas dans le cadre de l'analyse d'erreurs, dans la mesure où l'emploi inapproprié de *on* s'apparente davantage à une maladresse. Pour pouvoir identifier une utilisation inadaptée de *on*, il faudrait d'abord que l'identification automatique de la valeur de *on* dans un contexte donné soit possible.

3.2 Prétraitement du corpus TCFLE-8

Le corpus se présente sous la forme de deux fichiers CSV distincts. Le premier rassemble les productions des participants, accompagnées de leur identifiant unique et du niveau de maîtrise du français qui leur a été attribué par les évaluateurs d'après leur production écrite (de A1 à C2). Le second fichier (« metadata ») contient les métadonnées et les informations sociodémographiques et linguistiques des candidats, telles que leur identifiant, leur genre, leur L1, ainsi que le type de tâche d'écriture et le niveau CECRL.

L'objectif du prétraitement était d'unifier, de nettoyer et de structurer ces informations afin de les rendre directement exploitables pour les analyses ultérieures. Cette étape a nécessité d'une part, une harmonisation des formats et d'autre part, une sélection des informations pertinentes pour la recherche.

Dans un premier temps, les deux fichiers CSV ont été fusionnés de manière à ne conserver que les champs pertinents : identifiant du participant, L1, niveau de français et textes produits en réponse aux consignes. La procédure a consisté à extraire, depuis le fichier contenant les réponses, l'ensemble des lignes correspondant à chaque niveau de compétence (par exemple toutes les lignes de niveau A1), puis à exporter ces sous-ensembles vers de nouveaux fichiers CSV distincts. Ces différentes tables ont ensuite été importées dans une base de données SQLite, de même que la table metadata. Chaque table correspondant à un niveau a été reliée à la table metadata via l'identifiant des participants, de manière à enrichir les réponses des informations contextuelles nécessaires évoquées précédemment. Les tables obtenues après jointure ont finalement été exportées au format CSV pour servir de support pour la tâche d'annotation.

Au terme de ce traitement, nous obtenons six tables, correspondant chacune à un niveau, que nous considérons comme six sous-corpus distincts. Leur répartition est présentée dans le tableau 1. On observe des variations de taille entre ces sous-corpus, qui s'expliquent par le nombre inégal de candidats au Test de connaissance du français selon les niveaux, les effectifs étant plus réduits aux niveaux les plus faibles et les plus avancés. Précisons que seules certaines parties du corpus ont été retenues pour la présente étude ; les critères et choix méthodologiques correspondants seront détaillés ultérieurement.

Table 1. Les sous-corpus de TCFLE-8 par niveaux CECRL

Niveaux	A1	A2	B1	B2	C1	C2
Nombre de lignes	689	1375	1466	1427	1127	485

4 Méthodologie et résultats

4.1 Démarche générale

L'objectif de la chaîne de traitement mise en place est de caractériser automatiquement le pronom *on* selon deux emplois définis : *indéfini* et *nous*. La méthodologie adoptée est fondée sur l'apprentissage supervisé, une méthode fréquemment utilisée dans le domaine du TAL pour les tâches de classification automatique. Cette méthodologie est composée de plusieurs étapes décrites dans les sections suivantes : annotation manuelle portant sur le pronom personnel *on* et son évaluation, l'entraînement de système de la classification automatique et son évaluation.

4.2 Annotation manuelle

L'annotation manuelle a été effectuée sur les trois corpus correspondant aux niveaux A1, A2 et B2. Les niveaux C ont été jugés moins pertinents pour notre recherche car ils correspondent aux écrits d'apprenants ayant une très bonne maîtrise du français.

