Année universitaire 2020-2021

Evaluation d'outils d'annotation en entités nommées pour le français : que compare-t-on ?

Mémoire de Master 1 Langue et Informatique

Présenté par :
Alexane Jouglar

Sous la direction de :
Karën Fort,
Yoann Dupont,
Alice Millour

Remerciements

Je remercie chaleureusement les étudiants du Master 2 Langue&Informatique ayant pré-annoté le corpus qui a servi de base à ce travail. Je remercie également mes encadrants (Karën Fort, Alice Millour et Yoann Dupont) qui m'ont accompagnée pendant cette recherche et ont passé de longues heures avec moi à réaliser l'adjudication des annotations manuelles, ce qui ne fut pas sans peine. J'espère que ce mémoire sera porteur.

Table des matières

Table des figures

Liste des tableaux

Introduction

Le célèbre et très médiatisé Noam Chomsky écrivait dans un essai de linguistique de 1970 :

> « Le langage est un processus de libre création ; ses lois et principes sont fixés, mais la manière dont les principes de génération sont utilisés est libre et varie infiniment. Même l'interprétation et l'usage des unités lexicales induit un processus de libre création. » [2] [Chomsky, 1999]

Le processus d'interprétation invoqué par Chomsky joue un rôle prédominant tant en linguistique classique qu'en traitement automatique des langues. Il n'existe aucune manière d'accéder au système de la langue autrement que par ses manifestations. Or, ces manifestations sont possiblement infinies. De la simple segmentation en unités lexicales d'un texte à la reconnaissance d'entités nommées, il n'est donc pas un seul processus qui ne fasse appel à l'interprétation. Et plus la tâche est complexe, moins consensuelle se fait l'approche. Concernant les entités nommées, on peut sans doute admettre qu'il n'y a d'ailleurs de consensus que sur la notion de leur existence. Lorsqu'il s'agit en revanche de les circonscrire, les approches divergent. Pourtant nombreux sont les outils dont on vante la capacité de reconnaissance d'entités nommées. Parmi ceux-là, la librairie SpaCy [Honnibal et al., 2020], et la cascade de transducteurs CasEN [Maurel et al., 2011]. Nous verrons dans ce mémoire que la reconnaissance des entités nommées pose plusieurs problèmes non encore résolus et probablement impossibles à résoudre tant en l'état des connaissances actuelles qu'à la perspective de connaissances plus avancées. C'est ce qu'évoque l'article « Reconnaissance d'entité nommées - existe-t-il un plafond de verre ? » [3] [Stanislawek et al., 2019] : il y aurait un plafond de verre au sens où il pourrait y avoir un point à partir duquel on ne puisse pas

2. original : "*Language is a process of free creation ; its laws and principles are fixed, but the manner in which the principles of generation are used is free and infinitely varied. Even the interpretation and use of words involves a process of free creation.*"

3. original : "*Named Entity Recognition – Is there a glass ceiling*"

améliorer la performance des outils en reconnaissance des entités nommées. En faisant une analyse détaillée du jeu de données CoNLL 2003 d'un point de vue linguistique, les auteurs en ont tiré entre autres la conclusion suivante :

> « une phrase en elle-même n'est pas toujours suffisante pour reconnaître la classe d'une entité nommée » [4]

Nous nous proposons malgré tout d'exécuter les deux outils précédemment cités, CasEN et SpaCy, afin d'évaluer leurs sorties respectives et d'offrir une réflexion sur la tâche de reconnaissance des entités nommées (REN). Pour cela nous utiliserons un corpus préalablement annoté à la main et qui servira de référence. Notre travail consistera à évaluer dans quelle mesure on peut comparer des outils de REN alors même que, comme nous allons le voir, ils ne réalisent pas exactement la même tâche. Cela nous permettra de soulever une interrogation plus fondamentale : compte-tenu du flou qui entoure la définition des entités d'une part et de la tâche d'annotation en EN d'autre part, que savons-nous réellement de leur performance ? Pour ce faire, nous ferons un bref état de l'art dans le domaine des entités nommées et de leur reconnaissance. Cet état de l'art nous permettra d'évoquer les problèmes présents à la racine même de la notion d'entité nommée et qui rendent impossible une comparaison immédiate. Après cela, nous présenterons les méthodes que nous avons utilisées pour ce mémoire puis les résultats de nos recherches.

4. original : "*a sentence itself is not always sufficient to recognise a class of a NE.*"

Chapitre 1

Etat de l'art

1.1 Précédents dans la recherche sur les entités nommées

1.1.1 Emergence des entités nommées

La notion d'entité nommée naît lors des MUC (*Message Understanding Conference*), une série de conférences qui s'est déroulée aux Etats-Unis entre 1987 et 1998. Elles ont pour but de favoriser le développement de méthodes et outils permettant l'extraction d'information dans des énoncés écrits ou oraux. Organisées par la NRAD, une division du NOSC (*Naval Ocean Systems Center*) puis financées par la DARPA (*Defense Advanced Research Projects Agency*), leur but originel est militaire ; il s'agit de pouvoir repérer, au sein de larges corpus de documents, des informations qui soient d'utilité à l'armée [Grishman and Sundheim, 1996]. Plusieurs tâches de l'extraction d'information y sont abordées. Cet intérêt accru pour l'extraction d'information est liée à la société numérique et au traitement facilité de grandes quantités de données textuelles. On cherche à recueillir sans trop d'efforts des informations utiles ; c'est ce qu'on appelle les *low hanging fruits* [Church, 2011]. Et, de fil en aiguille, la notion d'entités nommées apparaît (lors de la MUC-6), pour ne plus jamais disparaître. On la retrouve lors du MUC-7 et Nancy A. Chinchor [Chinchor and Robinson, 1997] retranscrit la définition qui en a été donnée de la manière suivante :

> « Les entités nommées furent définies comme des noms propres et des quantités d'intérêt. Personnes, organisations, et noms de lieu furent annotés, de même que les dates, les durées, les pour-

centages, et les valeurs monétaires. » [1]

A leur tout début, les entités nommées recouvrent donc les personnes, organisations, lieux, dates, durées, pourcentages et les quantités monétaires. Les recherches réalisées lors de ces MUC dépassent les frontières de l'armée et deviennent une des tâches à part-entière du traitement automatique des langues. L'objectif de la reconnaissance des entités nommées est de pouvoir rendre compte du sens des textes en se focalisant sur les éléments les plus lourds d'information au sein d'un énoncé, d'où les items évoqués par Nancy A. Chinchor et que l'on peut regrouper dans trois grandes catégories : ENAMEX, NUMEX, et TIMEX. La catégorie ENAMEX regroupe toute entité nommée présente sous forme de chaîne de caractère. NUMEX fait référence aux quantités et informations chiffrées ; quant à TIMEX, cela regroupe tout ce qui se rapporte au temps (dates, heures, et indications temporelles). Cela dit, au fil du temps la notion d'entités nommées a été sujette à évolution et, lors des conférences CoNLL (en particulier CoNLL-2002), une définition très restrictive des entités nommées est apparue. Exit TIMEX et NUMEX, les entités nommées ne désignent plus lors de cette conférence que les ENAMEX [Tjong Kim Sang, 2002] soit les personnes, les organisations, les lieux, et tout ce qui est reconnu comme entité sans pour autant pouvoir y affecter d'étiquette (MISC, mis pour *miscellaneus*) :

