ELITEC : un corpus de conversations en microposts français annoté pour le liage d'entités Wikidata

Vivien Leonard[1, 2] Béatrice Markhoff[3] Jean-Yves Antoine[2,4]
(1) ATOS France
(2) LIFAT, Université de Tours, France
(3) UMR 7324 CITERES, CNRS Université de Tours, France
(4) LIFO, Université d'Orleans, France
{beatrice.markhoff,jean-yves.antoine}@univ-tours.fr
vivien.leonard@etu.univ-tours.fr

RÉSUMÉ

Nous présentons un corpus de microposts en français pour l'évaluation de la tâche de liage des mentions présentes dans le texte à des entités de Wikidata. Ce corpus est annoté à la fois pour la reconnaissance des mentions (Named Entity Recognition - NER) et leur liaison à des entités de Wikidata (Entity Linking - EL). Il s'agit d'une collection de 2 500 microposts, ciblés sur des termes liés à la vie en ville et regroupés en 618 conversations. Construit en suivant les conventions d'annotation de Impresso-Quaero, ce corpus a été pseudo-anonymisé afin d'être mis librement à disposition de la communauté. Nommé ELITEC (EL for mIcroposTs in FrEnCh), son objectif est de compléter les ressources spécifiques au français. ELITEC sert de base de tests pour les tâches NER et EL, et nous l'avons utilisé pour l'évaluation d'un système d'EL que nous avons développé.

ABSTRACT

ELITEC : A Corpus of French Micropost Conversations Annotated for Wikidata Entity Linking

We present a French corpus of micropost conversations, specifically tailored for assessing Wikidata Entity Linking (EL). It undergoes annotation for both Named Entity Recognition (NER) and Entity Linking (EL). It comprises a curated collection of 2,500 microposts, grouped into 618 conversations, selected by terms related to cities. The corpus was built following the Impresso-Quaero methodology, then pseudo-anonymized to make it freely available to the community. By introducing this corpus for the evaluation of EL for microposts in French (ELITEC), our aim is to help increase quantity and quality of language-specific resources for this language. ELITEC serves as a benchmark for NER and NEL, and have been used to evaluate an EL system that we developed.

MOTS-CLÉS : conversations de microposts, corpus annoté, français, liage d'entité, Wikidata.

KEYWORDS: Micropost Conversation, Annoted Corpus, French, Entity Linking, Wikidata.

1 Introduction

Les réseaux sociaux permettant la production de microposts représentent des sources de données volumineuses utilisées dans un large éventail de tâches d'apprentissage automatique (ML). En effet, la plupart des tâches de ML trouvent des applications pratiques sur ces plateformes, notamment les tâches de traitement du langage naturel (TAL) telles que la reconnaissance d'entités nommées

(NER), la résolution des coréférences (CR) ou l'analyse des sentiments (SA). Toutes nécessitent des ressources d'évaluation pour estimer les performances des algorithmes développés. Pour l'anglais, plusieurs corpus bien établis existent, tels que le Broad Twitter Corpus (BTC) (Derczynski *et al.*, 2016) et le Edinburgh Twitter Corpus (Petrović *et al.*, 2010). Ces ensembles de données ont été construits selon une méthodologie rigoureuse et sont largement reconnus au sein de la communauté TAL, facilitant ainsi l'évaluation des systèmes.

Pour la langue française, et en se concentrant sur la tâche spécifique d'Entity Linking (EL) (i.e. l'identification et l'extraction des entités nommées (NEs) dans un texte et la liaison aux entités correspondantes au sein d'une base de connaissances (KB)), on observe un manque de ressources facilement disponibles. À notre connaissance, un seul corpus existant a tenté de combler cette lacune : un sous-ensemble du corpus Microposts (Rizzo *et al.*, 2015). Cependant, en considérant les principes FAIR (Findable, Accessible, Interoperable, Reusable), il présente des lacunes en termes d'accessibilité et de réutilisabilité, car ses microposts sources ne sont plus disponibles. Pour la tâche de NER, à notre connaissance, une seule ressource satisfisant aux besoins existe : CAP 2017 (Lopez *et al.*, 2017). Elle sert de base de comparaison pour évaluer la qualité du corpus ELITEC présenté dans cet article.