L'annotation consistait dans l'ajout de balises autours du pronom *on* en le catégorisant selon deux types d'emploi : le *on indéfini* (avec les balises on) et le *on* substitut de *nous* (avec les balises <n>on</n>) dans un fichier Excel créé suite au prétraitement décrit dans la section 3.2. Les annotatrices2, étudiantes en Master TAL et locutrice natives du français, ont été formées à la tâche par un expert linguiste et ont rédigé le guide

[2] Les auteures remercient Pauline Fillols pour ses contributions au travail d'annotation et d'analyse.

d'annotation suite aux observations initiales du corpus et sur la base des travaux antérieurs [1], [2], [9].

Un certain nombre d'indices linguistiques ont été répertoriés dans le guide d'annotation, bien que le recours à ces indices n'ait pas toujours été nécessaire pour l'annotation. Par exemple, le pronom *on* était balisé <n> dans les phrases lorsqu'il reprenait le pronom *nous*, comme dans l'exemple (2) ou lorsqu'il renvoyait sans ambiguïté à un groupe dont faisait partie le scripteur, comme dans l'exemple (3).

> (2) [...] donc pous nous <n>**on**</n> préfére voyager quand nous sommes seul pour avoir un bonne voyage. (sous-corpus A2)

> (3) [...] jai ami il sappllet psacal <n>**on**</n> est ami ca fait pluis 10 ans (sous-corpus A1)

En revanche, il était balisé dans des constructions où *on* est suivi d'un verbe d'opinion ou d'assertion, comme dans l'exemple (4) :

> (4) on dit que l'étanger fait grandire ces vrai paresque je lai vecu (sous-corpus A2)

Lors de l'annotation, nous avons pu rencontrer des difficultés quant à l'interprétation du pronom. Dans certains cas, il n'était pas possible de trancher clairement entre les deux étiquettes, soit parce que les deux interprétations semblaient plausibles en contexte (5), soit parce que les formulations en français des apprenants rendaient la compréhension difficile, comme dans l'exemple (6).

> (5) j'ame bien reste à la maison pour organizer mon plannigue de la semaine,faire les devoir avec les enfants,arranger des placar ,faire une bon repar,regarder un bon filme,enfain profité meme si je suis à la maison ,on trouvè toujours une chouse à faire,j'adore etrè à la maison (sous-corpus A2)

> (6) beacoup des enfant aimer au joué tous le temp avec a les amis à l'ecole et des parents, mais on vou va à joue avec un chien ou un chat. (sous-corpus A2)

En effet, dans l'exemple (5), *on* pourrait se référer à un groupe de personnes faisant partie de la famille du scripteur ; toutefois, la présence de l'adverbe *toujours*, indiquant une généralisation de l'énoncé, laisse ouverte la possibilité d'un étiquetage *indéfini*. Dans l'exemple (6), en revanche, le référé ne peut être interprété en raison du contexte morphosyntaxique erroné.

Ainsi, nous avons décidé d'écarter ces occurrences ambiguës de notre recherche, au vu du faible nombre qu'elles représentaient (seulement 1,1% du corpus total). En effet, conserver une classe si petite n'aurait pas fourni assez d'exemples nécessaires au classifieur pour s'entraîner et aurait déséquilibré les résultats. Au total, 274 occurrences du pronom on ont été annotés pour le niveau A1, 767 pour le niveau A2 et 1175 pour le niveau B2.

4.3 Évaluation de l'annotation manuelle

L'évaluation de l'annotation manuelle a porté uniquement sur les niveaux A1 et A2. Ce choix s'explique par la nécessité de vérifier la cohérence des annotations sur les productions les plus susceptibles de contenir des erreurs linguistiques importantes pouvant

nuire à l'interprétation et à la compréhension. L'hypothèse sous-jacente est que si un bon niveau d'accord inter-annotateur est obtenu sur ces niveaux débutants, malgré la fréquence élevée d'erreurs dans les productions, l'accord sera au moins équivalent, sinon supérieur, pour les textes du niveau B2.

