Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

Le corpus LN-ATALA : 25 ans d'annonces du traitement automatique des langues modérées et catégorisées

Rémi Cardon[1, 3 *] Gaël Guibon[2, 3 *]
(1) Computer Science & Engineering Department, UC3M, Madrid, Espagne
(2) Université Sorbonne Paris Nord, CNRS, Laboratoire d'Informatique de Paris Nord, LIPN, F-93430 Villetaneuse, France
(3) ATALA, 45 rue d'Ulm, 75005 Paris, France
rcardon@inf.uc3m.es, guibon@lipn.fr

RÉSUMÉ

LN-ATALA est une liste de diffusion consacrée aux annonces en traitement automatique du langage naturel (TAL). Créée en 1990 et parrainée par l'association française ATALA, elle constitue depuis plus de trente ans un important canal de communication pour la communauté. Une grande partie des e-mails envoyés aux abonnés depuis 1999 est disponible en ligne. Dans cet article, nous présentons cette liste et décrivons la construction d'un jeu de données exploitable rassemblant les e-mails envoyés entre 1999 et 2024, afin de faciliter l'analyse de l'évolution du domaine du TAL au fil du temps. La ressource comprend 17 824 e-mails, répartis en 21 catégories annotées, et sera librement mise à disposition pour un usage non commercial et maintenue dans le temps.

ABSTRACT

The LN-ATALA Dataset : 25 years of Manually Moderated and Labeled NLP-related Announcements

LN-ATALA is a mailing list dedicated to announcements in natural language processing (NLP). Created in 1990 and sponsored by the French NLP association ATALA, it has served as an important communication channel for the community for more than three decades. A large portion of the emails sent to subscribers since 1999 is publicly available online. In this paper, we present the mailing list and describe the construction of a fully usable dataset gathering emails sent between 1999 and 2024, enabling the analysis of the evolution of the NLP landscape over time. The resource contains 17,824 emails grouped into 21 annotated categories and will be freely available for non-commercial use and maintained over time.

MOTS-CLÉS : corpus, liste de diffusion, e-mails, classification.

KEYWORDS: corpus, mailing list, e-mails, classification.

1 Introduction

Les listes de diffusion prennent différentes formes en fonction de leur objectif. Elles peuvent se retrouver sous une liste dédiée à un partage d'information bilatéral ou unilatéral. Ces listes sont souvent modérées automatiquement ou manuellement afin d'éviter la diffusion de spam mais ne

∗. Contribution égale.

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 1 : articles scientifiques originaux, pages 350–368.

comportent habituellement pas d'enrichissement manuel des e-mails envoyés. Bien que des travaux de recherche aient étudié les e-mails (Klimt & Yang, 2004) ou même les listes de diffusion (Wiese *et al.*, 2016), les chercheurs se retrouvent face à des difficultés dues à la variation de formats de messages, de sujets et de contextes. En pratique, étiqueter un corpus de mails a posteriori est coûteux. Avec ce travail, nous proposons un corpus d'une liste de diffusion modérée manuellement depuis 35 ans : LN-ATALA. Chaque message de corpus est étiqueté avec son type attribué manuellement lors de la diffusion. La liste LN-ATALA est la liste de diffusion française de référence pour les annonces liées au traitement automatique des langues (TAL), comme les offres d'emploi, les appels à communication et d'autres types d'annonce [1]. Bien que la liste soit basée en France, les messages en anglais y sont également diffusés en grand nombre.

Dans cet article, nous présentons la première version de ce corpus voué à être maintenu dans le temps. Bien qu'une grande partie des archives soit disponible en ligne, collecter et préparer des données pour en faire un corpus exploitable est une tâche non-triviale. Nous détaillons la méthodologie que nous avons suivie pour faciliter l'usage de ce corpus, 17 000 messages diffusés entre 1999 et 2024, par la communauté. Notre contribution principale est la production et la mise à disposition de ces données sous forme directement exploitable. Nous présentons également une analyse du jeu de données et quelques expériences pour illustrer un usage possible. En mettant à disposition le corpus LN-ATALA, nous permettons à la communauté d'étudier le domaine et son évolution à travers les années.

Nous passons en revue les travaux sur les corpus d'e-mails en Section 2. Dans la section 3, nous décrivons en détail la création de la ressource et ses contenus. Ensuite, nous rapportons des expériences qui illustrent comment la ressource peut être exploitée (Section 5), avant de conclure (Section 6).

2 État de l'art

Analyses de listes de diffusion. Les corpus d'e-mails et de listes de diffusion jouent un rôle central dans la recherche en TAL, permettant une grande variété de tâches comme la classification, la segmentation, la reconnaissance d'actes de dialogue, et le *topic modeling*.

L'un des corpus d'e-mails les plus utilisés est le corpus Enron (Klimt & Yang, 2004), utilisé pour de la classification en dossiers (c.-à-d. le type d'e-mail) avec des machines à vecteur de support à partir des métadonnées (expéditeur, destinataire, objet) et du contenu. Il a également été annoté en acte de dialogue (Taniguchi *et al.*, 2020). Bird *et al.* (2006) explorent des techniques de fouille d'archive d'e-mails, alors que Wiese *et al.* (2016) exploitent le corpus pour la désambiguïsation d'identité (retrouver les différentes adresses appartenant à une même personne).

Bettenburg *et al.* (2009) ont mené une étude empirique sur le traitement des listes de diffusion et ont identifié sept défis : l'extraction de messages, la suppression de doublons, l'identification de langue, la gestion des éléments MIME et des pièces jointes, la détection de citations et de signatures, la reconstruction de fils, et la résolution d'identités. Dans le cas de la liste LN-ATALA, sa modération et son fonctionnement préviennent la plupart de ces problèmes (p.ex. pas de diffusion de pièces jointes, standardisation du format, et il s'agit d'annonces et pas de discussions).

Hu *et al.* (2017) ont analysé 53 648 messages de la liste de diffusion MLA-L [2]. En faisant du *topic modeling* basé sur l'allocation de Dirichlet latente, ils ont analysé la distribution des thèmes et leur

1. https://www.atala.org/liste_ln
2. https://www.musiclibraryassoc.org/page/mlal

évolution de 2000 à 2016.

Pour l'analyse de dialogue, Bevendorff *et al.* (2020a) ont travaillé sur le corpus Webis Gmane (Bevendorff *et al.*, 2020b), et ont introduit un outil nommé Chipmunk (une combinaison de GRU bidirectionnels et de couches de CNN) pour segmenter le contenu des messages. Des chercheurs de l'université de Ryukoku au Japon ont travaillé sur des problématiques de question-réponse à partir d'un corpus d'e-mails en japonais, à des fins de vérification de connaissances (Watanabe *et al.*, 2005) ou de réponse automatique à des questions de type 'Comment' (Nishimura *et al.*, 2008).

Plusieurs autres corpus d'e-mails ont été créés dans d'autres langues que l'anglais. Guenoune *et al.* (2020) ont ainsi publié un corpus de 100 fils d'e-mails professionnels en français anonymisés et annotés pour la résolution d'anaphores. Krieg-Holz *et al.* (2016) ont publié deux corpus en allemand : un grand corpus de plus d'1,5 millions d'e-mails et un plus petit de moins de 1 000 e-mails enrichis avec des données démographiques.

Découpage en blocs. Le découpage en blocs (*zoning*) dénote la segmentation des e-mails en blocs fonctionnels comme les salutations, les signatures, le texte cité, et le corps du message. Cette tâche n'est pas homogène, différentes approches ont été explorées. Jardim *et al.* (2021) ont présenté un corpus d'évaluation multilingue de 625 e-mails en français, espagnol et portugais. Carvalho & Cohen (2004) ont étudié spécifiquement l'extraction de signatures et de réponses. Estival *et al.* (2007) ont proposé un schéma de découpage en 5 blocs (corps, signature, publicité, texte cité, et réponses) à partir d'un corpus de 9 863 e-mails. Lampert *et al.* (2009) ont introduit un modèle de segmentation hiérarchique avec trois blocs de haut niveau qui ont chacun leurs sous-blocs : expéditeur (auteur, salutations, conclusion), message cité (réponse, transfert) et modèle (signature, publicité, avertissement, pièces jointes). Repke & Krestel (2018) proposent une approche à deux niveaux : la séparation de l'en-tête et du corps, puis la segmentation du corps en en-tête, salutations, texte, conclusion, et signature. Leur corpus contient 1 300 e-mails. Bevendorff *et al.* (2020a) appliquent le découpage à un corpus massif de 153 millions d'e-mails, avec des blocs comme le contenu, le texte cité, l'en-tête, la signature, et les sections vides.