> « Le type "*miscellaneus*" est utilisé lors des conférences CONLL et inclut les noms propres qui ne relèvent pas des "*enamex*" classiques ». [2] [Nadeau and Sekine, 2007]

Cette définition est reprise lors de la conférence suivante :

> « Les entités nommées sont des expressions qui contiennent des noms de personnes, des organisations et des lieux. » [3]
> [Sang and De Meulder, 2003]

Mais avec la nécessité d'extraire des informations de plus en précises au sein des larges corpus disponibles aujourd'hui, une autre approche des

1. original : "*Named Entities were defined as proper names and quantities of interest. Person, organization, and location names were marked as well as dates, times, percentages, and monetary amounts.*"

2. Original : "*The type "miscellaneous" is used in the CONLL conferences and includes proper names falling outside the classic "enamex".*"

3. " *Named entities are phrases that contain the names of persons, organizations and locations.*"

entités nommées est par la suite apparue : les entités nommées structurées ou étendues. Lorsqu'on parle d'entitées nommées structurées, un annotable n'est plus seulement étiqueté par un type, c'est-à-dire une catégorie de sens large. Dans cette perspective, une étiquette fait apparaître une catégorie de sens large et une catégorie plus restrictive, soit un type et un sous-type. Par exemple en (1), « Paris » est un lieu et plus encore une ville donc l'étiquette pour cette entité ferait apparaître un type « lieu » et un sous-type « ville ».

(1) Je me rends à Paris.

Cette notion est expliquée en 2004 par Sekine [Sekine and Nobata, 2004] mais en 2012, une définition plus précise est donnée [Rosset et al., 2012] (et voir aussi [Grouin et al., 2011]). On ne considère plus seulement la hiérarchie entre les types et sous-types mais également ce qui compose les entités nommées et qu'on appelle en conséquence composants :

> « Nous avons défini une entité nommée étendue comme étant composée de deux sortes d'éléments : types et composants. Dans notre définition, les types se réfèrent à une segmentation générale du monde en de grandes catégories. Plus encore, nous considérons que le contenu d'une entité doit être structuré aussi. Dans cette perspective, nous avons défini un second niveau d'annotation pour chaque catégorie que nous appelons les composants. » [4]

Un exemple de cette structuration est donné dans la figure 1.1.

Diverses approches des entités nommées coexistent donc aujourd'hui, certaines plus restrictives que d'autres.

1.1.2 Reconnaître les entités nommées

Pour reconnaître les entités nommées, on pourrait tout simplement penser à faire des dictionnaires pour chaque catégorie. Au sein de ces dictionnaires, une entrée correspondrait à une entité nommée. Deux problèmes majeurs rendent cette méthode contre-productive. Le premier problème concerne l'actualisation des données. L'actualisation des données serait tout bonnement

4. "*We defined an extended named entity as being composed of two kinds of element : types and components. In our definition, types refer to a general segmentation of the world into major categories. Furthermore, we consider the content of an entity must be structured as well. From this perspective, we defined a second level of annotation for each category we call components.*"

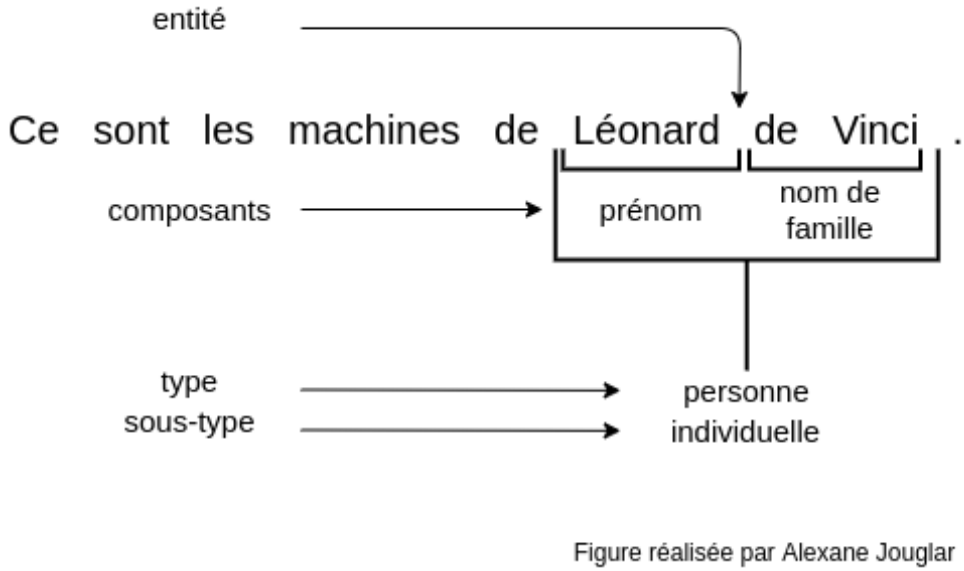

FIGURE 1.1 – Exemple d'une entité nommée structurée.

impossible puisqu'une même entité peut être produite de manière différente selon les locuteurs, que des noms propres apparaissent en continu et que le vocabulaire se renouvelle perpétuellement. Le deuxième problème concerne l'ambiguïté de certaines entités nommées. Dans ce cas, le contexte et/ou le cotexte sont éminemment importants pour reconnaître une entité nommée or un simple dictionnaire ne permettrait pas de les prendre en compte. Nous donnons ci-dessous deux exemples illustrant ce deuxième problème :

(2) a. Le ministre se sent très bien à Matignon.
b. Elle était nommée à Matignon.

Quand considérer que Matignon est un lieu (2-a) et quand considérer qu'il désigne le Premier ministre (2-b) ? Un humain n'aurait pas de problème à faire la distinction mais la machine, contrairement à un humain, n'a pas cette conaissance du monde. Il lui faut quelque chose de concret sur lequel s'appuyer. Lors des balbutiements de la REN, seuls les systèmes à base de règles existaient. Mais lorsque le *machine learning* a fait son apparition, il a également été utilisé pour la REN.

En 1991, Lisa F. Rau décrit un système à base de règles permettant d' « extraire les noms d'entreprise » [Rau, 1991]. Basé sur la transduction, ce système détecte premièrement les suffixes assignés aux différents types d'entreprise (*Inc*, *Corp*, *Co*, etc.) ; puis plusieurs étapes s'ensuivent avec différents

chemins à emprunter selon que l'état de chaque étape est observé ou non.