L'accord inter-annotateurs a été calculé au moyen du coefficient kappa de Cohen [10] pour chaque paire d'annotatrices. La moyenne des coefficients obtenus a ensuite été utilisée afin d'estimer l'accord global pour chaque niveau. Les valeurs moyennes de kappa obtenues sont de 0,984 pour le niveau A1 et de 0,969 pour le niveau A2. Selon l'échelle d'interprétation généralement admise dans la littérature [11] et qui est également considérée comme valable pour les recherches sur les corpus d'apprenants [12], ces valeurs correspondent à un accord presque parfait, indiquant une forte homogénéité dans le travail d'annotation.

4.4 Entraînement de classifieurs

Pour l'entraînement des classifieurs, nous avons sélectionné exclusivement les textes du niveau B2, qui contenaient au moins une occurrence de *on*, ce qui représente 1175 textes. Ce choix s'explique par les observations issues de l'annotation manuelle : à partir de ce niveau, conformément aux descripteurs du CECRL, les apprenants sont généralement en mesure de produire des textes argumentatifs portant sur des sujets complexes et abstraits, en mobilisant et en comparant différents points de vue, ce qui n'est pas le cas aux niveaux inférieurs. Ainsi, dans le sous-corpus B2, les occurrences de *on* sont plus variées, en effet la proportion de *on indéfini* est pratiquement égale à celle des *on nous* contrairement aux autres niveaux ou la proportion de *on nous* représente en moyenne 69,5 % du corpus. De plus, le niveau A contient de nombreuses erreurs de différents types, elles peuvent être lexicales, grammaticales, syntaxiques ou morphologiques. Ce bruit brouille les indices linguistiques utiles au modèle et rend difficile l'apprentissage : le classifieur reçoit des signaux incohérents et apprend moins bien. En travaillant uniquement avec le niveau B2, nous évitons les bruits et conservons un corpus où les deux usages de *on* sont mieux équilibrés, plus exploitables et surtout appuyés par un volume de données plus important. En effet, le sous-corpus B2 est plus grand en terme en nombre de mots (voir le tableau 1), ce qui le rend plus propice à l'entraînement du classifieur.

Afin de préparer les données, nous avons extrait pour chaque occurrence du pronom *on* un contexte gauche et un contexte droit. La fenêtre contextuelle a été fixée à dix mots à gauche et à droite. Ce contexte peut toutefois être réduit lorsque le script rencontre un signe de ponctuation finale, auquel cas l'extraction s'arrête à ce point. Ce sont ces contextes qui sont donnés aux classifieurs, pour les phases d'entraînement et de test. Les classifieurs se basent sur ce qui précède et suit le pronom pour prédire sa catégorie. Les tokens ont été directement vectorisés en utilisant le modèle TF-IDF de scikit-learn. Nous avons formulé la tâche sous la forme d'une classification binaire distinguant deux usages : *on* à valeur de *nous* et *on* à valeur *indéfini*.

Le premier modèle testé est une régression logistique en raison de ses capacités générales : il gère efficacement la classification binaire et offre un modèle simple et robuste même avec peu de données. Ce choix garantit une approche claire, stable et adaptée à une première étude. Le corpus de référence a été divisé en deux sous-corpus : 80 % des données ont été utilisées pour l'entraînement, et 20 % pour le test.

Les résultats obtenus sont présentés dans le tableau ci-dessous. Une précision et un rappel varient entre 0,84 et 0,94. La détection de deux usages obtient presque la même F-mesure, par contre l'usage de *on* comme substitut de *nous* obtient la meilleure précision alors que le *on indéfini* a le meilleur rappel. Ceci peut s'expliquer par la distribution et la nature linguistique des deux usages. L'usage de *on* comme substitut de *nous* apparaît

généralement dans des contextes plus reconnaissables (registre familier, indices lexicaux ou syntaxiques typiques), ce qui facilite des prédictions précises. À l'inverse, l'usage *indéfini* est plus fréquent et apparaît dans des contextes variés ; le modèle le détecte donc plus souvent, ce qui augmente le rappel. Les deux F-mesures proches montrent que ces avantages se compensent globalement.

