HAL open science

Un corpus libre, évolutif et versionné en entités nommées du français

Yoann Dupont

To cite this version:

Yoann Dupont. Un corpus libre, évolutif et versionné en entités nommées du français. TALN 2019 - Traitement Automatique des Langues Naturelles, Jul 2019, Toulouse, France. hal-02448590

HAL Id: hal-02448590

https://hal.science/hal-02448590v1

Submitted on 22 Jan 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Un corpus libre, évolutif et versionné en entités nommées du français

Yoann Dupont
LIFO, université d'Orléans, 6 rue Léonard de Vinci BP 6759, 45067 Orléans cedex 2
yoann.dupont@univ-orleans.fr

RÉSUMÉ
Les corpus annotés sont des ressources difficiles à créer en raison du grand effort humain qu'elles impliquent. Une fois rendues disponibles, elles sont difficilement modifiables et tendent à ne pas évoluer pas dans le temps. Dans cet article, nous présentons un corpus annoté pour la reconnaissance des entités nommées libre et évolutif en utilisant les textes d'articles Wikinews français de 2016 à 2018, pour un total de 1191 articles annotés. Nous décrivons succinctement le guide d'annotation avant de situer notre corpus par rapport à d'autres corpus déjà existants. Nous donnerons également un accord intra-annotateur afin de donner un indice de stabilité des annotations ainsi que le processus global pour poursuivre les travaux d'enrichissement du corpus.

ABSTRACT
A free, evolving and versioned french named entity recognition corpus.

Annotated corpora are very hard resources to make because of the high human cost they imply. Once released, they are hardly modifiable and tend to not evolve through time. In this article we present a free and evolving corpus annotated in named entity recognition based on French Wikinews articles from 2016 to 2018, for a total of 1191 articles. We will briefly describe the annotation guidelines before comparing our corpus to various corpora of comparable nature. We will also give an intra-annotator-agreement to provide an estimation of the stability of the annotation as well as the overall process to develop the corpus.

MOTS-CLÉS : reconnaissance des entités nommées, annotation manuelle, corpus annoté.

KEYWORDS: named entity recognition, manual annotation, annotated corpus.

1 Introduction

La reconnaissance des entités nommées est une tâche importante du TAL. Elle permet *« l'accès à l'information »* (Nouvel *et al.*, 2015) pour d'autres tâches, comme par exemple la construction d'une base de connaissances (Surdeanu, 2013) ou les systèmes de questions-réponses (Han *et al.*, 2017). La notion d'entité nommée a évolué avec le temps. Tout d'abord considérées comme *« tous les noms propres et quantités d'intérêt »* dans la campagne MUC-6 (Grishman & Sundheim, 1996), où les entités cibles étaient les personnes, lieux, organisations, temps et pourcentages, leur reconnaissance devait aider au remplissage automatique de formulaire. La campagne ACE (Doddington *et al.*, 2004) a donné comme périmètre aux entités nommées les personnes, les organisations, les lieux, les bâtiments, les armes, les véhicules et les entités géo-politiques, dans le cadre de la détection d'événements. Sekine & Nobata (2004), quant à eux, définissent 150 types d'entités nommées organisés de façon

hiérarchique, afin de couvrir un maximum de cas d'utilisation. Ils préconisaient d'élaguer la hiérarchie afin de correspondre au mieux au cas d'usage particuliers. Grouin *et al.* (2011) proposent également une définition d'entités nommées généraliste et couvrante. Ils définissent, en plus des types d'entités, leurs *composants* ainsi que leur structuration. Les entités nommées ont un caractère référentiel et, comme nous venons de le voir, ont également une visée applicative très concrète et sont fortement liées à leur corpus. Un définition rendant compte de toutes ces caractéristiques est celle d'Ehrmann (2008) : *« étant donnés un modèle applicatif et un corpus, on appelle entité nommée toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus. »*.

