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

FENEC : un corpus à échantillons équilibrés pour l'évaluation des entités nommées en français

Alice Millour[1] Yoann Dupont[2] Alexane Jouglar[3] Karën Fort[3,4]
(1) UMR 6240 LISA - Università di Corsica, Av. Jean Nicoli, 20250 Corte, France
(2) ObTIC, Sorbonne Université, 4 place Jussieu, 75006 Paris, France
(3) Sorbonne Université, 28 rue Serpente, 75006 Paris, France
(4) Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy
millour_a@univ-corse.fr, yoann.dupont@sorbonne-universite.fr,
alexane.jouglar@gmail.com, karen.fort@loria.fr

RÉSUMÉ

Nous présentons ici *FENEC* (FrEnch Named-entity Evaluation Corpus), un corpus à échantillons équilibrés contenant six genres, annoté en entités nommées selon le schéma fin Quæro. Les caractéristiques de ce corpus nous permettent d'évaluer et de comparer trois outils d'annotation automatique — un à base de règles et deux à base de réseaux de neurones — en jouant sur trois dimensions : la finesse du jeu d'étiquettes, le genre des corpus, et les métriques d'évaluation.

ABSTRACT

FENEC : a balanced sample corpus for French named entity recognition

We present *FENEC* (FrEnch Named-entity Evaluation Corpus), a balanced sample corpus containing six genres and annotated with named entities according to Quæro, a rich annotation scheme. The characteristics of this corpus allow us to evaluate and compare three automatic annotation tools—one rule-based and two neural network-based—by playing on three dimensions of the evaluation: the precision of the label set, the genre of the corpora, and the evaluation metrics.

MOTS-CLÉS : Entités nommées, corpus, évaluation.

KEYWORDS: Named Entities, corpus, evaluation.

1 Introduction

La tâche de reconnaissance d'entités nommées (REN) comprend deux volets : la segmentation et la classification (ou typage) des entités. La spécificité des outils de reconnaissance automatique repose à la fois sur le type de technologie utilisé et sur le schéma d'annotation pour lequel ils ont été élaborés.

Pour pouvoir apprécier et comparer la performance de différents outils, il est donc nécessaire de les confronter à différentes métriques et sur un corpus d'évaluation suffisamment riche (en termes de taille, de finesse d'annotation, de genre). C'est dans cette perspective que nous introduisons *FENEC* [1] (FrEnch Named-entity Evaluation Corpus) : un nouveau corpus de référence pour le français librement disponible, à échantillons équilibrés appartenant à six genres différents, et annoté avec le schéma fin

1. Disponible à l'adresse suivante : https://github.com/alicemillour/FENEC, l'ensemble des documents est distribué sous licence libre, le détail est donné dans le tableau 1.

Quæro (Grouin *et al.*, 2011).

Nous présentons ensuite un échantillon des possibilités d'évaluation offerte par le corpus en y confrontant trois outils de reconnaissance d'entités nommées (EN) : `CasEN` (Friburger & Maurel, 2004; Maurel *et al.*, 2011), un système à base de transducteurs, puis `SpaCy` (Honnibal & Montani, 2017) et `Flair` (Akbik *et al.*, 2018), des modèles entraînés tous deux sur la section française de *WikiNER* (Nothman *et al.*, 2013). Notre corpus le permettant, nous proposons à la fois une évaluation globale et individuelle de ces outils.

2 Travaux connexes

Corpus existants et accessibles pour le français Il existe plus d'une dizaine de corpus annotés en entités nommées (EN) pour le français [2].

Seuls certains de ces corpus contiennent plusieurs genres, nous les présentons dans cette section. Nous pouvons notamment citer les corpus des campagnes *ESTER* (Galliano *et al.*, 2009), *ETAPE* (Gravier *et al.*, 2012), et les corpus Quæro (Rosset *et al.*, 2012) pour la parole transcrite. *ETAPE* contient quatre sous-corpus : les actualités, débats et divertissement télévisuels, ainsi que les émissions radio. Quæro contient quant à lui deux sous-corpus : presse ancienne et actualités télé-radio-diffusées. Nous pouvons y ajouter le corpus Quæro médical (Névéol *et al.*, 2014), annoté selon le même schéma d'annotation.

Pour l'écrit, nous pouvons citer le corpus *PolyCorp* (Tutin *et al.*, 2015), un corpus d'échantillons équilibrés qui contient deux sources textuelles de 2 000 tokens chacune : un texte littéraire et un rapport scientifique. Le cadre de création de ce corpus étant l'étude des expressions polylexicales, les EN ne sont pas typées. Plus récemment, Candito *et al.* (2020) ont proposé une annotation en EN du corpus *Sequoia* (Candito & Seddah, 2012) qui contient quatre sources textuelles : *Europarl* français, *L'Est Républicain*, *Wikipédia* français et *l'Agence Européenne du médicament*. Le schéma d'annotation utilisé sur ce corpus s'apparente à un schéma Quæro simplifié où ne sont gardés que les types non numériques, et où le typage utilisé est à « gros grain ». Une différence majeure avec Quæro vient de l'utilisation d'annotations discontinues : par exemple, « M. et Mme. Chirac » contient deux mentions, dont une discontinue.

À notre connaissance, il n'existait donc pas pour le français de ressource à la fois variée et accessible. C'est pourquoi, afin de faciliter la reproductibilité de notre expérience et des évaluations ultérieures, nous avons fait le choix de proposer un corpus provenant de différentes sources libres, et pouvant ainsi être redistribué et réutilisé.