ELITEC est un corpus français de microposts conçu pour l'évaluation de l'EL. Ce corpus est construit à partir de conversations liées à la vie en ville entre des utilisateurs de plateforme de type micropost et permet d'évaluer les algorithmes d'EL à la fois au niveau de la conversation et du post. ELITEC présente une concentration plus élevée d'entités par rapport à CAP 2017 et s'appuie sur un guide et un processus d'annotation rigoureux. La création de ce corpus a été initiée dans le cadre du projet *France 2030 Cloud Platform for Smart Cities* (*CP4SC*), dont l'objectif est de collecter et analyser des données liées aux environnements urbains à partir de différentes sources (par exemple des réseaux sociaux ou des réseaux de capteurs). ELITEC a été utilisé pour la mise au point d'un algorithme original d'EL pour microposts (Leonard *et al.*, 2024).

La section 2 détaille le processus de collecte des données. La section 3 se concentre sur les lignes directrices d'annotation qui ont été suivies. La section 4 décrit le processus d'annotation, sur la base de ces lignes directrices. La section 5 présente le contenu du corpus librement distribué [1] et le compare au corpus *CAP 2017*. La fiabilité de l'annotation est discutée dans la section 6, ainsi que les caractéristiques du corpus.

2 Collecte des données

Les données ont été collectées à partir des API Twitter (désormais X) et Mastodon en utilisant des noms de villes françaises comme mots-clés. Les mots-clés utilisés sont "Toulouse", "Paris", "Bordeaux" et "Nice", noms de villes suffisamment grandes pour nous permettre de recueillir une quantité significative de données. 2 500 tweets ont été extraits, répartis en 668 conversations. Nous avons exclu les conversations constituées d'un seul micropost, car notre intérêt porte sur les échanges comprenant au moins deux microposts. Cette décision repose sur la richesse des informations contextuelles contenues dans des échanges de microposts plus longs. L'objectif de ce corpus est d'évaluer de manière cohérente les algorithmes au niveau conversationnel, ce qui nécessite la conservation des conversations offrant suffisamment d'informations contextuelles. En ce qui concerne X, les données ont été collectées avant les modifications des politiques d'accès à l'API et en utilisant un accès universitaire officiel. Pour les deux API, nous avons opté pour une méthode d'interrogation

1. https://repository.ortolang.fr/api/content/elitec/head/

simple visant à capturer des données se rapprochant le plus possible de l'utilisation des mots-clés. Nous avons utilisé une structure de requête reposant uniquement sur le mot-clé, tout en intégrant des critères d'exclusion pour filtrer le contenu multimédia ainsi que les cas où le mot-clé apparaît dans des hashtags (i.e. #MOT) ou des mentions d'utilisateurs (i.e. @NomUtilisateur). Nous soutenons que de telles occurrences apportent plus d'ambiguïté que d'information, en raison de la nature incohérente de la construction et de l'utilisation des noms d'utilisateur et des hashtags. Le format de requête utilisé correspond au modèle suivant : *"KEYWORD lang :fr place_country :FR -has :media -is :nullcast"*. Les données issues des réseaux sociaux proviennent dans notre cas de deux sources principales : Twitter et Mastodon. Une première partie du corpus a été collectée à partir de Twitter, via un accès universitaire accordé dans le cadre du programme *Twitter API for Academic Research*. Cet accès permettait, jusqu'en 2023, de consulter de larges volumes de données à des fins de recherche, conformément à une procédure de demande formelle. Il convient de souligner que ce programme n'est désormais plus accessible, la politique d'accès aux données de Twitter ayant été profondément modifiée à la suite de changements structurels opérés sur la plateforme.