Les approches symboliques, c'est-à-dire celles qui fonctionnent à base de règles, sont celles qui ont l'air de donner les meilleurs résultats. C'est en tout cas ce que révèle la campagne Ester 2. Mais cette campagne révèle aussi que les résultats dépendent grandement de la nature du document [Phase, 2005].

L'apprentissage machine regroupe trois approches principales. Une première approche appelée apprentissage non-supervisé comme cela a été fait par Michael Collins et Yoram Singer en 1999 [Collins and Singer, 1999]. Une deuxième approche appelée apprentissage faiblement supervisé (comme réalisé par Edouard Grave [Grave, 2014]). Une troisième approche appelée apprentissage actif (*active learning*) qui a été réalisé par des étudiants de l'université de Singapour et de l'Institute for Infocomm Technology de Singapour [Shen et al., 2004]). Depuis CoNLL-2002, la méthode privilégiée est l'apprentissage supervisé.

1.2 Problèmes épistémologiques posés par les entités nommées

Selon Maud Ehrmann :

> « le traitement des entités nommées s'articule en deux processus : identification ou reconnaissance de ces unités dans les textes tout d'abord, catégorisation ou typage selon des catégories sémantiques larges prédéfinies ensuite. » [Ehrmann, 2008]

Comme nous l'avons vu précédemment, la notion d'entité nommée est peu précise et peut recouvrir différentes expressions. Dans cette perspective, il est difficile d'avoir une notion absolue d'entité nommée car ce qu'on va reconnaître dépend aussi de ce qu'on va en faire après.

Certaines parties du discours ne sont en elles-mêmes des entités nommées pour personne : nous pouvons dire dans ce cas qu'il y a consensus sur certaines parties du discours qui ne sont pas des entités nommées. Les mots grammaticaux ne sont pas des entités nommées mais ils peuvent apparaître au sein de l'une d'elles comme en (3) (tout dépend de ce que prescrit le guide d'annotation utilisé et c'est par exemple le choix fait par le guide Quaero [5]

5. Le lien ci-après donne accès au guide Quaero : `http://www.quaero.org/media/`

dont nous reparlerons un peu plus tard).

(3) « Nous nous retrouvons à 15h ».
   a. L'entité peut être «15h».
   b. L'entité peut être aussi «à 15h».

Les verbes conjugués n'en sont pas non plus, de la même manière que ne le sont pas les couleurs, etc. Le problème survient toutefois bien au sein des mots lexicaux. J'en fournis un exemple et l'atteste :

(4) Salut Pierre ! Toujours d'accord pour que l'on se voit à 14h au Musée du Louvre ?

Dans l'exemple (4), il y a plusieurs informations essentielles ; nous dirons : « Pierre », « 14h », et « Musée du Louvre ». Nous pourrions donner à ces éléments des types, respectivement : nom, temps, et lieu. On pourrait même donner des sous-types : prénom, heure, et bâtiment. Les autres mots de l'exemple ne sont pas vides de sens, mais ils ne contiennent pas l'information principale de l'énoncé. On pourrait presque les enlever, ce qui produirait l'énoncé suivant : « Pierre, 14h, Musée du Louvre ». Si l'on s'adresse ainsi à notre ami Pierre, il le prendra sûrement comme un rappel un peu autoritaire mais il comprendra ce que l'on a voulu dire. Ces informations sont très importantes parce qu'elles codent l'essence de notre message et elles ont une réalité extra-linguistique. Le reste sert à enrober l'information ultime, éventuellement à la préciser voire à la désambiguïser. L'information essentielle réside donc dans le cas des entités nommées dans des morceaux de discours plus ou moins longs ayant une réalité extra-linguistique : il y a bien un humain dans notre monde que j'appelle Pierre, une heure qui existe au sein d'un système créé par la communauté humaine pour se repérer dans le temps, et un édifice appelé Musée du Louvre qui se trouve au sein de la ville de Paris, en France. Cela serait la même chose si l'on prenait un exemple comme celui ci-après :

(5) Tiens, je te donne mon numéro : 01 00 00 00 00, n'hésite pas à m'appeler. Je serai devant le Panthéon.

Dans l'exemple (5) nous pouvons relever entre autres un numéro de téléphone. Selon le guide Quaero, cela est considéré comme une entité nommée. Nous pouvons admettre que cela est cohérent : un numéro de téléphone code une information essentielle du discours ayant une réalité extra-linguistique, à savoir qu'il y a bien en ce monde un objet dont le signifiant est *téléphone* et

`files/bibliographie/quaero-guide-annotation-2011.pdf`

que l'on peut joindre physiquement en tapant sur un clavier numérique les chiffres indiqués.

Mais tous les guides ne considèrent pas les numéros de téléphone comme une entité nommée puisque tout dépend de la définition d'entité nommée utilisée et en conséquence de son étendue.

Voici une définition de ce qu'est une entité nommée :

> « Les "mentions d'entités nommées" sont des unités textuelles qui renvoient à des "entités" du domaine mais qui peuvent relever de différentes catégories linguistiques : noms propres ("Air France"), pronoms ("elle"), et plus largement descriptions définies ("cette compagnie", "la principale compagnie aérienne française") ». [Omrane et al., 2010].

Le problème posé par ce genre de définition est qu'il s'agit d'une définition en extension (on énumère ce qui compose la catégorie entité nommée sans pourtant dire ce que ces choses ont en commun), que cette extension n'est pas suffisamment précise pour savoir ce que l'on y range exactement, et que par ailleurs ce genre de définition reprend le mot « entité » pour définir la notion même d'entité nommée.

Ailleurs dans la thèse d'Ehrmann, la définition proposée d'entité nommée est :

> « Etant 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. »

Autrement dit, une entité nommée désigne une information du discours qui code un observable ; cet observable doit être spécifique. Nous pourrons ajouter que l'information dont il est question doit être essentielle.

La notion de spécificité dont il est question dans les entités nommées n'est pas propre aux objets et s'étend à l'ensemble du vocabulaire. Prenons un exemple :

(6) Un humain est un individu appartenant à une espèce de primates nommée homo sapiens.