3 Le corpus LN-ATALA

3.1 La liste de diffusion LN-ATALA

La liste de diffusion LN-ATALA (*LN* pour *langage naturel*) a été créée en 1990. Elle se focalise sur le domaine du traitement du langage naturel, et est principalement utilisée pour diffuser l'information liée au TAL. Les messages peuvent être en français ou en anglais, et ne doivent pas comporter de pièces jointes. La liste est utilisée pour des informations comme les appels à communications, les annonces de conférences ou de séminaires, les offres d'emploi / de thèse / de stage ou encore les parutions de revues ou de livres. Chaque message est modéré manuellement. La modération consiste non seulement à valider la pertinence de la diffusion (c.-à-d. non-spam et adéquation avec les thèmes de la liste), mais aussi à appliquer un processus de formatage. Chaque e-mail reçoit une étiquette (p. ex. *Appel*, *Stage*, *Job*) qui apparaît dans son sujet. Le reste du sujet est rédigé manuellement, suivant un modèle qui dépend de l'étiquette (p. ex. le sujet commence toujours par l'information de durée pour une offre d'emploi ou de stage, suivi de l'intitulé, suivi de l'employeur et de sa ville). Le jeu d'étiquettes et la rédaction des sujets ont évolué dans leur forme, mais pas dans leur signification qui elle, est restée la même au fil des ans. Le tableau 1 indique les 20 étiquettes actuellement utilisées

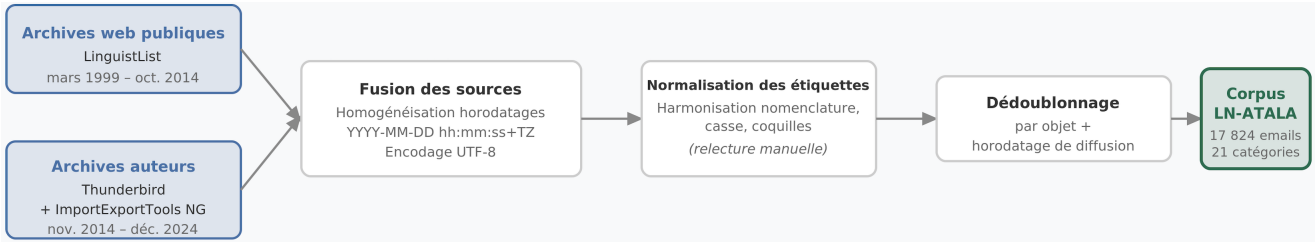

FIGURE 1 – Résumé du processus de création du corpus

et qui ont été appliquées à l'ensemble du corpus, normalisant ainsi les formes des étiquettes. Les messages sont aussi convertis en texte brut encodé en UTF-8 (le formatage HTML est donc supprimé et tous les encodages ou chiffrements sont normalisés).

3.2 Création du corpus

Nous avons collecté les e-mails à partir de deux sources différentes. Les e-mails de mars 1999 à octobre 2014 sont librement disponibles en ligne [3]. Pour les e-mails ultérieurs (de novembre 2014 à décembre 2024), nous avons exporté les archives de nos propres clients e-mail avec une extension de Mozilla Thunderbird appelée ImportExportTools NG [4]. Les e-mails diffusés sont actuellement archivés et disponibles en ligne [5]. Nous n'avons pas exploité cette source, car elle commence en octobre 2019 et nos propres archives couvrent une plus longue période.

Nous avons fait un travail d'intégration des deux sources pour obtenir un fichier unique pour tous les e-mails, qui contient les champs suivants : l'étiquette, l'objet, le gestionnaire de la liste qui a traité/diffusé le message, le nom et l'adresse e-mail de la personne à l'origine de l'annonce, la date et l'heure de diffusion de l'annonce, la date et l'heure d'envoi originales de l'annonce, et le message intégral. Nous détaillons ici les étapes spécifiques qui furent nécessaires pour aboutir à un jeu de données homogène.

Étiquettes. Le jeu d'étiquettes du tableau 1 est celui du corpus, et correspond à celui actuellement utilisé. Ces catégories ont toujours existé, mais leur dénomination a connu des variations. Par exemple l'étiquette *Stage* a parfois été *Offre de stage*. De plus, comme l'attribution de l'étiquette est faite manuellement, des coquilles peuvent apparaître. Enfin, dans l'objet l'étiquette est toujours suivie de deux points, parfois saisis par erreur en point-virgule, ou en virgule. Tous les cas qui ne correspondaient pas au jeu standardisé ont été passés en revue et corrigés manuellement.

Horodatage. Le format des dates et heures dans les archives n'est pas stable, car il dépend de la configuration du client e-mail utilisé pour l'expédition du message original, ainsi que de celui du diffuseur. Nous avons tout homogénéisé vers le format `datetime` suivant :`YYYY-MM-DD hh:mm:ss+TZ`. Dans certains cas, l'horodatage original du message n'était pas disponible. Pour ces cas, nous avons répliqué l'horodatage de la diffusion.

Un même message pouvait apparaître plusieurs fois dans nos sources, nous avons donc filtré le corpus pour supprimer les doublons, sur base de la combinaison de l'objet et de l'horodatage de diffusion.

3. https://listserv.linguistlist.org/pipermail/ln/
4. https://services.addons.thunderbird.net/en-US/thunderbird/addon/importexporttools-ng/
5. https://groupes.renater.fr/sympa/arc/ln

La figure 1 résume le processus de création du corpus à partir des sources.

3.3 Statistiques du corpus

Après les étapes décrites précédemment, nous avons obtenu 17 824 messages. La distribution par étiquette est visible au tableau 1 et la distribution par année en Figure 2. Pour chaque catégorie, le tableau montre les informations suivantes : le nombre de documents, le nombre moyen de tokens avec l'écart-type et le nombre absolu de tokens, et ces mêmes informations pour les phrases. La liste est triée par nombre de documents.

La distribution des étiquettes est grandement déséquilibrée. L'étiquette *Appel* est la plus représentée, avec 7 019 documents, suivie de *Job* (2 527 documents) et Spam (1 969 documents). Ces trois catégories représentent plus de 60% du corpus. Les catégories comme *Liste* et *Enquete* sont très peu représentées, avec moins de 25 occurrences. La longueur moyenne des documents varie grandement par catégorie, les documents les plus longs appartenant aux catégories *Appel*, *These* et *Ecole* (respectivement 735, 635 et 166 tokens en moyenne). La ressource compte actuellement 9,67 millions de tokens et plus de 2,22 millions de phrases, avec une moyenne globale de 533,93 tokens et 122,39 phrases par document. Les écart-types élevés observés pour toutes les catégories montrent une grande variabilité intra-classe en termes de longueur de document et de structure.