La seconde partie des données a été extraite de l'instance `piaille.fr`, un serveur Mastodon public hébergé en France [2]. Cette instance repose sur le logiciel libre Mastodon, distribué sous licence AGPLv3. Conformément à sa politique de confidentialité [3], les contenus publiés en mode public peuvent être consultés librement via l'API, dans le respect du cadre fixé par l'instance et des droits des utilisateurs. Seuls les statuts explicitement publics ont été collectés, à l'exclusion de toute information privée. L'ensemble de la collecte a été réalisé dans une démarche éthique et conforme aux bonnes pratiques en vigueur dans le domaine de la recherche en traitement automatique des données sociales.

Les données ont été pseudonymisées conformément aux recommandations CNIL [4]. Pour Mastodon, seuls les messages publics ont été extraits, les auteurs ont été supprimés, et les textes sont modifiés afin d'éviter toute réidentification. Cette approche est compatible avec la diffusion sous licence CC BY.

3 Méthodologie d'annotation

Le processus d'annotation est divisé en deux parties : l'annotation NER et l'annotation NEL.Pour l'annotation NER, nous avons suivi les lignes directrices fournies par **Impresso** (Ehrmann *et al.*, 2020), qui est un sous-ensemble du projet plus large **Quaero** (Rosset *et al.*, 2011). Ces deux initiatives ont introduit une stratégie d'annotation structurée complète s'appuyant sur un cadre hiérarchique englobant différents types et composants.

Pour illustrer le schéma d'annotation utilisé dans **Impresso**, considérons l'exemple de l'annotation de "le roi Mohamed VI." Dans ce cas, le processus d'annotation divise le texte en composants distincts, chacun étant associé à une paire type/sous-type spécifique ainsi qu'à un composant. Le terme "roi" est annoté avec la balise "<func.ind>", indiquant son rôle fonctionnel. "Mohamed VI" est annoté avec "<pers.ind>", et chacune de ses parties constitutives est également annotée : "Mohamed" est annoté avec "<name.first>", et "VI" est annoté avec "<qualifier>", soulignant ainsi sa fonction de composant qualifiant au sein de l'entité.

2. https://piaille.fr
3. https://piaille.fr/privacy-policy
4. https://www.cnil.fr/fr/reglement-europeen-protection-donnees/chapitre1#Article4

Dans les données de type micropost, les entités nommées (NE) sont relativement rares et il existe peu d'opportunités pour mettre en œuvre une structure hiérarchique aussi détaillée. En utilisant le contexte disponible dans ces données, nous ne disposons pas d'assez d'informations pour désambiguïser les niveaux d'annotation inférieurs (e.g. org.ent) par rapport aux catégories générales (e.g. loc, org, pers) de l'organisation hiérarchique définie dans **Impresso**. En nous basant sur **Impresso**, nous avons annoté une entité avec un type principal et avons également annoté chacun de ses composants. Par exemple, "Emmanuel Macron" est annoté avec le type d'entité *pers*, "Emmanuel" étant désigné comme le composant *firstname* et "Macron" comme le composant *lastname*. Cette décision a conduit à la création de nouvelles lignes directrices d'annotation [5].

De plus, notre approche diffère significativement des autres corpus de micropost existants pour le NER sur un point fondamental : nous avons délibérément choisi de ne pas annoter les noms d'utilisateur (@USER) et les hashtags (#HASHTAG). Cette décision repose sur l'observation que les microposts présentent généralement une utilisation incohérente de ces deux attributs, les rendant ainsi peu fiables comme indicateurs d'entités référencées. Par conséquent, ils ne sont pas considérés comme pertinents pour les tâches de NER. Bien que les noms d'utilisateur et les hashtags puissent occasionnellement contenir du contenu informatif (e.g. @EmmanuelMacron, @elonmusk, #Covid19), la majorité des occurrences de ces éléments tend à manquer d'information substantielle, présenter des formes instables ou être dépourvues de pertinence contextuelle (e.g. CMLyon, OMso).