Table 2. Résultats d'évaluation du modèle régression logistique

	nous	***indéfini***
Rappel	0,85	0,93
Précision	0,94	0,84
F-mesure	0,90	0,88

La matrice de confusion révèle la principale marge d'amélioration, concernant le rappel de la classe *nous*. Les 19 faux négatifs identifiés indiquent une tendance du modèle à être prudent. En cas d'ambiguïté, le classifieur semble privilégier l'étiquette *indéfini*.

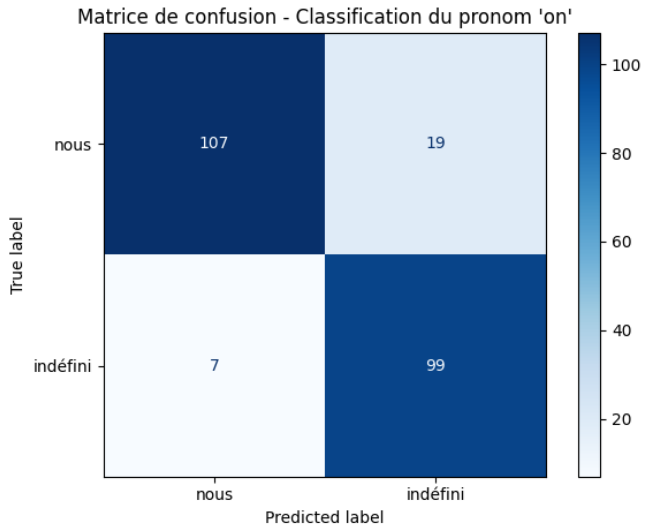

Fig. 1. Matrice de confusion pour la régression logistique

Le système développé permet de prédire la valeur de *on* dans une phrase donnée même si elle contient de nombreuses erreurs dans le contexte immédiat du pronom *on*, ce qui est le cas fréquent dans le corpus des apprenants :

> (7) Selon les études sur les effets que les jeux-vidéos a produit sur l'homme, qui signifie que les jeux-vidéos pouvent pousser le cerveau de l'adulte plus développé, à moins que on ne jousse pas les jeux ineffice pour le développement de l'homme. (sous-corpus B2)

Étant donné les bons résultats obtenus avec la régression logistique, nous l'avons utilisée comme base line pour la suite et avons décidé d'entraîner un modèle CamemBERT [13]. En effet, CamemBERT est particulièrement adapté à cette tâche car il exploite le contexte complet de la phrase et capture des nuances syntaxiques et sémantiques. Nous avons procédé à un finetuning du modèle avec 80% des données pour l'entraînement, 10% des données pour la validation et enfin les derniers 10% des données pour l'évaluation.

La matrice ci-dessous représente les résultats obtenus pour le modèle CamemBERT. On observe que le modèle a correctement identifié 58 cas comme *indéfini* et 48 cas comme *nous*. Cependant, il a commis quelques erreurs : 3 cas ont été prédits comme *nous* alors qu'ils étaient en réalité *indéfini* (faux positifs), et 7 cas ont été prédits comme *indéfini* alors qu'ils appartenaient à la classe *nous* (faux négatifs).