Étiquette	Signification	nb docs	nb moyen de tokens	nb tokens	nb moyen de phrases	nb de phrases
Appel	Appels à communication	7 019	$753{,}88_{\pm 469{,}99}$	5 291 455	$166{,}13_{\pm 93{,}41}$	1 166 070
Job	Offres d'emploi	2 527	$459{,}92_{\pm 337{,}46}$	1 162 221	$107{,}34_{\pm 69{,}63}$	271 238
Spam	Spam	1 969	$80{,}90_{\pm 111{,}70}$	159 288	$20{,}84_{\pm 30{,}58}$	41 234
Conf	Conférences/workshops	1 296	$535{,}95_{\pm 453{,}98}$	694 587	$135{,}61_{\pm 8{,}87}$	175 753
Seminaire	Séminaires	752	$408{,}69_{\pm 229{,}83}$	307 332	$98{,}78_{\pm 48{,}28}$	74 285
Stage	Offres de stage	688	$586{,}22_{\pm 320{,}11}$	403 320	$124{,}73_{\pm 62{,}28}$	85 816
Offre de these	Offres de thèse	611	$587{,}35_{\pm 405{,}14}$	358 871	$132{,}21_{\pm 79{,}65}$	80 780
Info	Type 'autre'	506	$349{,}19_{\pm 473{,}59}$	176 689	$88{,}60_{\pm 72{,}38}$	44 830
Revue	Publications de revue	390	$364{,}26_{\pm 324{,}01}$	142 062	$115{,}65_{\pm 79{,}13}$	45 104
Journee	Journées d'études	384	$440{,}41_{\pm 362{,}32}$	169 118	$111{,}55_{\pm 74{,}65}$	42 837
Ressource	Publication de ressources	324	$355{,}00_{\pm 224{,}05}$	115 020	$97{,}46_{\pm 54{,}98}$	31 578
Ecole	Écoles d'été/d'hiver	313	$625{,}48_{\pm 362{,}90}$	195 774	$139{,}76_{\pm 68{,}62}$	43 745
These	Soutenances de thèse	303	$635{,}00_{\pm 267{,}30}$	192 404	$132{,}16_{\pm 47{,}71}$	40 043
Livre	Parutions de livre	236	$406{,}65_{\pm 320{,}87}$	95 970	$98{,}96_{\pm 58{,}85}$	23 355
Cursus	Promotions de cursus universitaires	186	$496{,}54_{\pm 312{,}37}$	92 357	$119{,}24_{\pm 58{,}53}$	22 178
ATALA	ATALA	89	$502{,}33_{\pm 387{,}97}$	44 707	$117{,}35_{\pm 82{,}61}$	10 444
Question	Questions	74	$166{,}30_{\pm 127{,}48}$	12 306	$50{,}09_{\pm 28{,}42}$	3 707
Habilitation	Défenses de HDR	71	$503{,}23_{\pm 203{,}79}$	35 729	$107{,}70_{\pm 45{,}34}$	7 647
Reponse	Réponses	41	$225{,}17_{\pm 238{,}41}$	9 232	$71{,}37_{\pm 67{,}90}$	2 926
Enquete	Enquêtes / sondages	23	$299{,}91_{\pm 149{,}27}$	6 898	$75{,}96_{\pm 28{,}17}$	1 747
Liste	Annonces concernant LN-ATALA	22	$165{,}36_{\pm 134{,}47}$	3 638	$49{,}50_{\pm 28{,}13}$	1 089
Total	-	17 824	$533{,}93_{\pm 446{,}50}$	9 668 978	$122{,}39_{\pm 91{,}04}$	2 216 406

TABLE 1 – La liste des étiquettes et leur explication, le nombre de documents, le nombre moyen de tokens avec l'écart-type et le nombre absolu de tokens, et ces mêmes informations pour les phrases. La liste est triée par nombre de documents.

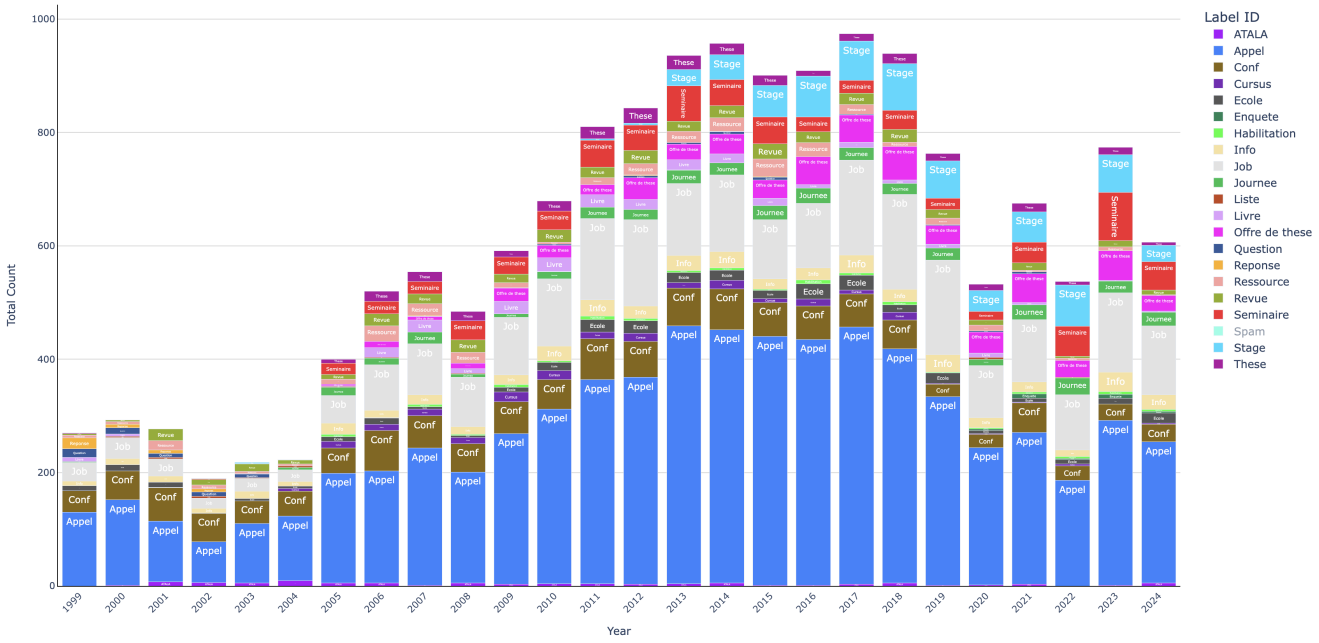

FIGURE 2 – Distribution des étiquettes (hors spam) par année, ordre alphabétique de bas en haut.

4 Exploration terminologue par plongements

Analyse de mots-clés. Dans cette section, nous rapportons une expérience préliminaire sur deux sous-corpus ciblés, les Appels et les Offres (combinaison des labels *Job*, *Offre de these* et *Stage*), visant à identifier des termes distinctifs de périodes pour chaque groupe.

Pré-traitement. Les adresses et les tokens numériques (dates, numéros de téléphone, identifiants...) sont supprimés, car ils tendent à dominer les mesures basées sur la fréquence sans apport sémantique. Nous combinons des listes de *stopwords* de l'anglais et du français pour filtrer les textes. Nous avons enrichi la liste de *stopwords* avec des mots spécifiques très fréquents et peu ou non informatifs, par exemple les trois termes `english`, `version` et `below`. Toutes les étapes décrites ci-après sont appliquées sur les messages pré-traités.

Plongements de documents et de termes. Nous calculons des plongements pour chaque document et chaque terme à l'aide d'un modèle *Sentence-BERT* (SBERT) : `all-MiniLM-L6-v2` [6]. Nous segmentons le corpus en sous-ensembles de périodes (1999-2004 puis des périodes de cinq ans). Pour chaque période p nous calculons un vecteur C_p qui est la moyenne arithmétique des plongements des documents de p : $C_p = \frac{1}{|D_p|} \sum_{d \in D_p} \mathrm{emb}(d)$, où D_p est l'ensemble des documents qui contiennent le terme t et $\mathrm{emb}(d)$ est le plongement du document d.

Pour chaque terme t, nous calculons son plongement comme la moyenne des plongements des documents qui le contiennent : $\mathrm{emb}(t) = \frac{1}{|D_t|} \sum_{d \in D_t} \mathrm{emb}(d)$, où D_t est l'ensemble des documents qui contiennent t. Nous procédons ainsi pour produire un vecteur qui reflète la distribution des contextes dans lesquels le terme apparaît dans le corpus entier.

Score terme × période. Pour mesurer à quel point un terme t est associé à une période p nous utilisons une combinaison de similarité sémantique et de fréquence locale. Soit $\mathrm{df}_{t,p}$ la fréquence de document

6. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

1999-2004	2005-2009	2010-2014	2015-2019	2020-2024
linguistics	paris	paris	analysis	learning
fax	programme	special	computer	université
email	communication	issues	université	machine
abstract	cnrs		science	science
computational	sciences		machine	publication
electronic	questions		learning	analysis
word	word		special	methods
	computer			nlp

TABLE 2 – Top 10 des termes spécifiques par période pour *Appel*

1999-2004	2005-2009	2010-2014	2015-2019	2020-2024
experience	linguistique	linguistique	données	learning
linguistics	connaissance	langue	data	données
areas	langue	linguistique	learning	équipe
computational	candidat	candidat	équipe	python
computer	java	java	python	compétences
knowledge	anglais	connaissance	compétences	développement
university	niveau	documents	développement	apprentissage
fax	informations	stage	stage	intelligence
work	knowledge	niveau	expérience	nlp
position	work	tal		sciences

TABLE 3 – Top 10 des termes spécifiques par période pour les *Offres*

du terme t dans la période p (le nombre de documents de p qui contiennent t), et $\cos(\mathrm{emb}(t), C_p)$ la similarité cosinus entre le plongement du terme t et celui de la période p. Le score terme x période est défini comme $s_{t,p} = \cos\big(\mathrm{emb}(t), C_p\big) \times \log\big(1 + \mathrm{df}_{t,p}\big)$.