Un inconvénient des corpus actuellement disponibles est leur côté figé dans le temps. En effet, le FTB (Abeillé *et al.*, 2003) annoté en entités nommées (Sagot *et al.*, 2012) contient des phrases extraites d'articles du Monde de 1989 à 1995, la partie anglaise du corpus CoNLL 2003 (Tjong Kim Sang & De Meulder, 2003) est un extrait du corpus Reuters d'Août 1996 à Août 1997, la partie allemande comprend, elle, des textes de l'année 1992. Des corpus sur des données plus récentes existent, comme par exemple le corpus d'oral transcrit ESTER2 (Galliano *et al.*, 2009) qui couvre les années 1999 à 2003, dont les transcriptions ont été utilisées pour le corpus Quaero (Galibert *et al.*, 2010; Rosset *et al.*, 2012). Il nous parait important de diffuser des corpus dont les sources sont aussi récentes que possible, pour plusieurs raisons. La première est d'élargir l'éventail des entités couvertes dans les corpus que nous avons à notre disposition, en effet, de nombreuses nouvelles entités nommées apparaissent régulièrement dans les textes journalistiques et il est important de pouvoir les attester. Un autre intérêt, qui découle du précédent, est que cela permet d'évaluer les systèmes construits pour ces tâches particulières et d'évaluer dans des conditions réalistes leur qualité et leur robustesse.

Au-delà des divergences définitoires, comme le fait d'annoter ou non le titre d'une personne, les corpus annotés ne sont pas exempts d'erreurs. Il existe divers articles les évoquant ou les corrigeant, nous pouvons citer Finkel *et al.* (2004) pour le corpus GENIA (Kim *et al.*, 2003) et Nouvel *et al.* (2010) pour ESTER2. Pourtant, il a été montré que les erreurs humaines étaient une source d'erreur pour les systèmes à base d'apprentissage (Boisen *et al.*, 2000). Pour ces raisons, il semble important de créer des ressources évolutives pour qu'elles soient capables de rendre compte des phénomènes émergeant, plus ouvertes afin de faciliter leur mise à jour. Il semble aussi important de créer des ressources versionnées, afin d'intégrer au mieux les corrections d'erreurs trouvées au fil des travaux.

Nous proposons ici un corpus libre et évolutif annoté en entités nommées [1] afin de remédier au mieux aux problèmes cités plus haut. Ce corpus contient des textes très récents. Nous nous plaçons donc ici dans la continuité des travaux effectués par Salmon-Alt *et al.* (2004); Hernandez & Boudin (2013), dont l'idée est de fournir des corpus annotés libres. Nous nous plaçons dans la continuation des travaux d' Hernandez & Boudin (2013) en décidant d'annoter des données issues de la partie française de Wikinews, et que l'annotation en entités nommées que nous proposons se veut enrichir celle en parties-du-discours déjà présente. Cependant, nous nous en distinguons de par le type d'annotations que nous effectuons, nous avons ici privilégié une annotation manuelle sur une partie plus petite. Nous souhaitons, à terme, que ce corpus puisse être utilisé afin d'entraîner des systèmes automatiques pour des tâches d'extraction d'information, qui impliquent souvent d'effectuer en amont la reconnaissance des entités nommées, l'*entity linking* ainsi que l'extraction des relations qui lient les entités nommées.

1. disponible à l'adresse : `https://github.com/YoannDupont/WiNER-fr`

2 Le corpus et son annotation

Le corpus contient actuellement les contenus textuels des articles de Wikinews français des années 2016 à 2018, pour un total de 1191 articles. Nous avons pris l'ensemble des articles Wikinews, à l'exception d'un certain type de document : les tableaux de résultats sportifs. Ces documents contiennent typiquement une unique phrase décrivant la compétition ayant eu lieu suivi d'un tableau de scores.