Si l'on parle d'humains, ou bien si l'on prend un énoncé scientifique comme en (6), on a là le singulier mais on ne fait pas référence à Pierre ou à Marie dans leur individualité, on fait référence au groupe humain en tant qu'espèce,

il n'y a donc là aucune forme de spécificité. Maintenant, si chacun s'accorde pour dire que « mouchoir » ou « les humains » ne rentrent pas dans le champ des entités nommées, tout le monde ne s'accorde pas sur le fait de savoir ce qui rentre dans le champs des entités nommées. En effet, il y a véritablement une notion d'essentialité de l'information lorsque l'on parle d'entités nommées ; or qu'est-ce qu'une information essentielle, qu'est-ce qui ne l'est pas ? La source du problème se trouve bien ici. En effet, si l'on admet que les fonctions font partie des entités nommées (et nous nous référons là au guide Quaero qui invite à annoter *func.ind* ou *func.col* pour respectivement fonction individuelle - comme « ministre » - et fonction collective – comme « pompiers »), on n'aura aucun mal à dire que « Président de la République », « ministre », ou « Directeur des opérations » sont des fonctions mais la limite devient très floue lorsque l'on prend des exemples tels que « boulanger », « livreur », « secrétaire », « consultant », « ingénieur ». Sont-ce là des fonctions ? Si l'on admet que ce n'en sont pas, il semble que l'on passe à côté de quelque chose. En effet, nous pouvons sans aucun doute établir la différence pratique entre un Président de la République et un boulanger (le premier a une fonction publique et officielle, il est élu par la population ; le deuxième non). Mais pourrait-on dire la différence entre un directeur des opérations et un boulanger ? Devrait-on là définir un degré d'importance entre les fonctions ? En procédant ainsi, nous faisons intervenir une très grande part d'interprétation, peut-être même une trop grande part d'interprétation. Admettons donc que « boulanger », « livreur » et « secrétaire » soient des fonctions. Il convient alors de savoir ce que l'on fait du danseur, du peintre, ou du chanteur. Si nous parlons là de domaines artistiques, certains en font pourtant leur métier. Prenons un exemple :

(7) a. Il est danseur au bolchoï.
b. Il est danseur à ses heures perdues.

Dans l'exemple (7-a), on pourrait admettre que « danseur » est une fonction. Mais qu'en est-il de l'exemple (7-b) ? Si l'on fait bien référence à une activité extra-curriculaire, il ne s'agit pas d'une fonction. Nous remarquerons ici la perméabilité des catégories. Et concernant les entités pour lesquelles il n'y aurait aucun doute, reprenons l'exemple (4) : dans cet exemple, on peut annoter « 14h » ou « à 14h » (difficulté des mots grammaticaux évoqué précédemment) ; la délimitation pose donc également un problème. Évoquons aussi la question des étiquettes multiples : une seule entité peut-être à la fois deux ou plusieurs choses. Par exemple « Musée du Louvre » est à la fois une organisation et un lieu. Doit-on mettre les deux annotations à chaque fois que l'on rencontre cette entité ou simplement celle qui correspond dans le

contexte en question ? Ci-dessous, en (8-a), il est question du lieu ; en (8-b), il est question de l'organisation.

(8) a. On se voit au Musée du Louvre.
b. Le Musée du Louvre a racheté une peinture vieille de mille ans.

Les bases de notre travail sont ainsi posées. Les différents problèmes liés à la reconnaissance des entités nommées font encore aujourd'hui l'objet de recherches et de diverses conférences. Il n'existe pas de consensus au sein de la communauté scientifique sur ce que recouvre la notion d'entité nommée. C'est ce que doit entre autres prendre en compte la comparaison d'outils qui n'ont pas nécessairement les mêmes paramètres de recherche. Nous tenterons dans les pages à venir d'apporter notre pierre au vaste édifice que constitue la tâche de reconnaissance des entités nommées.

Chapitre 2

Méthode

Le travail décrit dans ce chapitre se divise globalement en trois grands mouvements. Dans un premier temps, nous détaillons la création du corpus. Dans un deuxième temps, nous parlons du processus d'annotation manuelle. Suite à cela, nous parlons des annotations automatiques avec SpaCy [Honnibal et al., 2020] et CasEN [Maurel et al., 2011]. Enfin, nous en profitons pour faire quelques remarques sur les annotations et les problèmes pratiques posés par ce travail ainsi que pour faire une proposition d'unification des annotations.

2.1 Corpus

Le corpus pour ce mémoire a été élaboré par Karën Fort et Yoann Dupont et est composé de sept textes comprenant chacun environ 1 000 tokens. Il est constitué dans la lignée du *brown corpus* [Francis and Kucera, 1979], le premier corpus connu complet rassemblant une grande variété de genres de textes existant en anglais. Publié pour la première fois en 1961 à l'Université de Brown, il est divisé en 500 échantillons et la taille de ces échantillons est d'environ 2 000 tokens chacun.

La langue du jeu de données est le français.

Afin que le corpus soit complet, il a fallu recourir à des textes d'origine et de genre divers. Le corpus Sequoia [Candito and Seddah, 2012] [1] comporte des phrases de quatre origines : Europarl français, le journal l'Est Républicain, Wikipédia Fr et des documents de l'Agence Européenne du Médicament. Au total, cela fait 3 204 phrases et 69 246 tokens. Il est de genre documen-

1. `https://universaldependencies.org/#download`

taire. Les corpus Aij-wikiner [Nothman et al., 2013] [2] et GSD (pour *Google Stanford Dependencies*) [McDonald et al., 2013] [3] sont de genre documentaire également. Publiés par Universal Dependencies (UD)[Nivre et al., 2016], Aij-wikiner et GSD regroupent tous deux des phrases issues de différents articles Wikipedia. Le corpus Spoken [Gerdes and Kahane, 2017] [4] comporte, comme son nom l'indique, des énoncés oraux. Egalement diffusé par Universal Dependencies, c'est un corpus de genre dialogique. Le corpus APIL [5] comporte des informations de l'ordre du tourisme (programme culturel) : c'est un corpus de genre informatif. Wikinews [6] n'est pas un corpus mais le nom donné à un document reprenant des pages Wikinews. Enfin, des phrases extraites du *Ventre de Paris* [7], d'Emile Zola, ont été utilisées. Elles constituent la partie narrative de notre corpus. Le document est nommé pg6470. Le *parsing* ayant ajouté une phrase en anglais dans pg6470, nous l'avons remplacée à la main par une phrase en français de même taille (également extraite du roman bien entendu). Nous utilisons des ressources libres pour pouvoir utiliser seule une partie des textes ou corpus.

Document	Année de publication	Source	Licence
Aij-wikiner	2017	UD	CC-BY 4.0
GSD	2015	UD	CC-BY-SA 4.0
pg6470	1873	Gutenberg	Domaine public
Sequoia	2017	UD	CC-BY-NC-SA 4.0
Spoken	2018	UD	CC-BY-SA 4.0
Wikinews (extraits)	2018	Wikinews	CC-BY 2.5
APIL	2011	Groupe d'Action Locale Othe-Armance	Licence LGPLLR

TABLE 2.1 – Informations sur les documents utilisés pour le corpus.