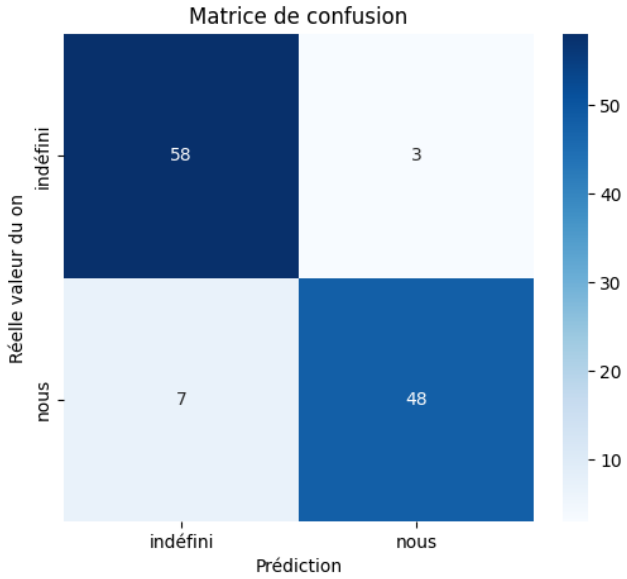

Fig. 2. Matrice de confusion pour CamemBERT

Nous avons obtenu une exactitude de 0.91 sur les données d'évaluation, ce qui est un très bon résultat. La figure ci-dessous reprend les calculs de rappel, précision et F-mesure. Nous obtenons la même structure que pour le modèle de régression logistique, l'usage de *on* comme substitut de *nous* dans un registre familier obtient la meilleure précision alors que le *on indéfini* a le meilleur rappel.

Table 3. Résultats d'évaluation du modèle Camembert

	nous	***indéfini***
Rappel	0.87	0,95

Précision	0,94	0,89
F-mesure	0,90	0,92

Nous proposons dans les exemples (8) et (9) des cas où le classifieur a fait une identification erronée et prédit une occurrence de *on* comme étant un substitut de *nous*, alors que l'annotation manuelle avait attribué l'étiquette *indéfini* (faux positif) :

(8) d' affilée , mais aucun nettyoage n' aidra si on continue à cette lancé…

(9) pu rencontrer ma résponsable elle est souvant en déplacement on m' a dit) et les autres membres de…

A l'inverse, dans l'exemple 10, le classifieur a prédit un *on indéfini* au lieu d'un substitut de *nous*, malgré la présence d'un indice linguistique fort, à savoir la reprise avec le pronom *nous* (faux négatif) :

(10) particularité de cet événement réside dans le fait qu on va , nous les éléves , accullir les éléves…

Ainsi, bien qu'il subsiste quelques erreurs, surtout au niveau des *on indéfinis* où le classifieur a fait le plus d'erreurs (7 faux positifs), nous obtenons de très bons résultats. On peut également remarquer que notre modèle avec régression logistique a un moins bon rappel pour le *on* substitut de *nous* que pour le *on indéfini*.

5 Conclusion

Le travail présenté constitue une première étape dans l'exploitation d'une partie du corpus TCFLE-8 au sein d'un programme de recherches plus large. Il met en évidence la possibilité de prédire la valeur d'un pronom personnel potentiellement ambigu, en mobilisant des méthodes d'apprentissage supervisé, ici dans le cadre d'une classification binaire. Les résultats obtenus sont encourageants et montrent que ce type d'approche permet d'atteindre des performances tout à fait satisfaisantes pour la tâche étudiée. L'étude montre que le classifieur est capable d'identifier les contextes où *on* se substitue à *nous*, ce qui correspond à un registre de langue familier, et les contextes où il se réfère à une entité indéfinie.

Aucune étude, à notre connaissance, ne combine une annotation fine de *on* (avec des catégories d'usage) et une classification automatique centrée sur ce pronom. Les résultats obtenus (f-mesure 0,89) sont très bons pour une tâche de classification binaire sur un pronom ambigu, ce qui montre que même des modèles simples (régression logistique) suffisent pour capturer une grande partie du signal sémantique/pragmatique. Compte tenu des travaux de [14] qui ont développé un modèle automatique (deep learning) pour détecter des usages « atypiques » d'indéfinis (*they*, *someone*…) chez des scripteurs non natifs en anglais, nos résultats indiquent que ce type de tâche est réalisable dans différentes langues. Bien que le développement de l'outillage informatique soit une tendance nette dans les études sur les corpus d'apprenants [15], les analyses manuelles restent dominantes dans les travaux actuels, en particulier pour analyser certaines dimensions du langage. Notre outil automatique pourrait être utilisé pour analyser de plus grands corpus, ou pour des

évaluations pédagogiques (ex. : repérer automatiquement les usages non natifs/inhabituels de on dans des rédactions d'apprenants, permettant ainsi d'évaluer si le registre de langue employé est adapté à la tâche d'écriture).