Un unique annotateur ayant de fortes connaissances dans la reconnaissance des entités nommées a annoté l'ensemble des documents. Afin d'accélérer le processus d'annotation, nous avons utilisé un outil spécifique codé en python avec la librairie Tkinter (Shipman, 2013). L'outil permet de sélectionner des empans de texte et de leur attribuer un type d'une simple touche de clavier, avec la possibilité de diffuser l'annotation à l'échelle du document. Afin d'accélérer encore le processus d'annotation, nous avons également régulièrement entraîné un système par apprentissage afin de donner une pré-annotation pour un document. Nous nous sommes inspirés des systèmes décrits par Dupont (2017), qui donnent à notre connaissance des performances état-de-l'art sur le FTB. Plus précisément, nous avons utilisé le modèle utilisant des CRF (Lafferty *et al.*, 2001). Nous avons fait ce choix en prenant en compte le rapport entre la correction et le temps de traitement. Bien que la correction du système utilisant les Bi-LSTM-CRF (Lample *et al.*, 2016) soit meilleure, la rapidité à l'entraînement et à l'annotation du système utilisant des CRF permettaient d'avoir un processus d'annotation des documents globalement plus rapide. Fort *et al.* (2009) indiquent que la pré-annotation *« introduit un biais en faveur de la correction des pré-annotations, au détriment de la recherche de nouvelles EN »*. Pour réduire ce biais, nous avons procédé à l'annotation de chaque document pré-annoté en deux temps : un premier temps où les pré-annotations ont été corrigées et un second où le document était repris depuis le début à la recherche d'annotations manquées. Une fois le corpus annoté, nous avons utilisé un script pour détecter de potentielles incohérences. Le script crée d'abord un lexique par type d'entités en utilisant les différentes mentions annotés comme entrées. Le lexique ainsi créé est alors appliqué sur le corpus, les différences en type ou en frontière sont alors indiquées pour vérification par l'annotateur. Grâce à l'ensemble de ces méthodes afin d'accélérer le processus d'annotation, l'annotation manuelle, c'est-à-dire sans compter les temps d'entraînement du modèle par apprentissage, a pris une semaine à l'annotateur. Une révision manuelle de l'ensemble des annotations du corpus n'a pas encore été effectuée.

Nous avons annoté le corpus en nous basant sur un jeu simplifié des étiquettes du Quaero. Nous avons annoté les types suivants : les dates, les événements, les heures, les lieux, les organisations, les personnes et les produits (sans hiérarchie). Notre ensemble d'étiquette est globalement un sous-ensemble du jeu d'étiquettes défini par Quaero. Ce jeu d'étiquettes est comparable à celui de CoNLL-2003, où des entités temporelles et les événements remplacent le type "MISC". Les dates sont ici des dates dites absolues, c'est-à-dire que nous annotons uniquement les dates référant à un jour unique (1er janvier 2019), référant à un jour ou un mois identifiable sur un calendrier (1er janvier, janvier, janvier 2019), les décennies, les siècles, millénaires et les périodes clairement identifiées (par exemple, « les trente glorieuses » désignent une période précise). Nous n'annotons cependant pas les dates relatives comme « la veille », « le mois prochain ». Dans le cas d'une date relative par rapport à une date absolue (selon la définition précédente) comme « 1er janvier prochain », il a été décidé d'annoter « 1er janvier » comme une date absolue. La notion d'événements dans notre corpus est identique à celle de Quaero. Elle désigne, entre autres, les tournois sportifs (championnats du monde de patinage), les congrès, les événements annuels, les fêtes. Nous avons décidé d'annoter les événements climatiques ainsi que les affaires politico-juridiques, ces derniers étant laissés à l'appréciation de l'annotateur.

Nous avons donc ici décidé d'être le plus couvrant possible, ce qui permet de simplifier le travail d'annotation en réduisant les incertitudes par rapport à certains types d'événements. Les heures sont, à l'instar des dates, les heures dites absolues. Les lieux ici sont équivalents à l'ensemble des types de lieux définis dans le Quaero. Les organisations correspondent ici aux types « organisation » et « entreprise ». Les personnes désignent les personnes aussi bien réelles que fictives. Nous n'annotons que les prénoms, noms et surnoms. Les titres, fonctions, nationalités, etc. ne sont pas annotés. Ne sont pas annotés non plus les groupes de personnes, comme les familles lorsqu'elles sont désignées par leur nom de famille. Les produits désignent les différents produits qui sont présents dans le jeu d'étiquettes Quaero. Parmi eux, nous trouvons principalement les objets physiques (Airbus A380), les logiciels (jeux vidéos, etc.) et les produits médiatiques (émission de radio, TV, etc.).