2. `https://figshare.com/articles/dataset/Learning_multilingual_named_entity_recognition_from_Wikipedia/5462500`
3. `https://universaldependencies.org/#download`
4. `https://universaldependencies.org/#download`
5. Lien de téléchargement non disponible
6. `https://fr.wikinews.org/wiki/Accueil`
7. `http://www.gutenberg.org/ebooks/6470`
7. `https://spacy.io/models/fr`

Document	Type	Nombre de tokens	Nombre de phrases
Aij-wikiner	didactique	859	37
GSD	didactique	839	34
pg6470	narratif	821	51
Sequoia	informatif	818	35
Spoken	dialogique	1 003	1
Wikinews (extraits)	informatif	854	45
APIL	informatif	620	45

TABLE 2.2 – Caractéristiques des documents du corpus.

Le corpus fera l'objet d'un dépot sur la plateforme Ortolang (Outils et Ressources pour un Traitement Optimisé de la LANGue ; la plateforme Ortolang se veut réservoir de données langagières ainsi que fournisseuse d'outils d'analyse linguistique [8]).

2.2 Annotations manuelles

2.2.1 Guide et outil

Guide Quaero

Les annotations manuelles ont été réalisées sur la base du guide Quaero [Le Pevedic and Maurel, 2016] afin d'obtenir des annotations aussi précises que possible. Le détail de l'annotation manuelle est importante. Du détail d'une annotation dépend sa réutilisation. En effet, une annotation très détaillée sera plus facilement adaptable à divers outils annotant de manières différentes. Par exemple si les annotations manuelles réalisées pour ce travail ne contenaient pas d'étiquettes TIMEX et NUMEX, il n'aurait pas été possible de juger de la qualité des annotations CasEN qui est un outil utilisant un jeu d'étiquettes relevant entre autres des TIMEX et NUMEX.

Nous en parlions dans l'introduction, il n'est possible d'accéder au système de la langue que par ses manifestations. Par conséquence, aucun guide d'annotation ne saurait être complet. Nous relevons ainsi quelques ambiguïtés que nous avons pu trouvées dans le guide Quaero :

8. Le site dédié au projet Ortoloang : `www.ortolang.fr`.

— les événements climatiques sont laissés à la discretion de l'annotateur ; par exemple, pour le texte 3, « El Niño » peut être interprété comme un *event* ou bien comme *loc.phys.hydro*,
— aucune indication n'est donnée pour les entités du type « la gauche » (en tant que parti politique) ou « la droite » (ces occurrences apparaissent dans notre corpus),
— certaines occurrences peuvent avoir deux annotations différentes sans que l'une ne soit supérieure à l'autre dans la hiérarchie. Par exemple, « 8 assaillants » est tantôt annoté comme *amount* tantôt comme *pers.coll.*

WebAnno

La campagne d'annotation a été réalisée sur la plateforme WebAnno [Yimam et al., 2013] [9] :

> « un outil d'annotation polyvalent en ligne permettant un large choix d'annotations linguistiques » [10] [De Castilho et al., 2016] et [Yimam et al., 2013].

La figure 2.1 présente l'accueil de la plateforme WebAnno et les onglets disponibles pour un annotateur donné. Nous y trouverons notamment un onglet *Annotation* pour réaliser les annotations et un onglet *Curation* pour réaliser l'adjudication. Nous reviendrons sur ces deux éléments un peu plus tard.

L'outil WebAnno intègre le moteur d'annotation d'un autre outil en ligne. Cet outil permet également l'annotation participative et utilise le *machine learning* pour proposer des choix d'annotation et réduire le temps du processus. Cet outil se nomme Brat [Stenetorp et al., 2012].

Plusieurs genres d'annotations sont disponibles sur la plateforme ; celles utilisées pour ce travail sont les annotations Quaero. Ce genre d'annotation est basé sur le guide Quaero évoqué précédemment.

9. `https://webanno.github.io/webanno/`
10. "*a general purpose web-based annotation tool for a wide range of linguistic annotations*"

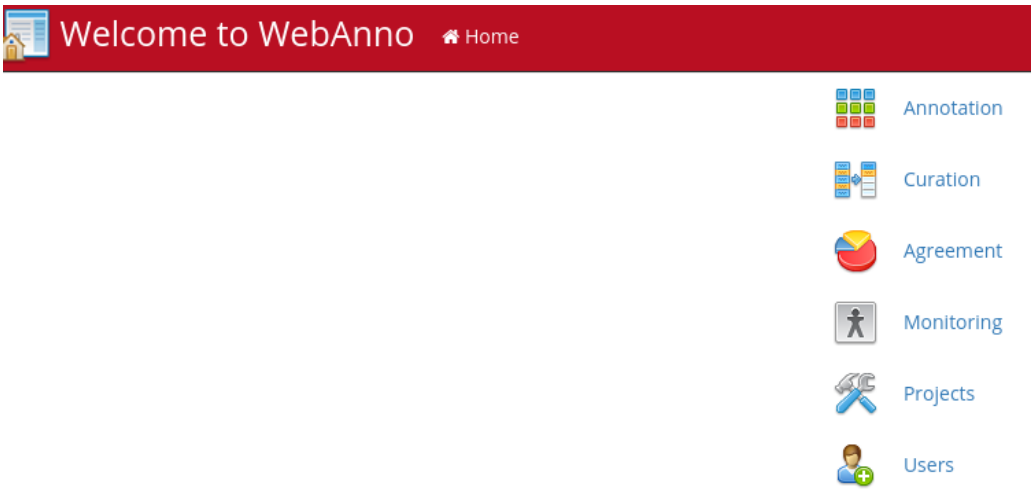

FIGURE 2.1 – La plateforme WebAnno.

2.2.2 Méthodologie

Si l'on suit à la lettre le guide Quaero, doivent figurer dans une étiquette le type et le sous-type, ainsi que chaque composant de l'entité reconnue. Il a été fait le choix de ne faire figurer que les types et sous-types. Prenons un exemple issu du document *Aijwikiner* :

(1) John Foxx

L'étiquette assignée à cette entitée doit faire figurer la catégorie *pers* et la sous-catégorie *ind* (*pers.ind*) car il s'agit d'une personne seule et non d'un groupe de musique par exemple. En revanche, les composants *name.last* et *name.first* n'apparaissent pas. Le guide Quaero s'appuie sur la définition des entités nommées donnée lors de la campagne d'annotation de 2008 appelée Ester-2 [Le Pevedic and Maurel, 2016], ce qui comprend :

— la définition des entités nommées de la thèse de Maud Ehrmann redonnée dans le premier chapitre de ce mémoire.
— la définition des entités temporelles donnée dans le projet Time ML [Pustejovsky et al., 2003]
— la taxinomie de la thèse de Mickaël Tran [Tran and Maurel, 2006]
— la hiérarchie des entités nommées étendues de Satoshi Sekine également évoquée dans le premier chapitre.