Références

1. E. Le Bel, Le statut remarquable d'un pronom inaperçu. La Linguistique, **27(2)**, 91‑ 109(1991).
2. F. Atlani, ON l'illusionniste, in A. Grésillon & J.-L. Lebrave, La langue au ras du texte, Lille : Presses Universitaires de Lille, 13‑ 29, (1984).
3. M. Delaborde, F. Landragin, En quoi le pronom « on » a-t-il une valeur anaphorique? Le cas des successions d'occurrences de « on ». Cahiers de praxématique, **(72)**, 1-18, (2019). https://journals.openedition.org/praxematique/5464
4. P. Charaudeau, Compréhension et interprétation. Interrogations autour de deux modes d'appréhension du sens dans les sciences du langage, in Achard-Bayle, G. & al. Les Sciences du langage et la question de l'interprétation (aujourd'hui), Fribourg : Lambert-Lucas, 21-55, (2018). https://www.patrick-charaudeau.com/IMG/pdf/-5.pdf
5. S. Aykurt-Buchwalter, L'acquisition du pronom on en FLE : une étude sur les essais argumentés d'apprenants turcophones. Langue française, **218(2)**, 73-88, (2023). https://shs.cairn.info/revue-langue-francaise-2023-2-page-73?lang=fr
6. R. Wilkens, A. Pintard, D. Alfter, V. Folny & T. François, TCFLE-8 : A Corpus of Learner Written Productions for French as a Foreign Language and its Application to Automated Essay Scoring, in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 3447‑ 3465, (2023). https://aclanthology.org/2023.emnlp-main.210/
7. C. Perelman & L. Olbrechts-Tyteca, Traité de l'argumentation (6e édition), Bruxelles : Éditions de l'Université de Bruxelles, (2008).
8. D. Biber, S. Conrad, Register, genre, and style, (Cambridge : Cambridge University Press, 2019).
9. J. Jacquin, Le pronom ON dans l'interaction en face à face : Une ressource de (dé)contextualisation. Langue française, **193(1)**, 77‑ 92, (2017). https://doi.org/10.3917/lf.193.0077
10. J. Cohen, A Coefficient of Agreement for Nominal Scales, Educational and Psychological Measurement, **20(1)**, 37–46, (1960).
11. J.R. Landis, G.G. Koch, The measurement of observer agreement for categorical data. Biometrics, **33(1)**, 159-74, (1977).
12. T. Larsson, M. Paquot, L. Plonsky, Inter-rater reliability in Learner Corpus Research : Insights from a collaborative study on adverb placement. International Journal of Learner Corpus Research, **6(2)**, 237‑ 251, (2020).
    https://www.jbe-platform.com/content/journals/10.1075/ijlcr.20001.lar
13. L. Martin, B. Muller, P.J. Ortiz Suárez, Y. Dupont, L. Romary, E. Villemonte de La Clergerie, D. Seddah, B. Sagot, CamemBERT : a Tasty French Language Model, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 7203–7219, (2020). https://aclanthology.org/2020.acl-main.645/
14. E. Rabinovich, J. Watson, B. Beekhuizen, S. Stevenson, Say Anything : Automatic Semantic Infelicity Detection in L2 English Indefinite Pronouns, in Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), Hong Kong, Chine. Association for Computational Linguistics, 77–86, (2019). https://aclanthology.org/K19-1008/
15. S. De Cock, H. Tyne, Corpus d'apprenants et acquisition des langues. Recherches en didactique des langues et des cultures, **11(1),** Article 1, (2014). https://journals.openedition.org/rdlc/1716