Le corpus FENEC est donc la première ressource librement disponible pour le français contenant une variété de genres (six, dont de la prose, de la poésie et de la parole transcrite), présents en proportions similaires. Outre ces spécificités, nous proposons par ailleurs dans ce corpus un typage fin des entités nommées permettant d'enrichir le cadre de l'évaluation.

Évaluation de l'annotation en EN Il existe différentes métriques prenant en compte l'évaluation de différents cas de figure. Chinchor & Sundheim (1993) ont introduit cinq mesures de comparaison

2. Une liste non exhaustive et non maintenue depuis 2020 est disponible ici : `https://github.com/juand-r/entity-recognition-datasets`.

Document	Période	Genre	Nb. phrases (Nb. tokens)	Licence
42131-0 (*Traité sur la Tolérance*, Voltaire)	XVIIIe	prose	40 (1 020)	Project Gutenberg [3]
pg6470 (*Le Ventre de Paris*, Émile Zola)	XIXe		51 (1 002)	Project Gutenberg
pg6099 (*Les Fleurs du Mal*, Baudelaire)	XIXe	poésie	30 (1 014)	Project Gutenberg
56708-0 (*Œuvres d'Arthur Rimbaud - Vers et proses*)	XIXe		52 (1 027)	Project Gutenberg
UD French GSD	XXIe	multisources	35 (1 021)	CC BY-SA 4.0
Sequoia (Candito & Seddah, 2012)	XXIe		44 (1 002)	Licence LGPL-LR
French Question Bank (Seddah & Candito, 2016)	XXIe		102 (1 006)	Licence LGPL-LR
APIL (office du tourisme Othe-Armance)	XXIe	informations	29 (1 002)	Licence LGPL-LR
Wikinews	XXIe		46 (1 024)	CC BY 2.5
WikiNER français	XXIe	encyclopédie	36 (1 003)	CC BY 4.0
Spoken (Rhapsodie (Lacheret *et al.*, 2014))	XXIe	parole	70 (1 028)	CC BY-SA 4.0

TABLE 1 – Contenu du corpus annoté (échantillons de 1 000 tokens dans six genres).

des étiquettes produites par rapport aux étiquettes attendues en se plaçant au niveau des entités. Ces mesures distinguent notamment les annotations incorrectes des annotations partiellement incorrectes. Au cours de la *shared task* de CoNLL-2003 (Tjong Kim Sang & De Meulder, 2003), précision, rappel et F-mesure stricts ont été utilisés au niveau des tokens. Segura-Bedmar *et al.* (2013) reprennent les mesures de Chinchor & Sundheim (1993) au sein de quatre schémas d'évaluation. Ils obtiennent ainsi une évaluation reposant sur le calcul de quatre F-mesures qui traduisent la correspondance plus ou moins stricte des frontières et des types. Caubrière *et al.* (2020) raffinent l'extension proposée dans la campagne Quæro (Galibert *et al.*, 2011) du Slot Error Rate (Makhoul *et al.*, 1999) en proposant un typage hiérarchique des erreurs, chaque type ayant un poids qui lui est propre. Dans le cadre de l'oral transcrit, nous pouvons également citer le *Named Entity Error Rate* (NEER), proposé par Mdhaffar *et al.* (2022), qui consiste en un *Word Error Rate* calculé à l'échelle des entités nommées.

3 Annotation selon le schéma Quæro

Le choix des documents du corpus a été guidé par deux contraintes visant à obtenir un corpus (i) équilibré et (ii) pouvant être redistribué sous forme annotée. Le tableau 1 indique la composition du corpus, formé de 11 documents totalisant 11 149 tokens et appartenant à six genres principaux. La détermination de ces genres a été réalisée en fonction du contenu des textes. La principale faiblesse de cette catégorisation est que nous avons dû créer une catégorie spécifique pour les corpus multisources, alors que ceux-ci contiennent de l'encyclopédique et des informations (*news*).

Nous avons utilisé une version allégée du schéma d'annotation Quæro (Grouin *et al.*, 2011). Plus précisément, nous avons annoté uniquement les entités du schéma et pas les composants. Nous avons également annoté le type *event*, un type optionnel du schéma d'annotation, absent de la campagne d'évaluation Quæro (voir Figure 1).

Les annotations ont été réalisées au cours de deux années académiques sur la plateforme WebAnno (de Castilho *et al.*, 2014) par des étudiants en Master linguistique et informatique de Sorbonne Université ayant à leur disposition le guide d'annotation Quæro [4] et ayant reçu une brève formation à

3. Voir : https://www.gutenberg.org/policy/license.html.

4. Guide disponible à l'adresse : http://www.Qu\T1\ae{}ro.org/media/files/bibliographie/Qu\T1\ae{}ro-guide-annotation-2011.pdf.

amount	*event*	*func*	*loc*	*org*	*pers*	*prod*	*time*
8 %	3 %	9 %	31 %	8 %	20 %	7 %	14 %

TABLE 2 – Répartition des types Quæro au sein du corpus *FENEC*.

l'annotation et au schéma d'annotation. Chaque texte a été annoté par deux étudiants. Les annotations ont ensuite été validées (adjudication) par les auteurs de cet article, experts en EN.