Concernant la partie NEL, nous avons choisi la base de connaissances Wikidata [6] pour ses qualités de bonne couverture, en particulier des entités d'actualité (les microposts sont essentiellement à propos de l'actualité), son caractère disponible et ouvert et ses mises à jour très régulières grâce à une communauté de contributeurs importante. Il est à noter que l'annotation NEL n'a pas représenté un défi majeur, car les entités nommées réellement ambiguës étaient relativement peu nombreuses.

4 Processus d'annotation

Le processus d'annotation s'est déroulé en deux phases distinctes, correspondant aux deux sources de données : Twitter et Mastodon.

La première phase a porté sur les données issues de Twitter. Elle a été réalisée par deux annotateurs : une annotatrice expérimentée, ayant déjà participé à une campagne d'annotation similaire avec les mêmes lignes directrices, et un annotateur débutant, spécifiquement formé par cette dernière. Un corpus pilote a d'abord été annoté afin d'identifier d'éventuels problèmes logiciels ou ambiguïtés dans les catégories, et de garantir une compréhension partagée des lignes directrices. À l'issue de chaque sous-ensemble annoté, une phase de curation était menée. En cas de désaccord persistant entre les annotateurs, un médiateur expert extérieur au processus d'annotation intervenait pour décider. Cette personne jouait un rôle essentiel d'arbitrage et de validation finale.

La seconde phase d'annotation, réalisée ultérieurement, visait à intégrer les données extraites de l'instance Mastodon `piaille.fr`. Elle a mobilisé deux nouveaux annotateurs débutants, encadrés par un groupe de trois personnes, incluant le médiateur de la première phase. Grâce à sa connaissance fine des lignes directrices, ce dernier assurait la cohérence des décisions, tranchait les divergences, et garantissait l'alignement méthodologique entre les deux phases.

5. https://repository.ortolang.fr/api/content/elitec/head/
6. https://www.wikidata.org/wiki/Wikidata:Main_Page

L'ensemble du jeu de données a été divisé en quatre sous-ensembles, sans séparation des conversations au sein d'un même bloc, afin de préserver l'intégrité contextuelle des dialogues. Après annotation de chaque sous-ensemble, une étape de curation était systématiquement menée, comme le présente la figure 1. Cette étape assurait un contrôle qualité rigoureux et permettait, en cas de désaccord entre les annotateurs, l'intervention d'un médiateur final pour harmoniser les annotations.

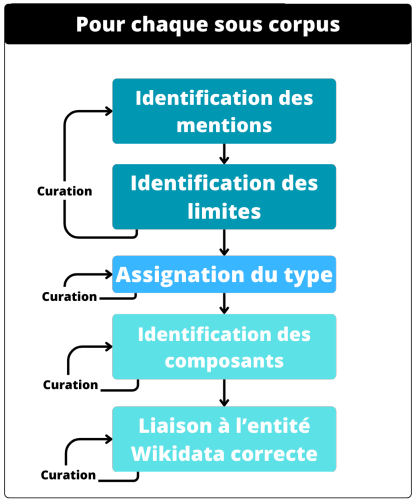

FIGURE 1 – Processus d'annotation des données

Le logiciel *INCEpTION*[7] a été utilisé pour l'annotation et l'export des données au format *xmi*. Ce logiciel permet l'annotation NER et NEL par l'ajout de couches sémantiques sur le texte brut.

5 Description du corpus

Le jeu de données a été annoté en utilisant six types : **Pers** pour Personne, **Loc** pour Lieu, **Event** pour Événement, **Prod** pour Production Humaine, **Nat** pour Naturel et **Org** pour Organisation. Ces types suivent un standard universel couramment utilisé dans divers corpus de reconnaissance d'entités nommées (NER) et se retrouvent dans les lignes directrices d'**Impresso**[8], à l'exception de la catégorie «Naturel». Cette dernière catégorie est destinée aux événements ou objets naturels. Par exemple, les mentions de *Covid* sont classées comme Naturel. Le type **Nat**, introduit pour isoler les entités du monde naturel non façonnées par l'humain et non bornées temporellement, permet de distinguer clairement ces entités des productions humaines (**Prod**), des lieux géographiques localisables (**Loc**) ou des événements ponctuels (**Event**), facilitant ainsi la compatibilité avec d'autres jeux de données et la projection vers des schémas d'annotation existants.