Nous avons décidé d'annoter les composants des entités si ces derniers étaient également des entités nommées dans l'absolu. Par exemple, « Tour de France 2016 » est une entité de type événement contenant deux composants, à savoir « France » qui est annoté comme un lieu et « 2016 » qui est annoté comme une date. Nous avons décidé de le faire de manière systématique, même lorsque le lien n'est pas forcément évident. Par exemple dans « Université Leland Stanford Junior », « Leland Stanford Junior » réfère au fils des fondateurs, nous avons donc décidé de l'annoter. Un intérêt à cette annotation, malgré le biais positif qu'elle donne aux systèmes automatiques, est de faciliter le futur passage au schéma d'annotation Quaero. Des exemples d'annotations sont fournis dans la figure 1.

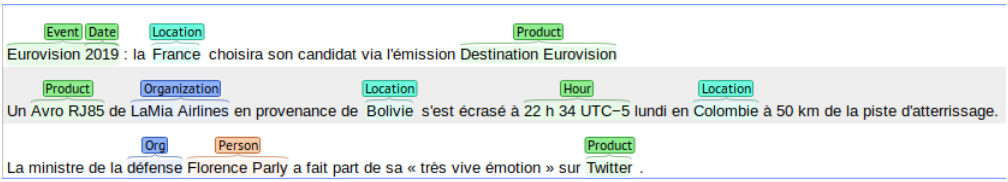

FIGURE 1 – Quelques exemples d'annotations visualisés avec l'outil BRAT.

Le corpus est en accès libre et se présente sous la forme d'un projet git hébergé sur GitHub. Nous utilisons un versionnement sémantique, c'est-à-dire un schéma "majeur.mineur.correctif". Une version majeure représente une annotation du corpus revue manuellement dans son intégralité. Une version mineure, l'ajout de nouveaux documents annotés avec des annotations non revues manuellement en vue d'une prochaine version majeure. Une version corrective implique uniquement la correction manuelle de l'annotation du corpus, aucun document ne pourra être ajouté.

La mise en ligne d'une nouvelle version majeure du corpus implique que ce dernier soit révisé. Si des portions du corpus ont déjà été vérifiées pour une version précédente du corpus et n'ont pas été modifiées depuis, leur vérification est optionnelle. Une version majeure du corpus ne pourra plus se voir attribuer de nouveaux documents, mais pourra toujours recevoir des *patchs* afin d'améliorer la correction des annotations. Des versions béta du corpus peuvent être diffusées de manière régulière selon l'avancement des annotations. À l'heure actuelle, nous souhaitons que ces versions béta correspondent à l'ajout d'années "pleines" (mois ou années) afin de limiter le nombre de sorties. Ces versions béta peuvent contenir des annotations non révisées.

Les ajouts et le corrections se font sur la base du volontariat. Afin de gérer au mieux les conflits, une fois la première version corrigée du corpus sortie, nous maintiendrons des branches séparées de la branche principale afin d'intégrer les modifications. Les mainteneurs du projet pourrons intégrer leurs changements simplement. Les participations d'une personne extérieure devront être faite via une

type entité	compte	compte uniques
Date	3817	1386
Event	720	340
Hour	945	426
Location	8719	2187
Organization	4215	1628
Person	5055	2499
Product	673	207
global	24144	8673

TABLE 1 – les comptes des entités par type. « uniques » compte les formes de surface différentes.

requête d'audit (*pull request*) et devront être vérifiées par un mainteneur du projet. Ces vérifications devront s'assurer de la cohérence des annotations avec le schéma.

3 Mesures relatives au corpus

Le corpus des années 2016 à 2018 de Wikinews français comporte un total de 1191 documents pour 322931 tokens (selon une segmentation automatique). Comme indiqué dans le tableau 1, le corpus comprend un total de 24144 annotations (pour 1122 annotations imbriquées, soit environ 4% du volume total des annotations).

Le corpus est actuellement annoté au format BRAT (Stenetorp *et al.*, 2012) : un premier fichier contient le contenu textuel du document et un second fichier contient les annotations déportées. Ce choix a deux motivations principales. Premièrement, il permet d'obtenir un historique très clair des modifications faites aux annotations dans un logiciel de gestion de version. Le second intérêt est que ce format est que nous pouvons utiliser directement l'outil BRAT afin de lier les mentions à une base (désambiguïsation des entités, ou *named entity linking*) ainsi que les relations qui lient les différentes entités entre elles. Ce choix est donc également motivé par les ajouts prévus au corpus.