En prenant en compte l'ensemble des annotateurs sur la première campagne, nous avons obtenu un α de Krippendorff (Krippendorff, 2013) moyen avec un intervalle de confiance de 0,95 de 0,78±0,06. En supprimant l'annotateur ayant le plus faible accord général, nous avons obtenu un α moyen de 0,82±0,03. Pour la seconde campagne, plus petite et contenant des textes du genre poétique difficiles à annoter, nous avons obtenu un α moyen de 0,35±0,43 [5]. La seconde campagne montre la difficulté supplémentaire de la tâche de REN sur des types de textes plus spécialisés.

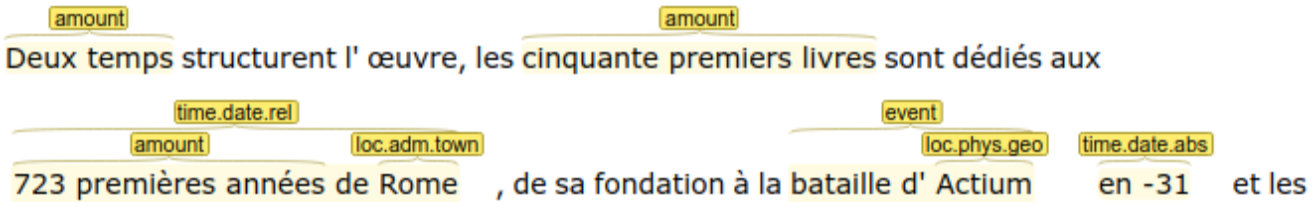


FIGURE 1 – Exemple de phrase annotée sur le corpus *WikiNER* via la plateforme WebAnno.

Après validation, 875 entités ont ainsi été annotées avec 31 étiquettes différentes pouvant être regroupées en 23 sous-types et huit types principaux. Nous donnons dans le tableau 2 la répartition des entités parmi ces derniers.

4 Évaluation des systèmes

Nous avons choisi de comparer trois systèmes de reconnaissance des EN. Le premier est `CasEN` (Friburger & Maurel, 2004; Maurel *et al.*, 2011), un système à base de cascades de transducteurs initialement évalué sur le corpus ESLO 1 (Abouda & Baude, 2006), un corpus oral de 300 heures d'interviews réalisées à Orléans, et la campagne ESTER 2 (Galliano *et al.*, 2009). Le deuxième système utilisé est `SpaCy` (Honnibal & Montani, 2017), à base de réseaux de neurones. Nous avons utilisé le modèle `fr_core_news_lg` qui a les meilleurs résultats rapportés sur la partie française de *WikiNER*. Enfin, nous avons aussi évalué le modèle `flair/ner-french`, mis à disposition par `Flair` (Akbik *et al.*, 2018) *via* la plateforme Hugging Face [6], qui utilise un réseau neuronal de type *Long short-term memory* (*LSTM*) (Hochreiter & Schmidhuber, 1997) bidirectionnel et des plongements contextuels. Comme celui de `SpaCy`, ce modèle a été entraîné sur la partie française de *WikiNER*.

5. Ces accords ont été calculés en utilisant la librairie *Rmisc*, voir : https://www.rdocumentation.org/packages/Rmisc/versions/1.5.

6. Voir : https://huggingface.co/flair/ner-french.

4.1 Faire correspondre les schémas d'annotation

Méta-étiquettes et corpus d'évaluation Afin d'évaluer individuellement les trois outils sur le corpus présenté, nous avons construit deux jeux de méta-étiquettes à l'intersection du schéma d'annotation Quæro (huit types permettant d'annoter 875 entités) et des jeux d'étiquettes utilisés par `CasEN` d'une part (huit types permettant d'annoter 843 entités) [7] et par `SpaCy` et `Flair` d'autre part (jeu minimal de quatre types permettant d'annoter 678 entités). Le schéma d'annotation correspondant au jeu minimal repose sur le guide de la campagne CoNLL-2003 [8] (Tjong Kim Sang & De Meulder, 2003). Les jeux de méta-étiquettes et leur correspondance avec les jeux d'origine sont donnés dans les annexes A et B.

Pour mener une comparaison stricte des trois outils, nous avons également réalisé une troisième mise en correspondance des étiquettes produites par `CasEN` vers le jeu minimal de quatre types.

Le schéma Quæro étant le plus complet, on constate une diminution du nombre d'entités après transformation du corpus d'évaluation selon les différents jeux de méta-étiquettes. En particulier, les entités de types *montants* et *dates* ont été retirées des évaluations pour `SpaCy` et `Flair`.

Comparaison des schémas d'annotation L'évaluation d'outils ayant été entraînés sur des corpus spécialisés sur un nouveau corpus pose le problème de l'adéquation des schémas d'annotation utilisés pour l'entraînement et pour l'évaluation respectivement.

Afin de mesurer cette correspondance de schémas, nous avons comparé un échantillon du corpus *WikiNER* ayant été annoté semi-automatiquement et utilisé pour l'entraînement de `SpaCy` avec deux échantillons annotés manuellement avec (i) le jeu Quæro, et (ii) le jeu minimal.

Le tableau 3 permet d'observer la forte précision entre les deux schémas : une entité du *WikiNER* est généralement une entité du schéma Quæro. Le rappel augmente naturellement lorsque la comparaison est menée sur le jeu minimal moins couvrant.