Concernant l'annotation NER, le seul corpus avec lequel nous pouvons comparer notre approche est *CAP 2017*. Ce corpus est composé de 6 685 tweets. Ses auteurs ont extrait tous les tweets d'une fenêtre temporelle, sans filtre, sans ciblage particulier et sans chercher à former des conversations.

7. https://inception-project.github.io/
8. https://zenodo.org/records/3604227

Pour l'annotation NER, ils emploient une plus grande variété de types que nous : *Musicartist*, *Product*, *Person*, *Geoloc*, *Movie*, *Org*, *Other*, *Media*, *Sportsteam*, *Facility*, *Tvshow*, *Event* et enfin *Transportline*. Afin de réaliser une comparaison significative entre notre corpus et le corpus *CAP 2017*, une étape essentielle consiste à harmoniser les types de données utilisés dans les deux corpus. Cette harmonisation implique de convertir ou de reformater les types du corpus CAP2017 pour qu'ils respectent nos conventions de typage établies.

En alignant les types de données, nous nous assurons que les deux corpus partagent un cadre commun, permettant ainsi une analyse comparative équitable et fiable. Voici comment nous agrégeons les types de CAP 2017 pour correspondre aux nôtres :

- **Pers** : *Musicartist*, *Person*
- **Loc** : *Geoloc*
- **Event** : *Event*
- **Prod** : *Product*, *Movie*, *Tvshow*
- **Nat** : Aucun type correspondant n'a pu être trouvé dans le corpus *CAP 2017*
- **Org** : *Org*, *Media*, *Sportsteam*, *Facility*, *Transportline*

	ELITEC	CAP 2017
Nb de tokens	58 180	80 750
Nb d'entités	3 258	3 753
Fréquence des NEs	5,6%	4,64%

TABLE 1 – ELITEC et CAP 2017 fréquence des entités

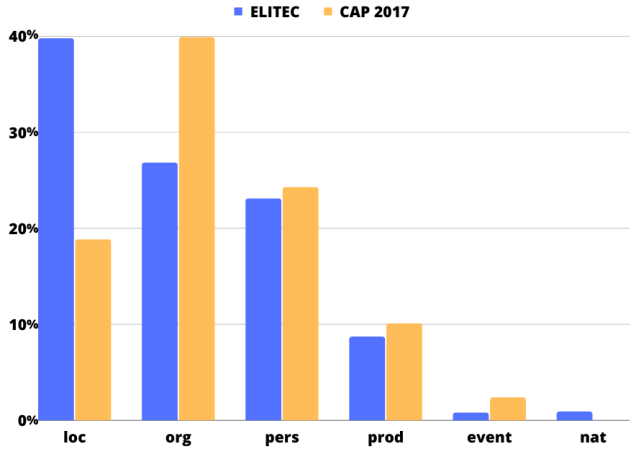

FIGURE 2 – Distribution des classes pour ELITEC et CAP 2017

En analysant la distribution des types d'entités dans les deux corpus, illustrée dans le tableau 1, nous observons une fréquence d'apparition plus grande des entités dans notre corpus, combinée à une quantité plus faible de microposts. Ce phénomène peut être attribué à notre méthode spécifique de sélection de données sur des noms de villes. En adoptant cette stratégie, nous garantissons que notre jeu de données contient, au minimum, l'entité nommée utilisée pour récupérer les microposts, en raison des moteurs de recherche stricts de Twitter et Mastodon (i.e. lorsqu'une recherche est effectuée avec un mot-clé, ils récupèrent chaque micropost contenant ce mot-clé).

Les deux corpus présentent une distribution des classes comparable, comme le montre la figure 2. La variation observée pour les types *loc* et *org* peut être attribuée aux différences dans les méthodes de collecte des données. Puisque nous avons collecté des données en utilisant des entités de type lieu, il est logique d'observer une plus grande prévalence de ces entités au sein du corpus ELITEC.