Comme indiqué dans la section 2, nous avons utilisé un script afin de trouver des erreurs d'annotation de manière semi-automatique. Afin d'évaluer l'apport de l'utilisation de ce script, nous avons comparé la première et la dernière version de chaque fichier. En évaluant les différences, nous pouvons donner une estimation de l'apport d'une recherche d'erreurs simple. Nous avons calculé cet accord selon deux points de repères : en prenant l'ensemble des documents et en prenant uniquement ceux ayant subi au moins une modification par l'annotateur. Au total, 183 documents ont subi des modifications depuis leur première version, soit 15%. La f-mesure entre la première et la dernière version de chaque article du corpus vaut 0.981 sur l'ensemble du corpus.

Afin de fournir un indice de la stabilité des annotations manuelles, nous avons calculé un accord intra-annotateur (Krippendorff, 2018). Nous avons sélectionné au hasard 1 document par mois, pour un total de 36 documents, en ignorant les documents ayant subi au moins une correction entre leur première et leur dernière version. Les documents ont été annotés sans pré-annotation. Nous avons calculé l'accord intra-annotateur à l'aide de la f-mesure plutôt que du κ de Cohen (1960). La raison de ce choix réside dans la nature des annotations, dont le nombre n'est pas fixé à l'avance et dont l'estimation de l'accord attendu servant de base de comparaison est difficile, comme noté par Grouin

type entité	précision	rappel	f-mesure
Date	0,9444	0,9754	0.9597
Event	0,625	0,7353	0,6757
Hour	0,9	1,0	0,9474
Location	0,9308	0,9528	0,9416
Organization	0,7576	0,7895	0,7732
Person	0,9439	0,9343	0,9391
global	0,9097	0,9371	0,9232

TABLE 2 – L'accord intra-annotateur pour chaque type et en global.

et al. (2011); Dalianis (2018). Nous avons ainsi pu calculer diverses métriques, données dans le tableau 2. Bien que la qualité globale atteigne 0,9232 de f-mesure, nous notons une variabilité de qualité en fonction du type d'entité. Ainsi, les dates, heures, lieux et personnes ont de très bons résultats, alors que ceux des organisations et des événements sont moindres.

La différence la plus fréquente est celle d'annoter plus d'entités (bruit), cette différence se répartit équitablement entre les lieux, organisations et événements. Les lieux étant plus nombreux, nous pouvons considérer que cette erreur est plus fréquente pour les organisations et les événements. Pour les organisations, ces différences sont principalement sur des mentions nominales et les ministères (« régime syrien », « économie »), pour les événements ces différences concernent surtout les événements politiques ou climatiques (« primaire socialiste », « Camp Fire »). Les erreurs de type, frontière ou silence ont des volumes comparables. Les erreurs de frontière concernent principalement le déterminant « le » (« l'armée syrienne » → « armée syrienne »), ou la nature d'un lieu (« l'autoroute E40 » → « E40 »). Ces erreurs semblent être principalement présentes dans les documents de l'année 2016, première année à avoir été annotée. Ces erreurs pourraient donc être dues à l'absence de mise-au-point d'une mini-référence et/ou d'une phase de rodage (Fort, 2012).

4 Comparaison avec d'autres corpus

Dans cette section, nous comparerons le corpus que nous avons produit avec d'autres corpus similaires, à savoir le FTB, Quaero, le Free-FTB et le corpus MEANTIME (Minard *et al.*, 2016).

Le corpus a une taille comparable au FTB en nombre de caractères. Parmi les types en commun, la principale différence est que nous ne faisons pas la distinction entre organisation et entreprise, différence faite dans le FTB. Une autre différence est que le FTB distingue les personnages fictifs des personnes réelles, distinction non faite ici. Notre schéma d'annotation est plus couvrant que celui du FTB : par exemple le FTB ne couvre pas les lieux géologiques, hydrologiques ou astrologiques.

Le corpus a une taille équivalente à environ un quart du Quaero. Un comparatif avec Quaero a été fait en section 2. La plupart des annotations imbriquées dans le corpus que nous proposons peuvent être remplacées par un composant de type "name" pour correspondre au modèle Quaero. Les autres composants doivent être récupérés manuellement.