échantillon	strict			partiel		
	P	R	F	P	R	F
WikiNER sur Quæro (sans typage)	0,91	0,56	0,69	0,95	0,58	0,72
WikiNER sur Quæro	0,83	0,51	0,63	0,88	0,53	0,66
WikiNER sur jeu minimal (sans typage)	0,91	0,83	0,87	0,95	0,86	0,90
WikiNER sur jeu minimal	0,84	0,79	0,82	0,89	0,84	0,87

TABLE 3 – Comparaison des annotations de référence de *WikiNER* et de notre corpus de référence.

Les différences observées entre les deux annotations proviennent principalement d'incohérences de segmentation dans l'annotation semi-automatique du corpus *WikiNER*, ce qui produit des erreurs aux

7. La finesse du jeu d'étiquettes Quæro permettrait une étude plus détaillée des performances de `CasEN` (par exemple *via* la mise en correspondance de l'étiquette *loc.adm.nat* avec l'étiquette *placeName* assortie de la composante *country*). Notons néanmoins que, comme il a été soulevé par un relecteur ou une relectrice que nous remercions ici, la mise en correspondance fine des étiquettes n'est pas toujours évidente, comme c'est le cas pour l'étiquette *demonym* qui n'a pas la même interprétation selon le jeu considéré. Dans le cadre de cet article porté sur la comparaison, nous nous sommes limités à des jeux de méta-étiquettes englobant ces différents types fins.

8. Disponible à l'adresse : `https://www.clips.uantwerpen.be/conll2003/ner/annotation.txt`.

niveaux des frontières des entités. On trouve parmi ces erreurs l'inclusion ou non du déterminant précédant l'EN, ou l'annotation sous forme séparée (deux annotations) ou regroupée (une annotation) de lieux du type "*Chine du Nord*".

Les exemples présentés dans les figures 2 et 3 permettent d'apprécier la plus forte correspondance avec le schéma minimal (*dates* et *amount* ne sont pas annotés ; l'annotation *prod* devient *MISC*). On y constate aussi certaines incohérences de typage [9] : c'est le cas de l'annotation en *ORG* d'une part et en *prod.media* [10] d'autre part de l'entité `New York Times`. Nous avons signalé par une astérisque les correspondances d'étiquettes que nous avons établies pouvant être discutées : nous avons fait le choix de faire correspondre *func* et *MISC*, mais l'étude du guide CONLL-2003 couplée à celle des annotations présentes dans *WikiNER* ne permet pas réellement de valider ou d'invalider cette correspondance. Le cas des annotations multiples *LOC+ORG* propres à la métonymie [11] conduit à une correspondance partielle et montre les limites de l'annotation du corpus *WikiNER*. Enfin, l'absence de l'étiquette *event* pour l'entité `bataille d'Actium` dans ce dernier est une erreur imputable au caractère semi-automatique de son annotation.

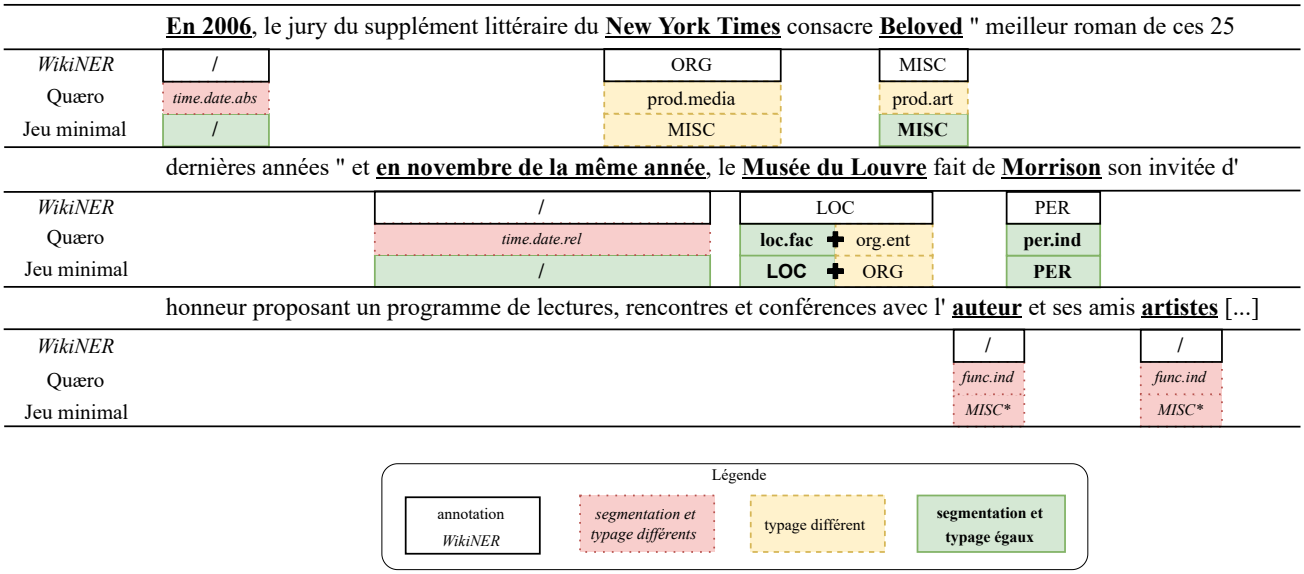


FIGURE 2 – Comparaison de *WikiNER* avec les schémas d'annotations de Quæro et du jeu minimal.