6 Fiabilité des données et expérimentation

6.1 Accord

Lors du processus d'annotation, nous avons considéré chaque sous-ensemble de données comme un jeu de données complet. Pour chaque sous-ensemble, et après le processus de curation, nous avons calculé l'alpha de Krippendorff (Krippendorff, 2018) afin d'évaluer l'accord inter-annotateurs pour l'annotation des types, des composants et des entités liées. Les scores sont présentés dans la figure 3. L'annotation réalisée, tant pour l'annotation des types d'entités que pour le liage, peut être considérée comme fiable : le taux d'accord inter-annotateurs observé est toujours supérieur au seuil de fiabilité de 0.8 défini par (Krippendorff, 2004). La partie la plus difficile était l'identification des composants. Cette difficulté découle du fait que cette tâche nécessite de répondre à la question : quelles sont les limites d'une entité ? Cela peut être simple, par exemple «Emmanuel Macron est un homme », où l'entité nommée se limite à «Emmanuel Macron ». Mais cela peut également être complexe dans d'autres cas, par exemple «La Fnac de Nice, près de la place principale ». Dans cet exemple, est-ce que la partie «près de la place principale »doit être prise en compte dans l'annotation de l'entité ? Contextuellement, il est possible qu'il n'y ait pas de nécessité d'inclure cette partie là afin de caractériser l'entité que l'on cherche à annoter. Mais il est également possible que cela soit nécessaire afin de lever toute ambiguïté. Ce choix est complexe, il dépend d'une connaissance précise de ces entités et est, comme dit précédemment, hautement contextuel. Aussi, par souci de clarté, nous estimons qu'il est nécessaire d'annoter l'ensemble de l'entité : dans l'exemple précédent toute l'expression permettant de caractériser précisément l'entité est sélectionnée.

À noter qu'à ce jour, il n'existe pas de mesure d'accord inter-annotateurs spécifiquement conçue pour évaluer de manière satisfaisante la fiabilité de l'annotation de la segmentation textuelle en unités — qu'il s'agisse d'entités nommées ou d'autres types d'unités. Les travaux de Krippendorff et al. (Krippendorff *et al.*, 2016) constituent une contribution notable sur ce sujet, notamment à travers la métrique γ, qui permet d'estimer de manière relative la fiabilité entre deux annotations. Toutefois, comme le souligne Mathet (Mathet, 2017), cette métrique ne permet pas d'évaluer de façon absolue la qualité intrinsèque d'une annotation isolée. Nous n'avons donc pas calculé de mesure d'accord inter-annotateurs pour cet aspect-là.

Notons que des gains d'efficacité ont été observés au fur et à mesure de l'avancement de l'annotation à travers les quatre sous-ensembles : les deux annotateurs ont construit une connaissance partagée qui s'est enrichie tout au long du processus d'annotation, réduisant ainsi le temps nécessaire à l'annotation. Il est important de noter que la grande majorité des erreurs d'ambiguïté provient de la difficulté à attribuer le bon type entre Lieu et Organisation. En particulier dans notre corpus, lorsque l'on parle de sport et de villes, il est fréquent d'utiliser le nom d'une ville pour faire référence à une de ses équipes sportives. Cela peut entraîner une forte ambiguïté, car il est nécessaire, à partir du contexte, de comprendre ce qui est réellement référencé derrière la mention.

	Sous corpus 1	Sous corpus 2	Sous corpus 3	Sous corpus 4	Moyenne
Types	0,80	0,82	0,81	0,83	0,815
Composants	0,70	0,68	0,67	0,72	0,692
Entités liées	0,82	0,80	0,81	0,89	0,83