Le Free-FTB est un corpus de textes issu d'articles de la partie française de Wikinews et de transcriptions d'Europarl. Ce corpus utilise les articles des années 2005 à 2012 de Wikinews. À l'heure

actuelle, le corpus Free-FTB ne comprend que des informations de segmentation et de POS, basées sur le FTB. Les annotations POS du Free-FTB sont les sorties d'un système par apprentissage entraîné sur le FTB. Le corpus que nous proposons ici, bien qu'ayant un schéma d'annotation différent du FTB, pourrait très bien être utilisé pour enrichir le Free-FTB d'information d'entités nommées.

Le corpus MEANTIME propose également une annotation en entités nommées sur les textes de 120 articles Wikinews et ses traductions en espagnol, italien et néerlandais (pour un total de 480 documents). Chaque corpus fait approximativement un dixième du corpus que nous proposons ici et les langues autres que l'anglais sont des traductions. Le corpus MEANTIME a cependant une annotation bien plus riche : les documents sont notamment annotés en chaines de coréférence des entités (intra- et inter-documents) et les relations entre entités. Le jeu d'annotation en entités nommées de MEANTIME se base sur celui de CoNLL 2003, enrichi d'expressions temporelles, produits et événements. L'une des différences notables entre MEANTIME et notre corpus est que dans MEANTIME les groupes sont également annotés, comme les groupes de personnes (*John Howard + Ian MacDonald*, *500 guests*, etc.) et les groupes d'organisations (*loss-making business*, etc.).

5 Conclusion et perspectives

Nous proposons ici un corpus annoté en entités nommées, libre et disponible. Il est également versionné, les ajouts, erreurs et corrections sont facilement traçables, des corrections peuvent également être proposées simplement. Nous souhaitons à cet effet que ce corpus soit collaboratif, des révisions (modification du corpus, du guide d'annotation, etc.) peuvent être effectuées sous réserve de validation. Étant donnés les points précédents, et tant que Wikinews sera alimenté de nouveaux articles, nous voulons ce corpus évolutif où de nouveaux textes seront ajoutés pour le maintenir à jour.

Il est prévu dans un premier temps de passer en revue de manière approfondie les annotations effectuées sur les années 2016 à 2018. Comme nous l'avons vu, bien que la plupart des types d'entités ont un fort accord intra-annotateur, les événements et les organisations ont un accord comparativement faible et méritent donc une plus grande attention. Nous prévoyons d'intégrer de la désambiguïsation des entités nommées ainsi qu'annoter les relations entre entités nommées, chose qu'il est possible d'effectuer avec l'outil BRAT. Le corpus continuera à être alimenté de nouveaux articles Wikinews afin de le maintenir le plus à jour possible. Lorsque ce dernier aura une taille comparable au corpus Quaero, nous pourrons l'annoter avec le même schéma d'annotation. Des campagnes d'évaluation pourront être effectuées en utilisant ce corpus à l'avenir, tant sur la reconnaissance des entités nommées que sur leur désambiguïsation. Il est également prévu d'enrichir le Free-FTB à l'aide de ce corpus afin d'obtenir un corpus segmenté et annoté en POS et entités nommées libre, accessible et de large volume. Nous pourrons à cette occasion refaire les expériences présentées par Hernandez & Boudin (2013) afin d'estimer si nous avons un apport similaire dans le cas des entités nommées par rapport au POS. Une autre expérience à mener serait d'utiliser notre corpus comme ressource supplémentaire pour des systèmes par apprentissage dans le cadre de tâches existantes.

Remerciements

Les travaux présentés ici bénéficient en partie du soutien financier du projet PARSEME-FR (ANR-14-CERA-0001).

Références

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). Building a treebank for French. In A. ABEILLÉ, Ed., *Treebanks*. Dordrecht : Kluwer.

BOISEN S., CRYSTAL M., SCHWARTZ R. M., STONE R. & WEISCHEDEL R. M. (2000). Annotating resources for information extraction. In *LREC*.

COHEN J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, **20**(1), 37–46.

DALIANIS H. (2018). Evaluation metrics and evaluation. In *Clinical Text Mining*, p. 45–53. Springer.