Concernant `CasEN`, le jeu d'étiquettes utilisé est très proche du jeu Quæro (voir B), tous deux ayant été construits en cohérence avec la définition d'*ESTER 2*. Notons néanmoins que `CasEN` utilise le type `ref` à la fois pour les références numériques (par exemple "[1]") et pour les titres d'ouvrages, seuls annotés dans Quæro. Ces annotations sont très minoritaires dans le corpus.

4.2 Évaluations et comparaisons

Les graphiques de la figure 4 permettent de visualiser les résultats des outils selon deux évaluations différentes. Le premier contient les F-mesures strictes obtenues par genre et globalement, le second

9. Nous en avons identifié une autre qui concerne les groupes de musique, annotés `ORG` dans *WikiNER* et `pers.coll` dans Quæro.
10. L'ambiguïté pouvant exister avec l'étiquette *org.ent* est résolue dans le guide Quæro, section 1.3.4. (Rosset *et al.*, 2011).
11. Voir la section 1.4 du guide Quæro (Rosset *et al.*, 2011).

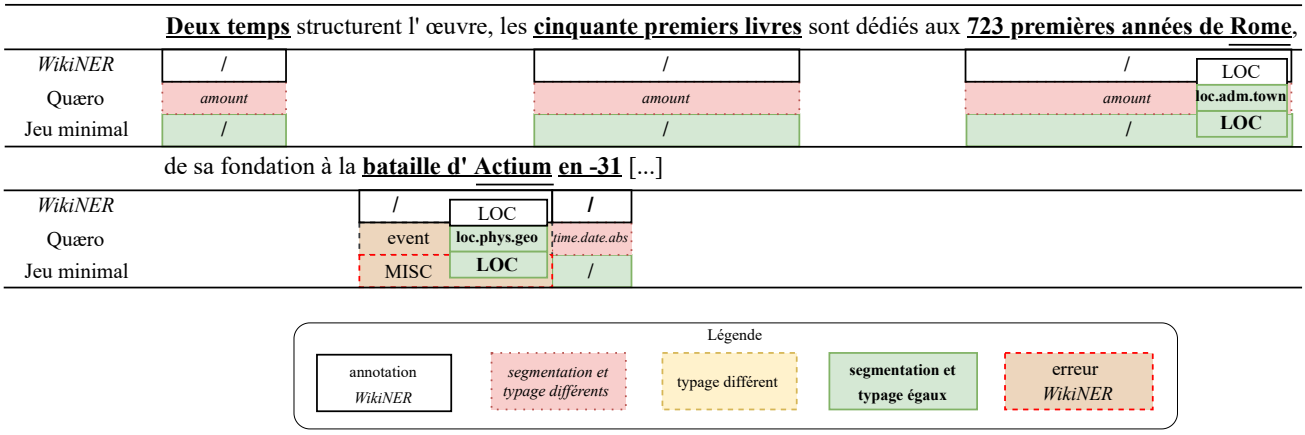


FIGURE 3 – Comparaison de *WikiNER* avec les schémas d'annotations de Quæro et du jeu minimal.

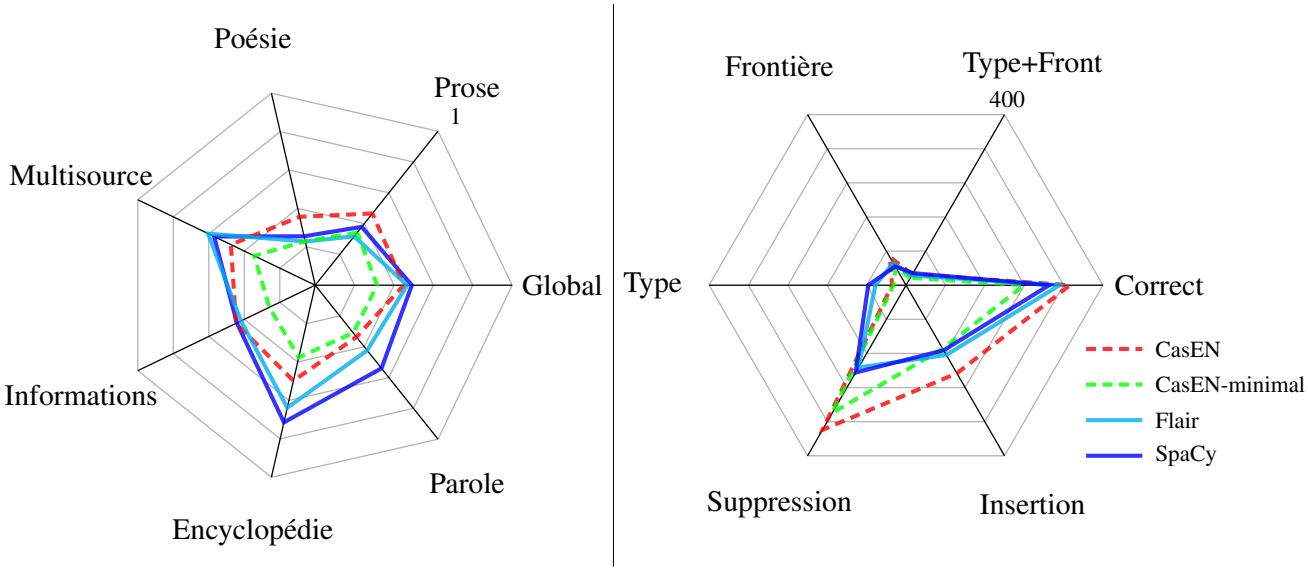


FIGURE 4 – À gauche, les scores de F-mesure stricte des trois outils selon les six genres et globalement. À droite, le nombre de résultats corrects et d'erreurs (type, frontière, type+frontière, suppression et insertion) pour les trois outils.

donne les nombres d'entités correctement reconnues, ainsi que les nombres d'erreurs de substitution, d'insertion et de suppression sur le corpus global selon la typologie définie par Galibert *et al.* (2011).