FIGURE 3 – Mesure de l'alpha de Krippendorff

Model	Recall	Precision	F1-score
bertweetfr_ner	0,664	**0,783**	**0,719**
roberta-large-NER	**0,681**	0,761	0,718
Davlan_bert-base-multilingual-cased-ner-hrl	0,613	0,746	0,672
magbert	0,634	0,679	0,656
jb-camembert-ner	0,623	0,644	0,634
distilcamembert-base-ner	0,598	0,607	0,603
wikineural-multilingual-ner	0,525	0,570	0,546
jb-camembert-ner-with-dates	0,537	0,521	0,529
distilbertmultilingual	0,477	0,580	0,524
camembert-base-finetuned-ner	0,433	0,398	0,415
julian-schelb_roberta-ner-multilingual	0,388	0,377	0,383

TABLE 2 – Résultat de l'évaluation des modèles NER

6.2 Evaluation NER

Afin d'évaluer la facilité d'utilisation d'*ELITEC*, nous avons choisi de mener une première expérience limitée dans laquelle nous avons évalué 11 modèles de reconnaissance d'entités nommées (NER) pré-entraînés disponibles sur la plateforme Hugging Face [9]. L'évaluation repose sur les métriques usuelles de rappel, de précision et de F1-score. Aucun modèle n'a été réentrainé avant d'être évaluer. Les performances obtenues sont synthétisées dans le tableau 2, tandis que le tableau 3 présente pour chaque modèle les types d'entités reconnus ainsi que les jeux de données d'entraînement utilisés.

Afin d'assurer une comparaison cohérente avec les annotations de notre corpus, une procédure de mise en correspondance a été définie entre les types d'entités reconnus par les modèles et les types d'ELITEC. Lorsque les libellés des types étaient identiques ou directement équivalents (par exemple `PER`, `ORG`, `LOC`), le mapping était direct. Pour certains modèles proposant des schémas plus riches (comme `bertweetfr-ner` ou `magbert`), seuls les types présentant une correspondance explicite avec ceux d'ELITEC ont été pris en compte. Les entités sans équivalent dans notre schéma typologique ont été ignorées lors de l'évaluation. Les règles de correspondances mises en place pour

9. https://huggingface.co/models

les cas non évidents sont présentées dans le tableau 4.

Cette première expérimentation n'a rencontré aucune difficulté de mise en œuvre, confirmant ainsi la bonne intégration technique et l'utilisabilité du corpus ELITEC dans un protocole d'évaluation standardisé, au bénéfice de la communauté de recherche en traitement automatique du langage.

Modèle	**Types**	**Corpus d'entraînement**
bertweetfr_ner	Person, MusicArtist, Org, GeoLoc, Product, TransportLine, Media, SportsTeam, Event, TVShow, Movie, Facility, Other	CAP2017
roberta-large-NER	PER, ORG, LOC, MISC	CoNLL-2003
Davlan_bert-base-multilingual-cased-ner-hrl	PER, ORG, LOC	Europeana Newspapers
magbert	PER, DATE, GPE, NORP, LOC, ORG, FAC, MONEY, TIME, PERCENT, EVENT, LAW	Corpus non publique
jb-camembert-ner	PER, ORG, LOC, MISC	WikiNER-fr
distilcamembert-base-ner	PER, ORG, LOC, MISC	WikiNER-fr
wikineural-multilingual-ner	PER, ORG, LOC	WikiNEuRal
jb-camembert-ner-with-dates	PER, ORG, LOC, DATE, MISC	Version enrichie de WikiNER-fr pour ajouter le type "DATE"
distilbertmultilingual	PER, ORG, LOC, MISC	Non spécifié
camembert-base-finetuned-ner	PER, ORG, LOC, DATE, Adresse postale	WikiNER-fr enrichi avec de nouveaux types
julian-schelb_roberta-ner-multilingual	PER, ORG, LOC	WikiANN

TABLE 3 – Types d'entités reconnues et jeux de données d'entraînement pour les modèles NER évalués