Le facteur cosinus multiplicateur capture l'alignement sémantique tandis que le facteur de fréquence logarithmique atténue l'impact des termes très rares en favorisant les termes qui sont à la fois sémantiquement pertinents et suffisamment fréquents pour être significatifs. Le facteur $\log(1 + \mathrm{df})$ est heuristique et fut choisi pour la stabilité : il a une croissance sous-linéaire avec la fréquence, ce qui empêche le score brut d'être dominé par des tokens à haute fréquence mais peu pertinents.

Sélection des termes distinctifs. Les scores sont la base pour une sélection en plusieurs étapes pour trouver les termes spécifiques à une période donnée, tout en écartant les termes courants dans plusieurs périodes. Pour chaque période, les termes sont triés par score en ordre décroissant et les top-K sont conservés. Nous utilisons $K = 60$ ici. Nous supprimons ensuite les termes courants : un terme est défini comme courant s'il apparaît plus de deux fois dans ces listes top-K.

Nous rapportons les termes les plus représentatifs de chaque période, selon notre approche, dans le tableau 2 pour *Appel* et le tableau 3 pour les *Offres* (emploi, thèse, stage).

Appel. Le plus grand nombre de termes distinctifs pour une période selon nos critères est de 8. Cela indique plutôt une stabilité du contenu des appels dans le temps.L'évolution notable réside dans la présence de *linguistics* en première position en 1999-2004 puis l'apparition de *machine* et *learning* en 2015-2019, qui deviennent plus importants en 2020-2024. La période 2010-2014 ne comporte que 3 termes distinctifs (*paris*, *special*, *issues*) après application du filtrage. Cela s'explique par le fait que la quasi-totalité des termes dominants du top-K de cette période le sont aussi pour les autres périodes. Des termes comme *information*, *notification* ou *dates*, bien que très saillants pour 2010-2014, sont également saillants pour les autres périodes et sont donc écartés comme non-distinctifs. En revanche,

les rares termes véritablement propres à 2010-2014, tels que *paris*, *special* et *issues*, révèlent avant tout que cette période ne génère pas de terminologie thématique nouvelle et forte : elle constitue une phase de transition entre la linguistique computationnelle classique et le tournant axé sur les données.

Offres. L'observation la plus frappante concerne l'évolution des profils recherchés. De 1999 à 2014, les offres mettent l'accent sur *linguistics/linguistique* et *langue*, reflétant la prédominance des postes de linguistes. Entre 2015 et 2024, le vocabulaire pivote vers *données*, *data*, *learning*, *python* et *intelligence*, marquant un glissement majeur vers des compétences informatiques et de science des données. Ce changement illustre la transformation du marché du travail vers des rôles plus orientés données et intelligence artificielle.

5 La classification au service de l'exploration des données

Dans cette section nous proposons de continuer l'étude exploratoire du jeu de données à l'aide d'apprentissage automatique, par le biais, cette fois-ci, d'une tâche de classification. Les étiquettes assignées manuellement au long des décennies représentent une des spécificités majeures du jeu de données LN-ATALA. Nous considérons donc ces étiquettes comme classes de références pour une classification multi-classes dans laquelle nous cherchons à prédire l'étiquette à partir du corps du mail. L'objectif de cette tâche n'est pas d'en obtenir les meilleures performances possibles mais plutôt d'identifier les schémas et les tendances de difficulté de prédiction que les modèles peuvent rencontrer afin d'en retirer des intuitions sur les catégories d'e-mails présents dans le corpus. Pour ce faire, nous considérons plusieurs approches, volontairement de familles différentes : (1) une classification habituelle combinant sac de mots en matrice numérique avec un vote majoritaire d'arbres de décisions, (2) l'affinage du modèle d'encodeurs Transformers pour prédire la classe, (3) l'utilisation en inférence de grands modèles de langage auto-régressifs pour la prédiction de l'étiquette par génération de texte.

Classifieurs. Comme indiqué précédemment, nous considérons trois familles de classifieurs. Le *Random Forest* (Breiman, 2001) représente une base de référence bien souvent performante pour la classification par arbre de décision. Il est constitué d'un vote majoritaire de multiples arbres de décision, ce qui le rend souvent rapidement performant sur de petits corpus. Nous utilisons une représentation TF-IDF et 100 arbres de décisions. Bien entendu, nous comparons cette approche avec une autre méthode de classification plus récente : l'utilisation de modèles d'encodeurs *Transformers* (Vaswani *et al.*, 2017) suivant l'architecture RoBERTa (Liu *et al.*, 2019) et entraînés pour le français, dans leurs versions initiales CamemBERT (Martin *et al.*, 2020) (version de base et version large), mais aussi dans une version plus récente inspirée de ModernBERT (Warner *et al.*, 2025) qui intègre toutes les nouvelles architectures issues des travaux sur les modèles génératifs, nommée Modern-CamemBERT (Antoun *et al.*, 2025). Nous affinons ces modèles d'encodeur avec la cross-entropie comme fonction d'erreur, et nous utilisons le jeton `CLS` pour entraîner une tête de classification supplémentaire à la sortie du bloc *transformer*. Pour chaque version, nous procédons à l'entraînement de 300 époques maximum après 500 étapes d'échauffement (*warmup steps*) avec une décroissance de poids à 0.01. Nous utilisons un arrêt anticipé (*early stopping*) en fonction de la F1-mesure et une patience de 5. Ces modèles sur-apprennent vite et nécessitent entre 5 et 15 époques d'entraînement.

Modèles génératifs sans affinage. Nous avons également considéré les modèles génératifs en *zero-shot*, c'est-à-dire sans affinage ni exemple fourni dans l'amorce, et cela, afin de permettre de quantifier leur performance sur cette division stratifiée du jeu de données, laissant alors la possibilité de les utiliser tels quels sur tout l'ensemble des données, sans séparation entre

entraînement-validation-test. Pour cette approche nous nous sommes restreints aux différentes versions du modèle Qwen et, plus précisément, du modèle de 2024, Qwen2.5 (Yang *et al.*, 2024; Team, 2024), ainsi que du tout récent modèle Qwen3.5 sorti fin février 2026 (Qwen Team, 2026). Bien que plusieurs approches existent pour classifier à l'aide de modèles génératifs, nous avons choisi l'approche totalement génératrice dans laquelle le modèle doit générer l'étiquette souhaitée et uniquement l'étiquette souhaitée. Ainsi, au décodage, nous nous limitons uniquement à la conversion en minuscule et à la suppression des accents, puisque ces derniers sont volontairement enlevés de notre ensemble d'étiquettes. Concrètement, nous demandons au modèle de générer un maximum de 30 jetons, lui laissant alors la possibilité d'aller au delà de la simple étiquette. Notre amorce est très simple et se résume à `You are an expert text classifier. Here is the mapping of each label with its corresponding definition: JSON. Classify the user's text into one and only one of the following categories: labellist. Your response must be the label itself and nothing else.` où le `JSON` représente le tableau associatif des étiquettes et de leur signification (Table 1) et `labellist` la liste des étiquettes. Dans le cas où nous ne considérons pas les définitions, nous enlevons la seconde phrase. Pour Qwen2.5, les configurations par défaut de *Transformers* sont utilisées, tandis que pour Qwen3.5 nous n'utilisons que le modèle en modalité textuelle sans raisonnement, avec les recommandations officielles : `temperature=0.7, top_p=0.8, top_k=20, min_p=0.0, presence_penalty=1.5, repetition_penalty=1.0.`

Modèle	Mode	Exactitude	Macro F1	Micro F1	wF1	MCC
Random Forest	train	81,27	53,52	81,27	78,67	0,7622
CamemBERT-base	ft	**86,37**	74,76	**86,37**	**86,39**	**0,8297**
CamemBERT-large	ft	84,46	**76,02**	84,46	84,40	0,8046
ModernCamemBERT	ft	83,12	72,06	83,12	82,86	0,7872
Qwen2.5-7B	zs	40,33	25,70	40,60	40,33	0,3989
Qwen2.5-32B	zs	54,51	47,08	54,53	57,44	0,5345
Qwen2.5-72B	zs	55,19	52,72	55,20	55,68	0,5539
Qwen3.5-4B	zs	48,63	32,80	48,72	50,08	0,4339
Qwen3.5-4B	zs+def	71,00	60,29	71,00	73,72	0,6693
Qwen3.5-9B	zs	47,67	46,27	47,67	43,61	0,4921
Qwen3.5-9B	zs+def	74,43	60,78	74,43	77,51	0,6980
Qwen3.5-27B	zs	65,40	52,75	65,40	67,09	0,6209
Qwen3.5-27B	zs+def	82,84	70,04	82,84	83,24	0,7895

TABLE 4 – Scores de classification des emails par étiquette (Table 1 colonne "étiquette"). 'train' signifie un entraînement complet, tandis que 'ft' (*finetuning*) et 'zs' (*zero-shot*) représentent l'affinage et la prédiction sans affinage ni exemple. Le suffixe 'def' représente l'ajout de définition des classes (Table 1 colonne "signification"). En gras, le meilleur classifieur, en souligné le meilleur générateur.