Le corpus a été annoté dans un premier temps par les étudiants du M2

Langue et Informatique de Sorbonne Université, promotion 2020-2021, lors du premier semestre universitaire et dans le cadre du cours de Karën Fort sur l'annotation collaborative de corpus. Chaque texte a été annoté par un binôme distinct ou trinôme formé au moins d'un francophone natif et d'un étudiant international. Dans un deuxième temps, nous avons nous-mêmes annoté les textes et cela a pris environ six heures de temps en tout et a été réalisé en trois étapes : lecture des textes et des annotations déjà faites, étude du guide Quaero, annotation de chacun des textes du corpus (cette étape, la plus longue, a duré quatre heures). Toutes ces annotations manuelles ont été réalisées sur l'outil WebAnno précédemment évoqué.

Afin d'avoir un aperçu de la manière dont on annote un texte en WebAnno, la figure 2.2 présente la fenêtre d'annotation qui s'ouvre pour un annotateur donné lorsque l'on clique sur l'onglet approprié.

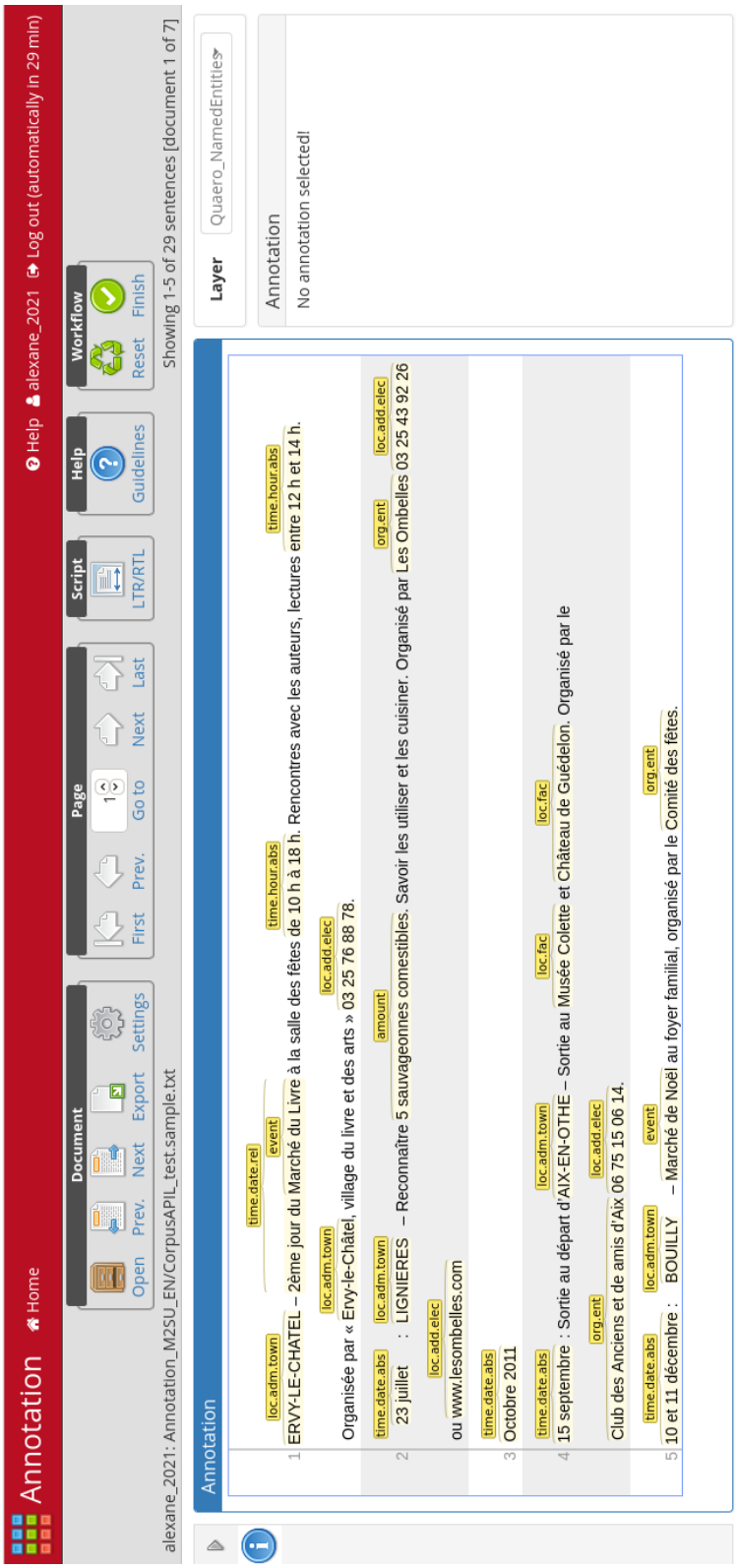

FIGURE 2.2 – L'onglet annotation de la plateforme WebAnno avec le corpus APIL.

Puis il a fallu faire l'adjudication. Nous définissons l'adjudication comme la mise en commun de toutes les annotations individuelles pour établir une référence sur laquelle nous appuyer pour évaluer les outils. Le processus d'adjudication a mis un peu plus de sept heures de temps et a été réalisé par les trois encadrants de ce mémoire - Alice Millour, Yoann Dupont et Karën Fort - et moi-même. La figure 2.3 présente l'onglet d'adjudication tel que peut le voir un annotateur du projet.