DODDINGTON G. R., MITCHELL A., PRZYBOCKI M. A., RAMSHAW L. A., STRASSEL S. & WEISCHEDEL R. M. (2004). The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation. In *LREC*, volume 2, p. 1–4.

DUPONT Y. (2017). Exploration de traits pour la reconnaissance d'entités nommées du français par apprentissage automatique. In *24e Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, p. 42–55.

EHRMANN M. (2008). *Les entités nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation*. PhD thesis, Paris 7.

FINKEL J., DINGARE S., NGUYEN H., NISSIM M., MANNING C. & SINCLAIR G. (2004). Exploiting context for biomedical entity recognition : From syntax to the web. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, p. 88–91 : Association for Computational Linguistics.

FORT K. (2012). *Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus*. PhD thesis, Université Paris-Nord-Paris XIII.

FORT K., EHRMANN M. & NAZARENKO A. (2009). Vers une méthodologie d'annotation des entités nommées en corpus ? In *Traitement Automatique des Langues Naturelles 2009*.

GALIBERT O., QUINTARD L., ROSSET S., ZWEIGENBAUM P., NÉDELLEC C., AUBIN S., GILLARD L., RAYSZ J.-P., POIS D., TANNIER X. *et al.* (2010). Named and specific entity detection in varied data : The quæro named entity baseline evaluation. In *LREC*.

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*.

GRISHMAN R. & SUNDHEIM B. (1996). Message Understanding Conference-6 : A Brief History. In *COLING*, volume 96, p. 466–471.

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 *Proceedings of the 5th Linguistic Annotation Workshop*, p. 92–100 : Association for Computational Linguistics.

HAN S., KWON S., YU H. & LEE G. G. (2017). Answer ranking based on named entity types for question answering. In *Proceedings of the 11th International Conference on Ubiquitous Information Management and Communication*, p. 71–74 : ACM.

HERNANDEZ N. & BOUDIN F. (2013). Construction automatique d'un large corpus libre annoté morpho-syntaxiquement en français. In *Traitement Automatique des Langues Naturelles (TALN)*.

KIM J.-D., OHTA T., TATEISI Y. & TSUJII J. (2003). GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, **19**(suppl 1), i180–i182.

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

LAFFERTY J., MCCALLUM A. & PEREIRA F. (2001). Conditional random Fields : Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML 2001*, p. 282–289.

LAMPLE G., BALLESTEROS M., SUBRAMANIAN S., KAWAKAMI K. & DYER C. (2016). Neural architectures for named entity recognition. *arXiv preprint arXiv :1603.01360*.

MINARD A.-L., SPERANZA M., URIZAR R., ALTUNA B., VAN ERP M., SCHOEN A., VAN SON C. *et al.* (2016). Meantime, the newsreader multilingual event and time corpus.

NOUVEL D., ANTOINE J.-Y., FRIBURGER N. & MAUREL D. (2010). An analysis of the performances of the casen named entities recognition system in the ester2 evaluation campaign.

NOUVEL D., EHRMANN M. & ROSSET S. (2015). *Les entités nommées pour le traitement automatique des langues*. Rapport interne, ISTE editions.

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 *Proceedings of the Sixth Linguistic Annotation Workshop*, p. 40–48 : Association for Computational Linguistics.

SAGOT B., RICHARD M. & STERN R. (2012). Annotation référentielle du corpus arboré de paris 7 en entités nommées. In *Traitement Automatique des Langues Naturelles (TALN)*, volume 2.

SALMON-ALT S., BICK E., ROMARY L. & PIERREL J.-M. (2004). La freebank : vers une base libre de corpus annotés. In *Traitement Automatique des Langues Naturelles-TALN'04*, p. 10–p.

SEKINE S. & NOBATA C. (2004). Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy. In *LREC*.

SHIPMAN J. W. (2013). Tkinter 8.4 reference : a gui for python.

STENETORP P., PYYSALO S., TOPIĆ G., OHTA T., ANANIADOU S. & TSUJII J. (2012). Brat : a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, p. 102–107 : Association for Computational Linguistics.

SURDEANU M. (2013). Overview of the TAC2013 Knowledge Base Population Evaluation : English Slot Filling and Temporal Slot Filling. In *Text Analysis Conference*.

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-Volume 4*, p. 142–147 : Association for Computational Linguistics.