Évaluation. L'évaluation de ces modèles est faite à l'aide de métriques habituelles de classification multi-classes. Nous utilisons une division stratifiée sur la distribution du corpus afin de permettre une comparaison directe des approches, le tout suivant une répartition de 80% des données pour l'entraînement, et 10% pour les jeux de validation et de test. Nous calculons ainsi les métriques d'exactitude (*accuracy*), de F-mesure pondérée (wF1), les scores de macro- et micro-F-mesure ainsi que la version multi-classes du coefficient de corrélation de Matthews (Matthews, 1975, MCC). Le score de macro-F1 nous renseigne sur la manière dont le modèle gère la plupart des classes, tandis que le MCC nous informe sur la distance des performances par rapport à l'aléatoire, en amoindrissant

l'importance du support par le fait qu'il mesure la corrélation de Pearson (Pearson, 1895) entre les prédictions et les références. Il se définit ainsi MCC $= \frac{TP/N - S \times P}{\sqrt{PS(1-S)(1-P)}}$ avec le nombre de vrais positifs (TP), le nombre total d'échantillons (N), les proportions de vrais positifs (TP/N), de positifs réels (P) et de prédictions positives par le modèle (S). Le MCC peut aller de -1 à +1, 0 représentant l'aléatoire. Ce choix est motivé par le grand déséquilibre des classes, conservé dans la division stratifiée du corpus, et visible dans la figure 2. Dans un objectif exploratoire, cette évaluation ne saurait suffire, c'est pourquoi nous regardons également les erreurs les plus communes propres aux classifieurs les plus performants afin d'obtenir des indices supplémentaires sur les spécificités du jeu de données proposé et l'ambiguïté potentielle de nos étiquettes manuellement assignées.

Analyse des prédictions. Une vue d'ensemble des métriques agrégées de classification et de prédiction d'étiquette est visible en table 5. Ces résultats semblent indiquer une difficulté peu élevée de la tâche considérée compte tenu des très bons résultats d'un Random Forest. Toutefois, ces derniers sont systématiquement dépassés par les encodeurs *transformers* affinés ('ft'), d'autant plus en voyant les résultats très élevés en MCC. Il est également intéressant de constater que les modèles génératifs sans affinage ni exemple ('zs') peinent à obtenir de bons résultats.

L'analyse des différences d'erreurs entre CamemBERT-base et Qwen3.5-27B (zs+def) montre que Qwen semble mieux comprendre l'information contenue mais se perd dans les étiquettes sémantiquement proches comme *Job* qui est confondue avec *Offre de these* et *Stage*. De son côté, CamemBERT a un comportement prédictif plus rigide et assigne à des catégories spécifiques, des catégories plus génériques vues plus souvent lors de l'affinage. Il montre un fort biais de préférence vers la distribution des données. Sans surprise, le premier a des difficultés avec les événements académiques et les stages, tandis que le second peine pour les éléments à part comme les offres spécifiques et les entités propres.

L'apport des définitions est considérable. En comparant Qwen3.5-27B avec et sans définition, l'ajout des définitions montre une amélioration de la prédiction globale avec la correction de certaines erreurs, particulièrement pour les catégories spécifiques. En effet, l'ajout des définitions corrige les mauvaises classifications des étiquettes *Appel*, *Conf* ou encore *Revue*, avec notamment une nette amélioration sur la prédiction des *Appel* qui passe de 62,54 à 90,31 en F1-mesure.

Cette classification n'est qu'une première étape d'exploration des dynamiques des étiquettes. En observant les matrices de confusion, nous voyons que *Job* peut souvent être confondu avec *Offre de these* et *Stage*, alors qu'*Appel* est souvent confondu avec *Conf*. Cela paraît peu surprenant au vu de leur proximité thématique, mais l'inverse n'est pas vrai, posant la question des spécificités de ces classes. Globalement, cette classification informe sur la difficulté de la tâche de classification (inégale selon les classes) et donne des pistes pour des exploitations ultérieures du corpus.

Limitations. Cette approche exploratoire par classification comporte toutefois des limites. La première est (1) la potentielle contamination des données, certes limitée par le fait que nos données soient pré-traitées et corrigées comparé aux sources potentiellement disponibles en ligne (voir Section 3). (2) La tâche se révèle simple pour les modèles sans doute influencés par le fait que les données soient corrigées, il pourrait être intéressant de voir les différences de prédiction entre notre version et les sources brutes. Enfin, (3) nous avons considéré des modèles génératifs *open-weights*, bien que l'utilisation de modèles open-source comme Olmo (Olmo *et al.*, 2025) ou Pythia (Biderman *et al.*, 2023) pourrait nous aider à vérifier les connaissances a priori des modèles.

6 Conclusion

Dans cet article, nous avons présenté le corpus LN-ATALA. Nous avons évoqué la motivation derrière ce travail, la création de la ressource, ses contenus, et décrit quelques expériences qui offrent un aperçu de ce que le corpus peut offrir. Le jeu de données et chaînes de traitement sont rendus publics via git [7]. Il sera enrichi avec le temps, avec l'arrivée continue de nouveaux messages. Cette ressource se démarque des autres corpus d'e-mails par sa facilité d'utilisation, rendue possible par (i) le processus de standardisation par lequel chaque message passe lors de sa diffusion, en plus de (ii) toutes les étapes d'homogénéisation des archives que nous avons opérées. Chaque message est accompagné de métadonnées comme la date d'expédition, la date de diffusion ou l'objet. Nous espérons que cette ressource suscitera l'intérêt de la communauté du traitement automatique du langage.

Remerciements

Nous remercions les relecteurs anonymes pour leurs commentaires qui ont contribué à améliorer la qualité du présent article. Nous remercions également les gestionnaires de la liste qui nous ont précédés, dans l'ordre chronologique Jean Véronis, Pierre Zweigenbaum, Philippe Blache, Alexis Nasr et Thierry Hamon.

Références

ANTOUN W., SAGOT B. & SEDDAH D. (2025). ModernBERT or DeBERTaV3 ? examining architecture and data influence on transformer encoder models performance. In K. INUI, S. SAKTI, H. WANG, D. F. WONG, P. BHATTACHARYYA, B. BANERJEE, A. EKBAL, T. CHAKRABORTY & D. P. SINGH, Éds., *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, p. 3061–3074, Mumbai, India : The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

BETTENBURG N., SHIHAB E. & HASSAN A. E. (2009). An empirical study on the risks of using off-the-shelf techniques for processing mailing list data. In *2009 IEEE International Conference on Software Maintenance*, p. 539–542 : IEEE.

BEVENDORFF J., AL KHATIB K., POTTHAST M. & STEIN B. (2020a). Crawling and preprocessing mailing lists at scale for dialog analysis. In D. JURAFSKY, J. CHAI, N. SCHLUTER & J. TETREAULT, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 1151–1158, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.108.

BEVENDORFF J., AL-KHATIB K., POTTHAST M. & STEIN B. (2020b). Crawling and Preprocessing Mailing Lists at Scale for Dialog Analysis. In *58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, p. 1151–1158 : Association for Computational Linguistics.

BIDERMAN S., SCHOELKOPF H., ANTHONY Q. G., BRADLEY H., O'BRIEN K., HALLAHAN E., KHAN M. A., PUROHIT S., PRASHANTH U. S., RAFF E. *et al.* (2023). Pythia : A suite for

7. Page du projet, avec informations et git public : https://www.atala.org/content/corpus-ln-atala

analyzing large language models across training and scaling. In *International conference on machine learning*, p. 2397–2430 : PMLR.