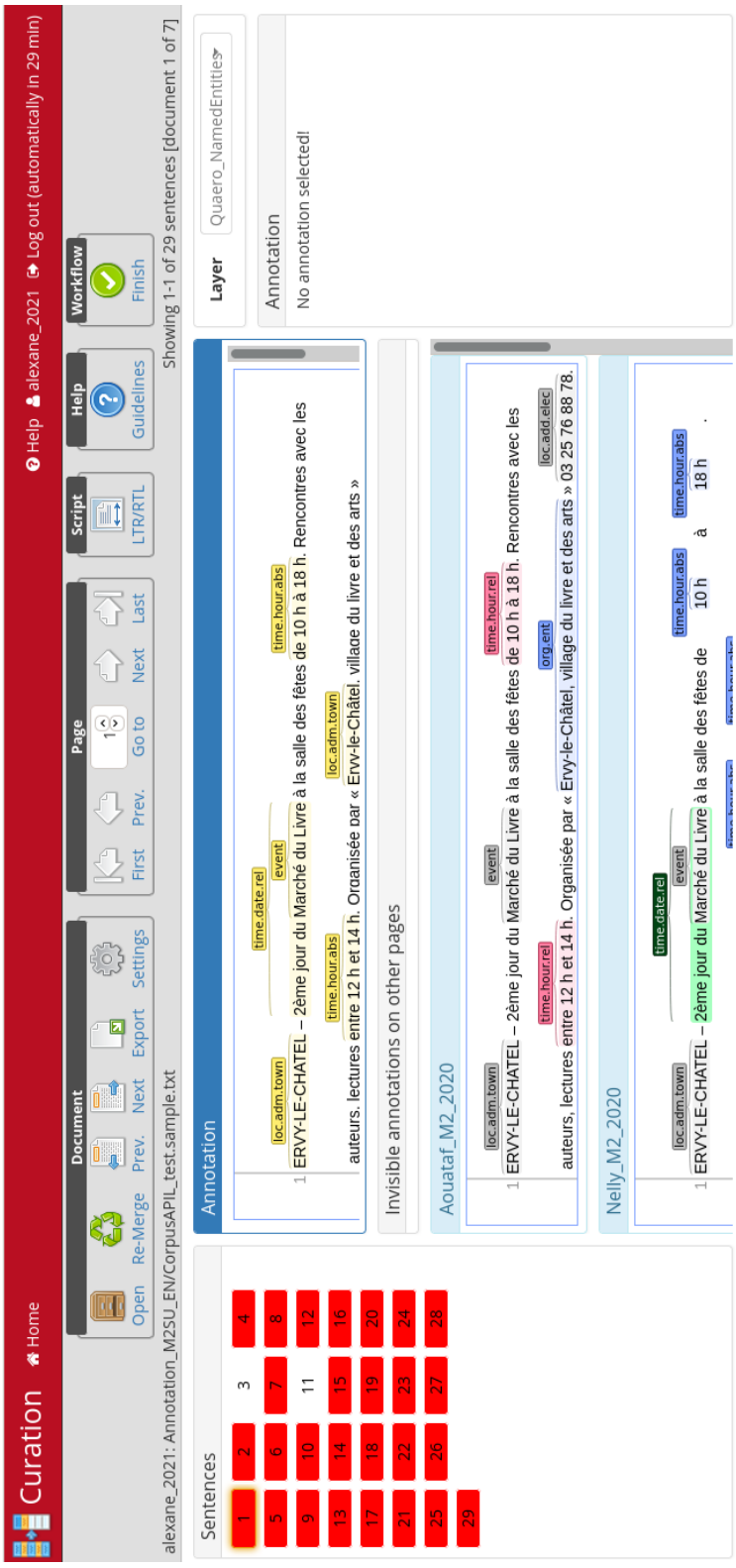

FIGURE 2.3 – Extrait de l'adjudication du corpus APIL sur la plateforme WebAnno.

Nous donnons dans la table 2.3 l'accord inter-annotateur pour chaque paire d'annotateur ayant travaillée sur le corpus. L'accord utilisé est l'alpha de Krippendorf et non le kappa de Cohen. En effet, ce dernier est davantage adapté à une annotation mot par mot, or dans l'exercice que nous réalisons seuls certains pans de phrases de longueurs variables sont étudiés.

La formule pour l'alpha de Krippendorff est donnée par l'expression suivante dans laquelle D_O est le désaccord observé et D_e le désaccord attendu [Artstein and Poesio, 2008] :

$$K_w = 1 - \frac{D_o}{D_e}$$

Les alphas de Krippendorff apparaissant dans le tableau 2.3 sont calculés par WebAnno et nous n'avons pas accès au détail du calcul. Le guide WebAnno nous explique simplement comment visualiser ces accords sur la plateforme [11].

	K	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	Ale
K								0,79							0,81
A1									0,91						0,90
A2					0,64		0,72							0,69	0,77
A3															0,91
A4															0,84
A5					[12]										
A6														0,77	0,86
A7															0,82
A8															0,92
A9												0,80			0,88
A10													0,38		0,52
A11															0,83
A12															0,93
A13															0,83

TABLE 2.3 – Accords inter-annotateurs (alpha de Krippendorff).

Quelques remarques sont de mise. Les annotateurs K et Ale ont un statut un peu particulier. Les deux sont à la fois annotateurs et adjudicateurs et K s'est également chargé de gérer la campagne d'annotation selon la typologie

11. `http://webanno.grew.fr/doc/user-guide.html#sect_monitoring_agreement`

12. Les cases grisées auraient dû faire apparaître un accord inter-annotateur dans l'une ou l'autre des cases mais il n'y a pas suffisamment d'annotation en commun entre A5 et A3 pour qu'il soit calculé par le programme.

des rôles donnés par [Fort, 2012]. Nous pouvons remarquer que les accords inter-annotateurs sont plutôt élevés (il est considéré qu'un alpha de Krippendorf supérant 0,8 marque un accord élevé), mis-à-part pour les couples incluant les annotateurs A2 et A10. Pour ces deux annotateurs, le français n'est pas la langue maternelle et ils ne le maîtrisennt pas aussi bien que les autres étudiants pour qui ce ne serait pas non plus la langue maternelle.

Les annotations manuelles ont été téléchargées depuis WebAnno en format texte. La figure 2.4 montre ce qu'on obtient après ce téléchargement. Nous avons à faire à un fichier tabulaire dont les différents champs sont :

— numéro de la phrase - numéro du fragment de la phrase
— où commence le token - où termine le token
— token en question
— annotation[numéro de l'annotation] (ou bien rien)

```
1-1 0-4 Deux    amount[1]
1-2 5-10    temps   amount[1]
1-3 11-22   structurent
1-4 23-24   l
1-5 24-25   '
1-6 26-31   œuvre
1-7 31-32   ,
1-8 33-36   les
1-9 37-46   cinquante   amount[2]
1-10    47-55   premiers    amount[2]
1-11    56-62   livres  amount[2]
```

FIGURE 2.4 – Annotations manuelles sur WebAnno_exemple de sortie.

Voici deux exemples de ce que nous avons cherché à obtenir et que nous avons obtenu après coup :

(2) Deux temps | amount

(3) cinquante premiers livres | amount

La table 2.4 montre les annotations que l'on a obtenu manuellement et sous quelle méta-étiquette elles ont été regroupées pour la comparaison (ces méta-étiquettes ont été définies pour les besoins de ce mémoire ; pour plus de précision les étiquettes font l'objet d'une partie ultérieure : dans cette partie, le détail des regroupements des donné). On a 31 étiquettes au total et 636 entités annotées.

Méta-étiquette	Etiquette(s)
LOC	loc.phys.geo, loc.adm.reg, loc.phys.hydro, loc.add.phys, loc.oro, loc.fac, loc.add.elec, loc.other, loc.adm.nat, loc.adm.town, loc.adm.sup, loc.phys.astro
PROD	prod.object, prod.art, prod.rule, prod.media, prod.soft, prod.award
TIME	time.date.rel, time.hour.rel, time.date.abs, time.hour.abs
FUNC	func.ind, func.coll
EVENT	event
AMOUNT	amount
PER	pers.ind, pers.coll
ORG	org.adm, org.ent

TABLE 2.4 – Liste des types et sous-types issus de l'annotation manuelle ainsi que les méta-étiquettes les regroupant par type.