Modèle	Type reconnu	Correspondance vers ELITEC	Utilisé
bertweetfr-ner	MusicArtist	PER	Oui
	GeoLoc	LOC	Oui
	TransportLine, Media, Sportsteam, TVShow, Facility	ORG	Oui
	Product, Event, Movie, Other	Non mappés	Non
magbert	GPE	LOC	Oui
	NORP, FAC	ORG	Oui
	MONEY, TIME, PERCENT, EVENT, LAW, DATE	Non mappés	Non
jb-camembert-ner-with-dates	DATE	Non mappé (absent d'ELITEC)	Non
camembert-base-finetuned-ner	Adresse postale	Mappé sur LOC (approximation)	Oui
	DATE	Non mappé	Non

TABLE 4 – Types non standards reconnus par certains modèles et règles de mapping appliquées vis-à-vis du schéma typologique ELITEC

7 Conclusion

Dans cet article, nous avons présenté ELITEC, un corpus de microposts en français contenant des entités nommées avec des annotations NER et NEL, conçu selon une méthode décrite dans un guide accessible compatible avec Quaero. Il a été utilisé pour une tâche d'EL décrite dans (Leonard *et al.*, 2024). Il s'agit du seul corpus conçu pour l'étude des tâches de NER et de NEL sur des données de type microposts en français qui soit actuellement librement accessible [10]. Il a été pseudo-anonymisé pour rendre cet accès libre possible. Ce corpus est spécifiquement axé sur les discussions liées à la vie urbaine et a été construit dans ce but. Bien qu'il ne s'agisse pas d'un corpus générique représentant tous les types de communication par microposts, il constitue néanmoins une ressource précieuse pour la recherche dans ce domaine. Il sera hébergé dans un référentiel public afin d'assurer sa trouvabilité, avec des métadonnées garantissant son accessibilité, exprimées à l'aide de vocabulaires standards pour assurer son interopérabilité. De plus, son processus de construction et d'annotation est clairement documenté afin d'en accroître la réutilisabilité, en accord avec les principes FAIR.

Références

DERCZYNSKI L., BONTCHEVA K. & ROBERTS I. (2016). Broad twitter corpus : A diverse named entity recognition resource. In *Proceedings of COLING 2016, the 26th International Conference on*

10. Sous licence CC-BY 4.0.

Computational Linguistics : Technical Papers, p. 1169–1179.

EHRMANN M., WATTER C., ROMANELLO M. & CLEMATIDE S. (2020). Impresso named entity annotation guidelines.

KRIPPENDORFF K. (2004). Reliability in content analysis : Some common misconceptions and recommendations. *Human communication research*, **30**(3), 411–433.

KRIPPENDORFF K. (2018). *Content analysis : An introduction to its methodology*. Sage publications.

KRIPPENDORFF K., MATHET Y., BOUVRY S. & WIDLÖCHER A. (2016). On the reliability of unitizing textual continua : Further developments. *Quality & Quantity*, **50**(6), 2347–2364.

LEONARD V., MARKHOFF B. & ANTOINE J.-Y. (2024). Ufel : a by-design understandable and frugal entity linking system for french microposts. In *International Semantic Web Conference*, p. 253–270 : Springer.

LOPEZ C., PARTALAS I., BALIKAS G., DERBAS N., MARTIN A., REUTENAUER C., SEGOND F. & AMINI M.-R. (2017). Cap 2017 challenge : Twitter named entity recognition. *arXiv preprint arXiv :1707.07568*.

MATHET Y. (2017). *A contribution to Computational Linguistics and Natural Language Processing : From the Semantics of Space and Time to Annotations and Agreement Measures*. Thèse de doctorat, Université de Caen Normandie, France.

PETROVIĆ S., OSBORNE M. & LAVRENKO V. (2010). The edinburgh twitter corpus. In *Proceedings of the NAACL HLT 2010 workshop on computational linguistics in a world of social media*, p. 25–26.

RIZZO G., BASAVE A. E. C., PEREIRA B., VARGA A., ROWE M., STANKOVIC M. & DADZIE A. (2015). Making sense of microposts (# microposts2015) named entity recognition and linking (NEEL) challenge. In *# MSM*, p. 44–53.

ROSSET S., GROUIN C. & ZWEIGENBAUL P. (2011). Entites nommées structurées : guide d'annotation Quaero.