BIRD C., GOURLEY A., DEVANBU P., GERTZ M. & SWAMINATHAN A. (2006). Mining email social networks. In *Proceedings of the 2006 international workshop on Mining software repositories*, p. 137–143.

BREIMAN L. (2001). Random forests. *Machine learning*, **45**(1), 5–32.

CARVALHO V. R. & COHEN W. W. (2004). Learning to extract signature and reply lines from email. In *Proceedings of the Conference on Email and Anti-Spam*, volume 2004.

ESTIVAL D., GAUSTAD T., PHAM S. B., RADFORD W. & HUTCHINSON B. (2007). Author profiling for english emails. In *Proceedings of the 10th conference of the Pacific Association for computational linguistics*, volume 263, p. 272.

GUENOUNE H., COUSOT K., LAFOURCADE M., MEKAOUI M. & LOPEZ C. (2020). A dataset for anaphora analysis in French emails. In M. OGRODNICZUK, V. NG, Y. GRISHINA & S. PRADHAN, Éds., *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*, p. 165–175, Barcelona, Spain (online) : Association for Computational Linguistics.

HU X., CHOI K., HAO Y., CUNNINGHAM S. J., LEE J. H., LAPLANTE A., BAINBRIDGE D. & DOWNIE J. S. (2017). Exploring the music library association mailing list : A text mining approach. In *ISMIR 2018*, p. 302–308.

JARDIM B., REI R. & ALMEIDA M. S. C. (2021). Multilingual email zoning. In I.-T. SORODOC, M. SUSHIL, E. TAKMAZ & E. AGIRRE, Éds., *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Student Research Workshop*, p. 88–95, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.eacl-srw.13.

KLIMT B. & YANG Y. (2004). The enron corpus : A new dataset for email classification research. In *European conference on machine learning*, p. 217–226 : Springer.

KRIEG-HOLZ U., SCHUSCHNIG C., MATTHIES F., REDLING B. & HAHN U. (2016). CodE alltag : A German-language E-mail corpus. In N. CALZOLARI, K. CHOUKRI, T. DECLERCK, S. GOGGI, M. GROBELNIK, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 2543–2550, Portorož, Slovenia : European Language Resources Association (ELRA).

LAMPERT A., DALE R. & PARIS C. (2009). Segmenting email message text into zones. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, p. 919–928.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized bert pretraining approach. *arXiv preprint arXiv :1907.11692*.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2020). CamemBERT : a tasty French language model. In D. JURAFSKY, J. CHAI, N. SCHLUTER & J. TETREAULT, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7203–7219, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.645.

MATTHEWS B. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, **405**(2), 442–451. DOI : https ://doi.org/10.1016/0005-2795(75)90109-9.

NISHIMURA R., WATANABE Y. & OKADA Y. (2008). Confirmed language resource for answering how type questions developed by using mails posted to a mailing list. In *Proceedings of the 6th Workshop on Asian Language Resources*.

OLMO T., ETTINGER A., BERTSCH A., KUEHL B., GRAHAM D., HEINEMAN D., GROENEVELD D., BRAHMAN F., TIMBERS F., IVISON H., MORRISON J., POZNANSKI J., LO K., SOLDAINI L., JORDAN M., CHEN M., NOUKHOVITCH M., LAMBERT N., WALSH P., DASIGI P., BERRY R., MALIK S., SHAH S., GENG S., ARORA S., GUPTA S., ANDERSON T., XIAO T., MURRAY T., ROMERO T., GRAF V., ASAI A., BHAGIA A., WETTIG A., LIU A., RANGAPUR A., ANASTASIADES C., HUANG C., SCHWENK D., TRIVEDI H., MAGNUSSON I., LOCHNER J., LIU J., MIRANDA L. J. V., SAP M., MORGAN M., SCHMITZ M., GUERQUIN M., WILSON M., HUFF R., BRAS R. L., XIN R., SHAO R., SKJONSBERG S., SHEN S. Z., LI S. S., WILDE T., PYATKIN V., MERRILL W., CHANG Y., GU Y., ZENG Z., SABHARWAL A., ZETTLEMOYER L., KOH P. W., FARHADI A., SMITH N. A. & HAJISHIRZI H. (2025). Olmo 3.

PEARSON K. (1895). Vii. note on regression and inheritance in the case of two parents. *proceedings of the royal society of London*, **58**(347-352), 240–242.

QWEN TEAM (2026). Qwen3.5 : Towards native multimodal agents.

REPKE T. & KRESTEL R. (2018). Bringing back structure to free text email conversations with recurrent neural networks. In *European Conference on Information Retrieval*, p. 114–126 : Springer.

TANIGUCHI M., UEDA Y., TANIGUCHI T. & OHKUMA T. (2020). A large-scale corpus of E-mail conversations with standard and two-level dialogue act annotations. In D. SCOTT, N. BEL & C. ZONG, Éds., *Proceedings of the 28th International Conference on Computational Linguistics*, p. 4969–4980, Barcelona, Spain (Online) : International Committee on Computational Linguistics. DOI : 10.18653/v1/2020.coling-main.436.

TEAM Q. (2024). Qwen2.5 : A party of foundation models.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. U. & POLOSUKHIN I. (2017). Attention is all you need. In I. GUYON, U. V. LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN & R. GARNETT, Éds., *Advances in Neural Information Processing Systems*, volume 30 : Curran Associates, Inc.

WARNER B., CHAFFIN A., CLAVIÉ B., WELLER O., HALLSTRÖM O., TAGHADOUINI S., GALLAGHER A., BISWAS R., LADHAK F., AARSEN T., ADAMS G. T., HOWARD J. & POLI I. (2025). Smarter, better, faster, longer : A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In W. CHE, J. NABENDE, E. SHUTOVA & M. T. PILEHVAR, Éds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 2526–2547, Vienna, Austria : Association for Computational Linguistics. DOI : 10.18653/v1/2025.acl-long.127.

WATANABE Y., NISHIMURA R. & OKADA Y. (2005). Confirmed knowledge acquisition using mails posted to a mailing list. In *Second International Joint Conference on Natural Language Processing : Full Papers*. DOI : 10.1007/11562214_12.

WIESE I. S., DA SILVA J. T., STEINMACHER I., TREUDE C. & GEROSA M. A. (2016). Who is who in the mailing list ? comparing six disambiguation heuristics to identify multiple addresses of a participant. In *2016 IEEE international conference on software maintenance and evolution (ICSME)*, p. 345–355 : IEEE.

YANG A., YANG B., HUI B., ZHENG B., YU B., ZHOU C., LI C., LI C., LIU D., HUANG F., DONG G., WEI H., LIN H., TANG J., WANG J., YANG J., TU J., ZHANG J., MA J., XU J., ZHOU J., BAI J., HE J., LIN J., DANG K., LU K., CHEN K., YANG K., LI M., XUE M., NI N., ZHANG P., WANG P., PENG R., MEN R., GAO R., LIN R., WANG S., BAI S., TAN S., ZHU T., LI T.,

LIU T., GE W., DENG X., ZHOU X., REN X., ZHANG X., WEI X., REN X., FAN Y., YAO Y., ZHANG Y., WAN Y., CHU Y., LIU Y., CUI Z., ZHANG Z. & FAN Z. (2024). Qwen2 technical report. *arXiv preprint arXiv :2407.10671*.