Nous présentons deux résultats avec `CasEN`. L'un où les correspondances ont été établies en fonction des sorties de l'outil (`CasEN` dans les légendes) et l'autre avec les annotations de `CasEN` réduites au schéma d'annotation de *WikiNER* (`CasEN-minimal` dans les légendes).

Les performances de `CasEN` sont généralement plus faibles avec le schéma minimal, ce que nous n'avons pu voir que grâce au choix que nous avons fait de réaliser une annotation fine des EN. Cependant, elles sont comparables globalement avec celles des autres outils (Global). Nous notons que les F-mesures de `CasEN` sont relativement stables d'un genre à l'autre et sont généralement meilleures sur des genres qui diffèrent de l'encyclopédique (poésie et prose). Par ailleurs, si les F-mesures des outils par apprentissage sont très similaires, `Spacy` semble cependant avoir de meilleures performances sur la parole. Ces outils présentent les meilleures performances sur les

genres multisources et encyclopédie. Le graphe de droite montre que la plupart des outils semble avoir des profils similaires d'erreurs, ce qui pourrait indiquer qu'il existe une base d'entités qui mettent les outils en difficulté.

5 Conclusion et perspectives

Notre expérience montre à la fois la difficulté de comparer des outils réalisant la même tâche, l'intérêt de disposer d'une ressource d'évaluation riche, et l'importance d'utiliser différentes métriques d'évaluation. L'évaluation commune d'outils sur un même jeu d'étiquettes minimal, si elle permet une comparaison stricte, pénalise les outils développés pour un schéma plus complexe, introduisant de ce fait un biais dans notre appréciation des performances. L'expérience demande à être enrichie, notamment par un analyse plus fine des performances par type.

Par ailleurs, nous souhaitons compléter nos évaluations en utilisant le paradigme d'évaluation proposé par Fu *et al.* (2020). Pour comprendre les contextes d'erreurs et les marges de progression des outils évalués, de nouvelles métriques intégrant l'observation de traits pertinents des entités (par exemple l'ambiguïté et la fréquence des tokens, la longueur des entités et des phrases, etc.) y sont employées. L'utilisation de cette méthodologie sur notre corpus permettrait de progresser vers une meilleure interprétation des performances observées. L'utilisation de correspondances de types strictes pour harmoniser les jeux d'étiquettes est source de certains biais dans l'évaluation, notamment pour les imbrications ou des différences sur des types d'entités spécifiques. Des systèmes plus complexes permettraient d'obtenir une correspondance plus fidèle sur ces cas.

À l'heure actuelle, notre corpus contient 11 000 tokens (à titre de comparaison, cela représente un volume équivalent à environ 30 % du corpus de test du corpus arboré de Paris 7 ou *French Treebank* (Abeillé *et al.*, 2003) tel que découpé dans Crabbé & Candito (2008)). Nous remercions le relecteur ou la relectrice qui nous a indiqué le corpus *80 jours* [12] (Lecuit *et al.*, 2013), disponible sous licence CC BY-NC-SA, et dont les noms propres, noms et adjectifs relationnels ont été annotés selon le schéma proposé par la Text Encoding Initiative Consortium (TEI P5) [13]. Celui-ci pourra en effet être intégré à FENEC après harmonisation du schéma d'annotation.

Par ailleurs, et bien que nos échantillons soient équilibrés, les genres dans lesquels ils sont distribués le sont moins. Ainsi, le genre dialogique est sous représenté dans le corpus par rapport aux autres et n'est pas présent dans les corpus multisources (à la différence du genre encyclopédique). Une perspective de notre travail est donc d'élargir *FENEC* en équilibrant davantage les genres.

Remerciements

Nous tenons à remercier les relecteurs pour leurs suggestions constructives, qui nous ont permis d'enrichir l'article. Nous remercions également les étudiants de M2 du Master Linguistique et informatique de Sorbonne Université qui ont réalisé une première annotation du corpus dans le cadre de leur cours sur l'annotation en 2020 et 2021.

12. Voir : https://tln.lifat.univ-tours.fr/version-francaise/ressources/corpus-80-jours.
13. Voir : https://tei-c.org/.

Références

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). Building a treebank for french. In *Treebanks*, p. 165–187. Springer.

ABOUDA L. & BAUDE O. (2006). Constituer et exploiter un grand corpus oral : choix et enjeux théoriques. le cas des eslo. In *Corpus en Lettres et Sciences sociales, Des documents numériques à l'interprétation*.

AKBIK A., BLYTHE D. & VOLLGRAF R. (2018). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, p. 1638–1649, Santa Fe, Nouveau Mexique, USA : Association for Computational Linguistics.

CANDITO M., CONSTANT M., RAMISCH C., SAVARY A., GUILLAUME B., PARMENTIER Y. & CORDEIRO S. (2020). A french corpus annotated for multiword expressions and named entities. *Journal of Language Modelling*, **8**(2), 415–479.