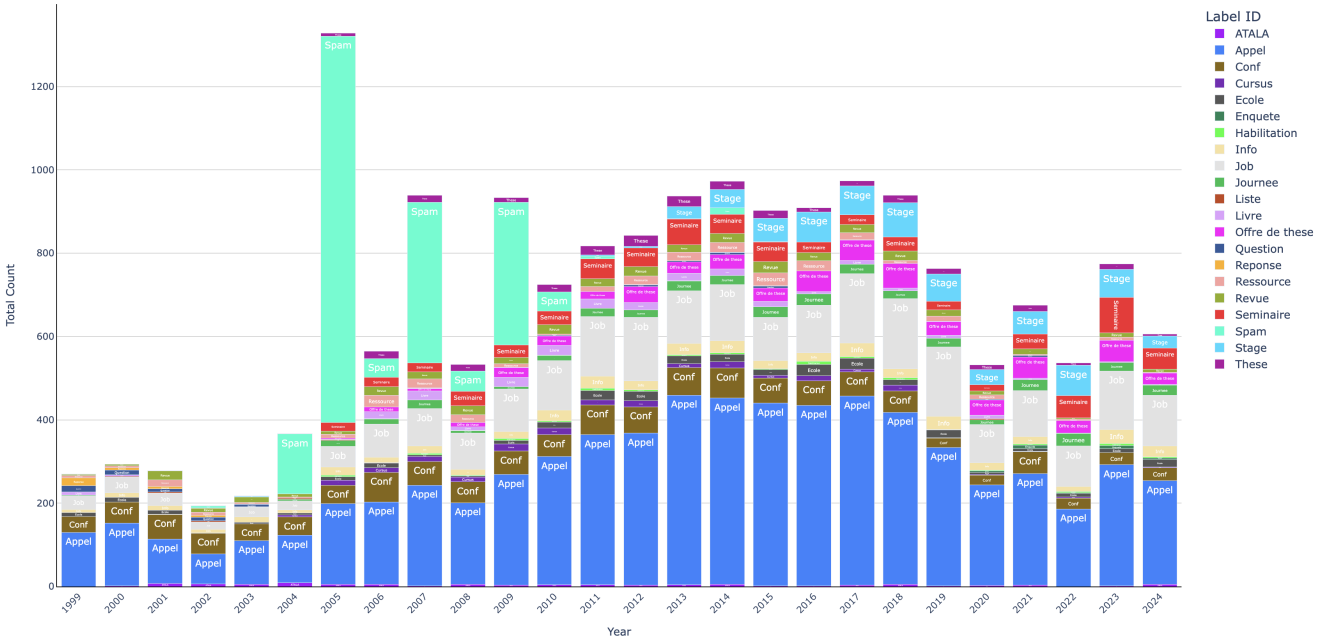

FIGURE 3 – Distribution des étiquettes au fil des années – avec spam

A Annexes

A.1 Considérations éthiques

Nous publions un corpus qui contient des informations identifiantes. Le corpus contient uniquement des informations qui sont destinées à être diffusées publiquement, le plus largement possible. Concernant la maintenance du corpus, nous appliquerons la même politique que celle du site qui héberge actuellement les archives, Renater [8], notamment la possibilité pour toute personne ayant envoyé un e-mail présent dans les archives de demander sa suppression. Nous publions donc une ressource dont le contenu est déjà accessible en ligne, afin d'en faciliter l'exploration pour la recherche. Dans cette optique, nous distribuons le corpus sous licence CC-BY-NC-SA [9].

A.2 Statistiques du corpus

Une visualisation supplémentaire des statistiques du corpus est visible en Figure 3. Elle contient les spams, contrairement à la Figure 2.

A.3 Classification pour l'exploration de données

La table A.3 représente les résultats agrégés de tous les modèles, y compris les modèles de taille intermédiaire.

8. Conditions générales d'utilisation : https://services.renater.fr/groupware/universalistes/cgu
9. https://creativecommons.org/licenses/by-nc-sa/4.0/deed.fr

Modèle	Mode	Exactitude	Macro F1	Micro F1	wF1	MCC
Random Forest	train	81,27	53,52	81,27	78,67	0,7622
CamemBERT-base	ft	**86,37**	74,76	**86,37**	**86,39**	**0,8297**
CamemBERT-large	ft	84,46	**76,02**	84,46	84,40	0,8046
ModernCamemBERT	ft	83,12	72,06	83,12	82,86	0,7872
Qwen2.5-1.5B	zs	19,97	03,86	20,66	10,99	0,1094
Qwen2.5-7B	zs	40,33	25,70	40,60	40,33	0,3989
Qwen2.5-14B	zs	47,56	32,55	47,57	50,35	0,4489
Qwen2.5-32B	zs	54,51	47,08	54,53	57,44	0,5345
Qwen2.5-72B	zs	55,19	52,72	55,20	55,68	0,5539
Qwen3.5-2B	zs	49,47	13,36	49,52	41,83	0,3136
Qwen3.5-4B	zs	48,63	32,80	48,72	50,08	0,4339
Qwen3.5-4B	zs+def	71,00	60,29	71,00	73,72	0,6693
Qwen3.5-9B	zs	47,67	46,27	47,67	43,61	0,4921
Qwen3.5-9B	zs+def	74,43	60,78	74,43	77,51	0,6980
Qwen3.5-27B	zs	65,40	52,75	65,40	67,09	0,6209
Qwen3.5-27B	zs+def	82,84	70,04	82,84	83,24	0,7895

TABLE 5 – Version complète de la table 5 avec toutes les versions de Qwen3.5. Scores de classification des emails par étiquette (Table 1 colonne "étiquette"). 'train' signifie un entraînement complet, tandis que 'ft' (*finetuning*) et 'zs' (*zero-shot*) représentent l'affinage et la prédiction sans affinage ni exemple. Le suffixe 'def' représente l'ajout de définition des classes (Table 1 colonne "signification"). En gras, le meilleur classifieur, en souligné le meilleur générateur.

Les matrices de confusion normalisées pour CamemBERT-base et Qwen3.5 27B en version zero-shot et avec définition sont visibles en Figures 4, 5 et 6.