CANDITO M. & SEDDAH D. (2012). Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *Actes de Traitement Automatique des Langues Naturelles (TALN)*, Grenoble, France.

CAUBRIÈRE A., ROSSET S., ESTÈVE Y., LAURENT A. & MORIN E. (2020). Where are we in named entity recognition from speech ? In *Proceedings of the 12th Language Resources and Evaluation Conference*, p. 4514–4520, Marseille, France : European Language Resources Association.

CHINCHOR N. & SUNDHEIM B. (1993). MUC-5 evaluation metrics. In *Fifth Message Understanding Conference (MUC-5) : Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993*.

CRABBÉ B. & CANDITO M. (2008). Expériences d'analyse syntaxique statistique du français. In *15ème conférence sur le Traitement Automatique des Langues Naturelles - TALN'08*, p. pp. 44–54, Avignon, France. HAL : hal-00341093.

DE CASTILHO R. E., BIEMANN C., GUREVYCH I. & YIMAM S. M. (2014). WebAnno : a flexible, web-based annotation tool for CLARIN. In *Proceedings of the CLARIN Annual Conference (CAC) 2014*, p. online, Utrecht, Netherlands : CLARIN ERIC. Extended abstract.

FRIBURGER N. & MAUREL D. (2004). Finite-state transducer cascades to extract named entities in texts. *Theoretical Computer Science*, **313**(1), 93–104.

FU J., LIU P. & NEUBIG G. (2020). Interpretable multi-dataset evaluation for named entity recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 6058–6069, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.489.

GALIBERT O., ROSSET S., GROUIN C., ZWEIGENBAUM P. & QUINTARD L. (2011). Structured and extended named entity evaluation in automatic speech transcriptions. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, p. 518–526, Chiang Mai, Thaïlande : Asian Federation of Natural Language Processing.

GALLIANO S., GRAVIER G. & CHAUBARD L. (2009). The ESTER 2 evaluation campaign for the rich transcription of french radio broadcasts. In *Tenth Annual Conference of the International Speech Communication Association*, p. 2583–2586, Brighton, Angleterre.

GRAVIER G., ADDA G., PAULSSON N., CARRÉ M., GIRAUDEL A. & GALIBERT O. (2012). The ETAPE corpus for the evaluation of speech-based TV content processing in the French language. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, p. 114–118, Istanbul, Turquie : European Language Resources Association (ELRA).

GROUIN C., ROSSET S., ZWEIGENBAUM P., FORT K., GALIBERT O. & QUINTARD L. (2011). Proposal for an extension of traditional named entities : From guidelines to evaluation, an overview. In *5th Linguistic Annotation Workshop*, p. 92–100, Portland, Oregon, USA. Poster.

HOCHREITER S. & SCHMIDHUBER J. (1997). Long short-term memory. *Neural computation*, **9**(8), 1735–1780.

HONNIBAL M. & MONTANI I. (2017). spaCy 2 : Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, **7**(1), 411–420.

KRIPPENDORFF K. (2013). *Content Analysis : An Introduction to Its Methodology*. Thousand Oaks, CA : Sage, 3rd edition édition.

LACHERET A., KAHANE S., BELIAO J., DISTER A., GERDES K., GOLDMAN J.-P., OBIN N., PIETRANDREA P. & TCHOBANOV A. (2014). Rhapsodie : un Treebank annoté pour l'étude de l'interface syntaxe-prosodie en français parlé. In *4e Congrès Mondial de Linguistique Française*, volume 8, p. 2675–2689, Berlin, Allemagne. DOI : 10.1051/shsconf/20140801305, HAL : halshs-01061368.

LECUIT É., MAUREL D. & VITAS D. (2013). Mise en ligne du corpus aligné des traductions du Tour du monde en quatre-vingts jours (Jules Verne, 1872) en français (annoté en entités nommées), anglais, allemand et serbe, sur le site TLN. Corpus aligné en quatre langues avec entités nommées annotées en français, HAL : hal-01229502.

MAKHOUL J., KUBALA F., SCHWARTZ R. & WEISCHEDEL R. (1999). Performance measures for information extraction. In *Proceedings of DARPA Broadcast News Workshop*, p. 249–252.

MAUREL D., FRIBURGER N., ANTOINE J.-Y., ESHKOL I. & NOUVEL D. (2011). Cascades de transducteurs autour de la reconnaissance des entités nommées. *Traitement automatique des langues*, **52**(1), 69–96.

MDHAFFAR S., DURET J., PARCOLLET T. & ESTÈVE Y. (2022). End-to-end model for named entity recognition from speech without paired training data. DOI : 10.48550/ARXIV.2204.00803.

NOTHMAN J., RINGLAND N., RADFORD W., MURPHY T. & CURRAN J. R. (2013). Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, **194**, 151–175.

NÉVÉOL A., GROUIN C., LEIXA J., ROSSET S. & ZWEIGENBAUM P. (2014). The QUAERO French medical corpus : A ressource for medical entity recognition and normalization. In *Proc of BioTextMining Work*, p. 24–30.

ROSSET S., GROUIN C., FORT K., GALIBERT O., KAHN J. & ZWEIGENBAUM P. (2012). Structured named entities in two distinct press corpora : Contemporary broadcast news and old newspapers. In *6th Linguistic Annotation Workshop (LAW VI)*, p. 40–48, Jeju, Corée.