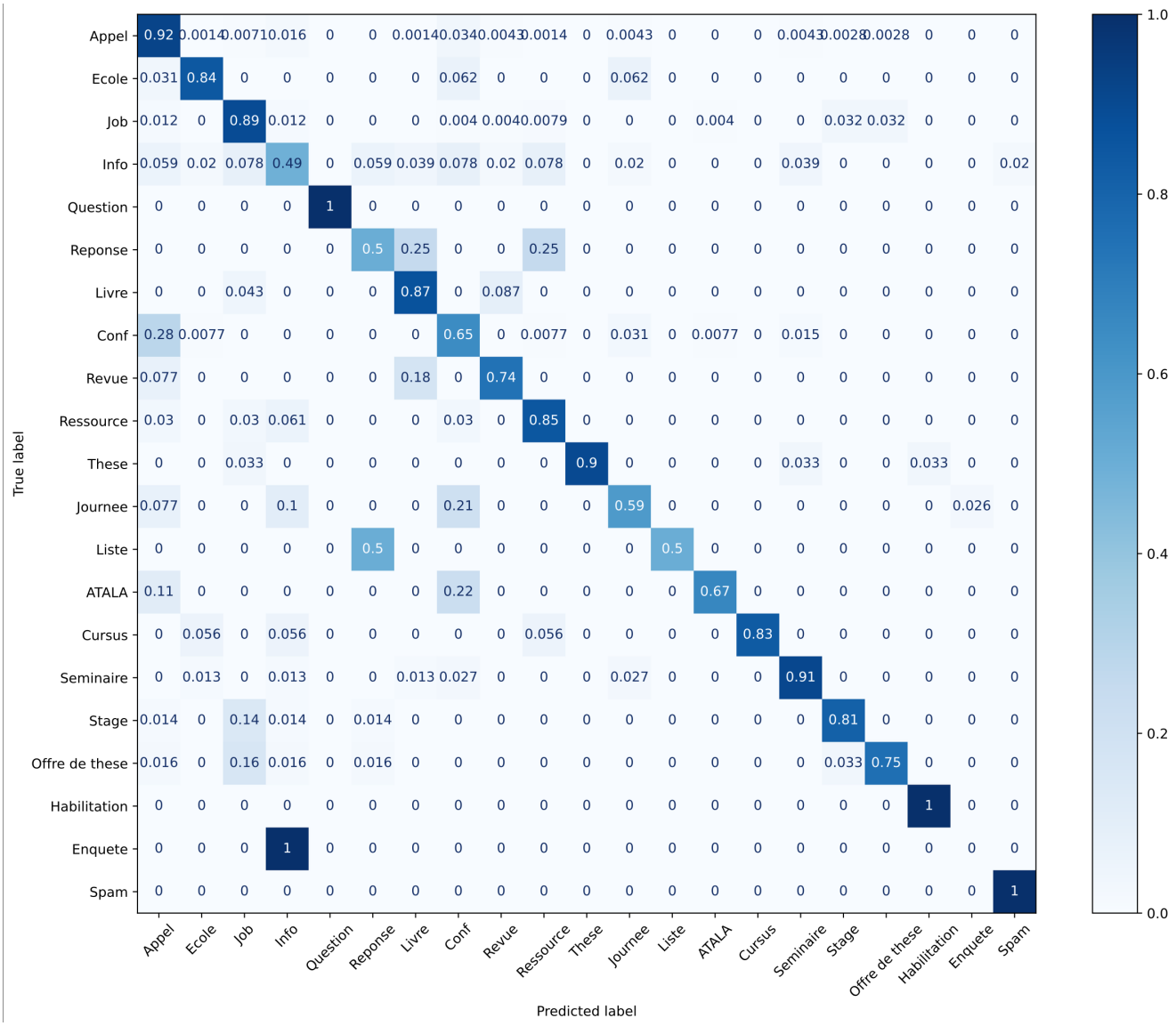

FIGURE 4 – Matrice de confusion normalisée de CamemBERT-base

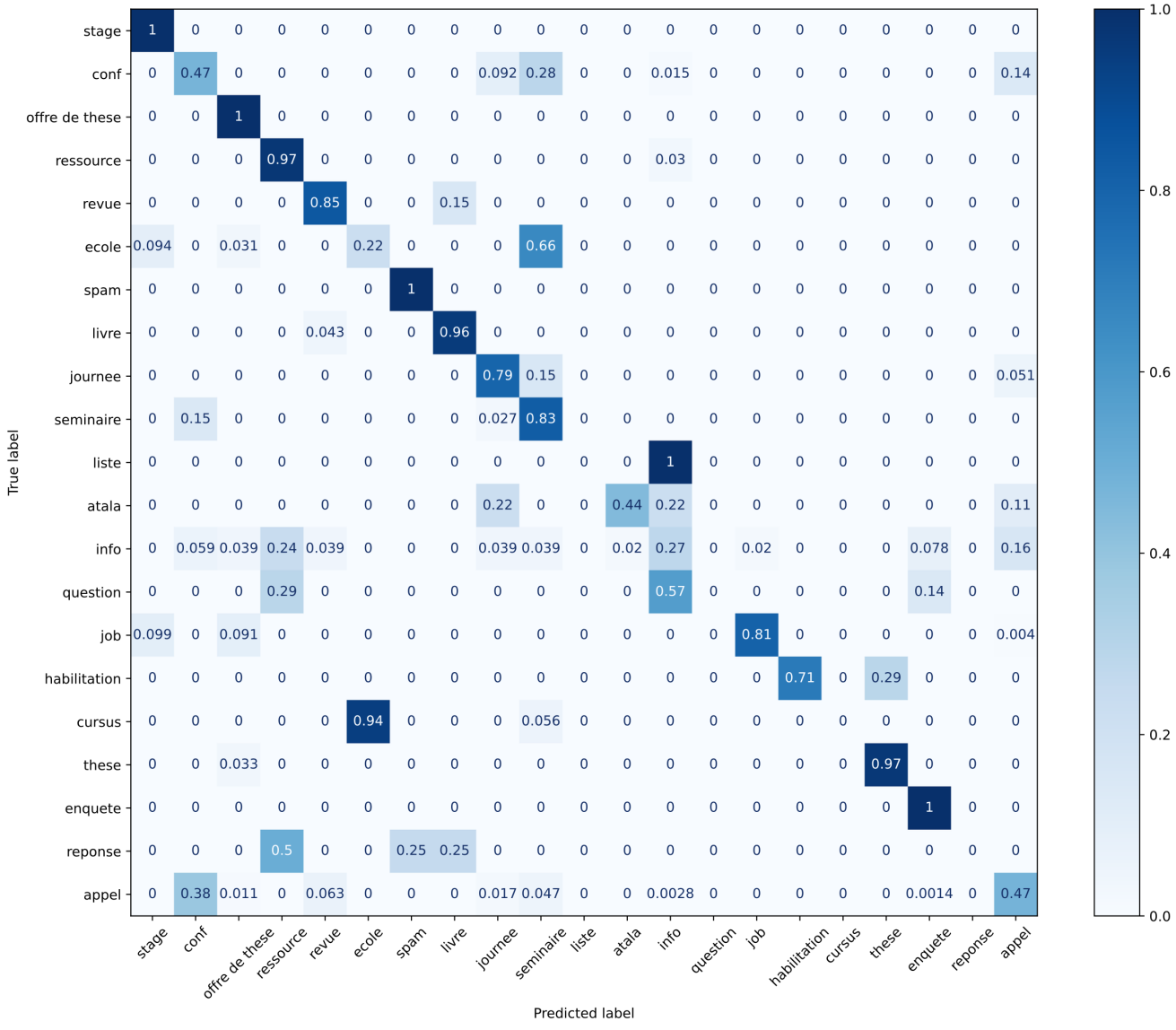

FIGURE 5 – Matrice de confusion normalisée de Qwen3.5-27B (zs)

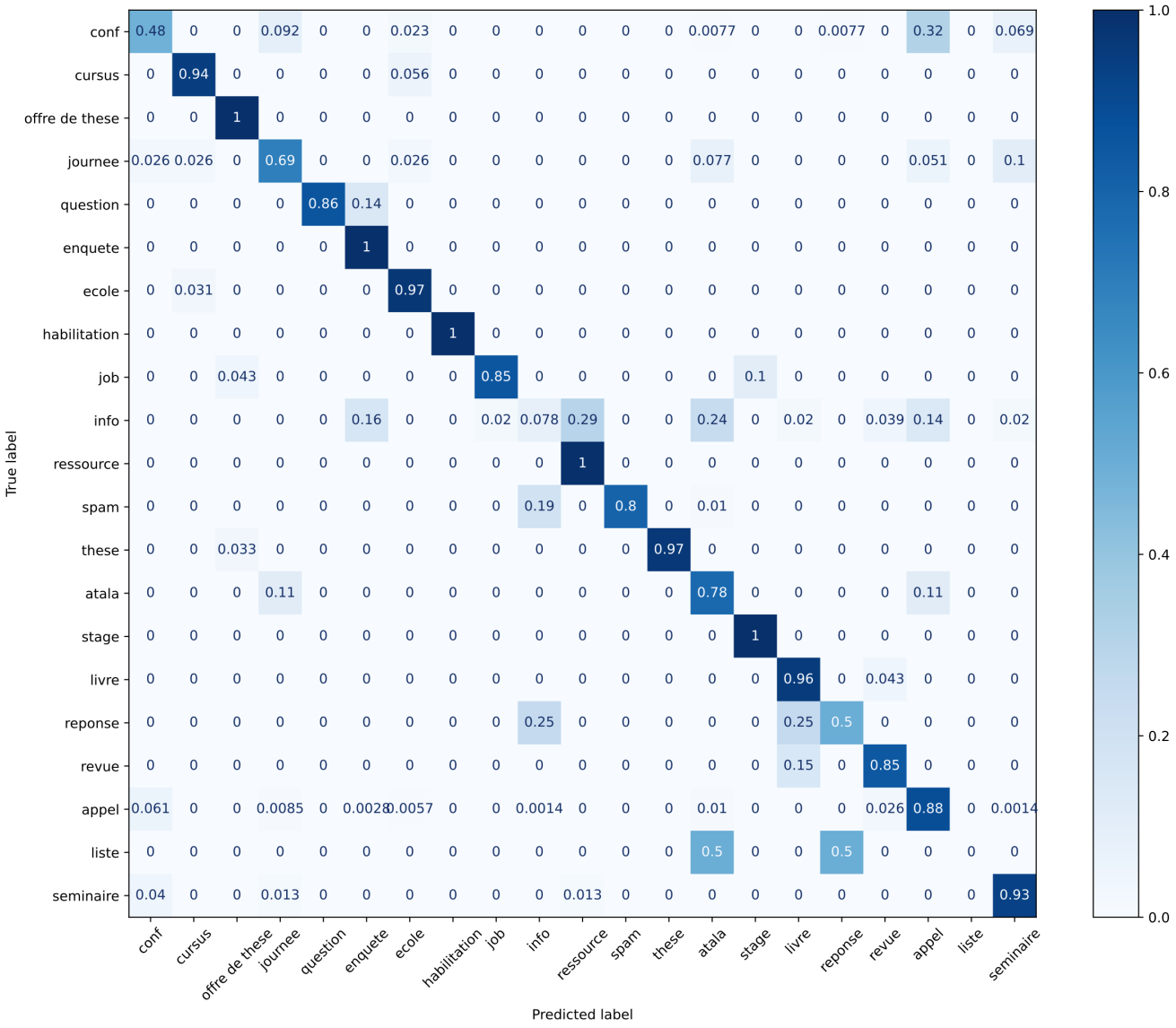

FIGURE 6 – Matrice de confusion normalisée de Qwen3.5-27B (zs+def)