ROSSET S., GROUIN C. & ZWEIGENBAUM P. (2011). *Entités Nommées Structurées : guide d'annotation Quaero*. Notes et Documents LIMSI N° : 2011-04. LIMSI-Centre national de la recherche scientifique. http://www.quaero.org/media/files/bibliographie/quaero-guide-annotation-2011.pdf.

SEDDAH D. & CANDITO M. (2016). Hard time parsing questions : Building a QuestionBank for French. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 2366–2370, Portoroz, Slovénie : European Language Resources Association (ELRA).

SEGURA-BEDMAR I., MARTÍNEZ P. & HERRERO-ZAZO M. (2013). SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In *Second Joint*

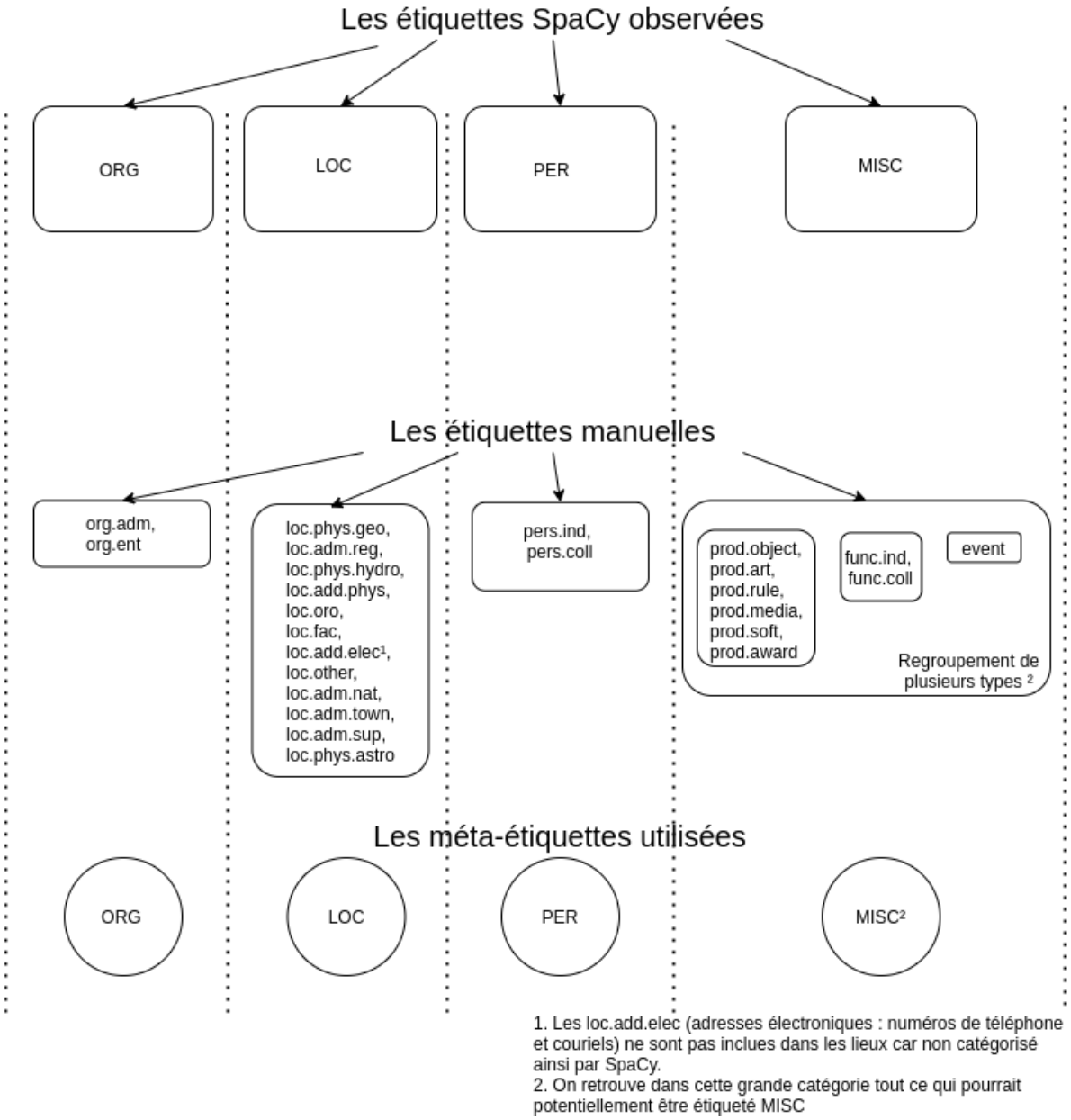
*Conference on Lexical and Computational Semantics (*SEM), Volume 2 : Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, p. 341–350, Atlanta, Georgia, USA : Association for Computational Linguistics.

TJONG KIM SANG E. F. & DE MEULDER F. (2003). Introduction to the CoNLL-2003 shared task : Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, p. 142–147.

TUTIN A., ESPERANÇA-RODIER E., IBORRA M. & REVERDY J. (2015). Annotation of multiword expressions in French. In C.-P. GLORIA, Éd., *European Society of Phraseology Conference (EUROPHRAS 2015)*, p. 60–67, Malaga, Espagne. HAL : hal-01348549.

Annexes

A Méta-étiquettes minimales utilisées pour évaluer SpaCy et Flair



B Méta-étiquettes minimales utilisées pour évaluer `CasEN`