2.3 Annotations automatiques

2.3.1 Outils

Les annotations automatiques ont été faites avec CasEN et SpaCy.

CasEN

CasEN est un outil d'extraction d'entités nommées sous licences LGPL-LR. Il fonctionne à base de règles utilisant CaSys [Friburger and Maurel, 2004]. CaSys est programme de création de cascades de transducteurs à états finis disponible sur Unitex. CasEN peut donc être utilisé *via* l'interface Unitex ou *via* scripts. La version de CasEN utilisée respecte la norme TEI ; cette cascade a été créée dans le cadre des projets Ortolang et Istex (mise à disposition de grandes quantités de ressources langagières [13]). S'il a existé une cascade CasEN respectant la norme Quaero, elle n'est aujourd'hui plus maintenue.

Le fonctionnement de la cascade CasEN :

> « La reconnaissance des entités nommées par la cascade CasEN utilise des ressources lexicales et des descriptions locales de motifs, des transducteurs qui agissent sur le texte par des insertions, remplacement ou suppressions. [...] Le principe d'une cascade est de pouvoir utiliser dans les descriptions suivantes les motifs déjà

13. Le site du projet : `https://www.istex.fr/`.

> détectés ou, au contraire, d'éviter un étiquetage non souhaité pour un motif déjà reconnu. L'ordre de passage de ces transducteurs est donc un paramètre important. » [Maurel et al., 2011]

Les annotations CasEN, puisqu'elles respectent la norme TEI, sont un mélange entre deux manières d'annoter. Conformément à ce qui fut indiqué lors des MUC, elles se répartissent en trois catégories : ENAMEX, TIMEX, NUMEX. Mais davantage d'annotations apparaissent et celles-ci correspondent davantage à ce qui a été décidé lors de la conférence IREX au Japon : toutes les catégories énumérées lors du MUC7 ainsi qu'une catégorie *artefact* pour annoter les productions humaines et les prix. En revanche, elle ne reprend pas l'étiquette *optional* décidée lors de cette conférence.

L'outil CasEN est un outil relativement complexe à prendre en main. Cette prise en main a duré une après-midi et a nécessité l'intervention de l'un des encadrants. Les performances de l'outil sont données dans le tableau 2.5 [Maurel et al., 2011].

Mesure	**Entités reconnues**	**Entités bien typées**	**Entités bien balisées**
Précision	94,0	88,4	87,5
Rappel	97,8	92,0	91,1
F-mesure	95,9	90,2	89,3

TABLE 2.5 – Performances de CasEN sur le corpus Eslo 1. Ces résultats sont issus de l'article « Cascades de transducteurs autour de la reconnaissance des entités nommées » [Maurel et al., 2011].

Les annotations CasEN sortent sous forme insérée et au format xml. C'est-à-dire qu'elles sont directement disponibles dans le texte original et entre balises xml. L'exemple (4) donne un aperçu de sortie de l'outil tel qu'obtenu avec le corpus utilisé pour ce travail.

(4) <persName><foreName>Claude </forename><surname>Nougaro </surname></persName>.

Parmi les annotations obtenues en CasEN, on supprime : *address-number* (parce qu'il n'y a en tout que trois annotations sur le corpus et ce sont des erreurs), *offset* qui ne regroupe que quelques mots (« début », « fin », « nord de » et « ouest de ») qui n'ont pas fait l'objet d'une annotation manuelle et donc qui induirait une inexactitude non justifiée, *demonym* (une seule annotation dans le corpus : « américaine ») et *nationality* (« françaises » seul).

SpaCy

SpaCy est sous licence LGPL-LR. C'est un outil qui fonctionne sur base de *machine learning* et qui est entrainé sur deux corpus : Sequoia et Wikiner. SpaCy n'est pas un outil spécialement réalisé pour la reconnaissance d'entités nommées. Il sert entre autres à la tokenization ou au *POS-tagging*. Pour la tâche de reconnaissance d'entités nommées, des modèles sont entraînés sur des corpus plus ou moins larges et vont donner, a priori, des annotations plus ou moins nombreuses et de plus ou moins bonne qualité ; ces modèles sont : *sm* (pour *small*), *md* (pour *middle*), *lg* (pour *large*). Les temps d'exécution varient respectivement du plus court au plus long et il existe une véritable différence de traitement de l'un à l'autre allant respectivement du moins performant au plus performant. Les créateurs indiquent, concernant la reconnaissance d'entités nommées, que SpaCy sert à :

> « étiqueter des objets du monde réel, comme les personnes, les organisations et les lieux » [14]

Les quatre étiquettes disponibles correspondent à quatre annotations réalisables avec SpaCy : LOC, PER, ORG, et MISC. Les constructeurs indiquent les performances de leur outil pour cette tâche. Nous les retranscrivons dans le tableau 2.6.

Mesure	**Score**
Précision	0,85
Rappel	0,84
F-mesure	0,85

TABLE 2.6 – Performances de SpaCy en reconnaissance d'entités nommées d'après la fiche informative disponible sur le site de l'outil [16].

Enfin, nous pouvons noter que l'installation de SpaCy a été très rapide, son utilisation aussi. C'est un outil très documenté et facile à prendre en main. Pour ce travail, la prise en main a duré une après-midi.

Les annotations SpaCy sortent sous forme déportée, c'est-à-dire qu'elles sont extraites de leur contexte (on ne voit que l'entité nommée et son étiquette).

14. `https://spacy.io/usage/spacy-101`
15. `https://spacy.io/models/fr`
16. `https://spacy.io/models/fr`

La mise en forme des données ainsi que les comparaisons ont été menées en Python.

2.3.2 Etiquettes

Les étiquettes manuelles et les étiquettes issues de CasEN ont fait l'objet d'une réorganisation. Les étiquettes appartenant à une même notion on été regroupées sous des méta-étiquettes. Ces méta-étiquettes permettent de comparer plus facilement les deux outils. Le choix fait pour ce mémoire est d'adopter deux schémas d'étiquettes différents : un schéma pour les étiquettes CasEN qui comprend huit méta-étiquettes et un schéma pour les étiquettes SpaCy qui comprend quatre méta-étiquettes (identiques aux étiquettes issues de SpaCy). Les figures 2.6 et 2.5 shématisent les regroupements effectués. Il a été fait le choix de la simplification au détriment de complexification pour deux raisons. La première raison est qu'en précisant les étiquettes issues d'un outil que l'on veut évaluer, on fausse les sorties de l'outil. La deuxième raison réside dans la difficulté de l'exercice. C'est d'ailleurs ce qu'évoque Cyril Grouin dans son article « Simplification de schémas d'annotation : un aller sans retour ? » [Grouin, 2018] :

> « Les règles de conversion ne suffisent pas pour retrouver le niveau de détail du schéma d'annotation d'origine. »

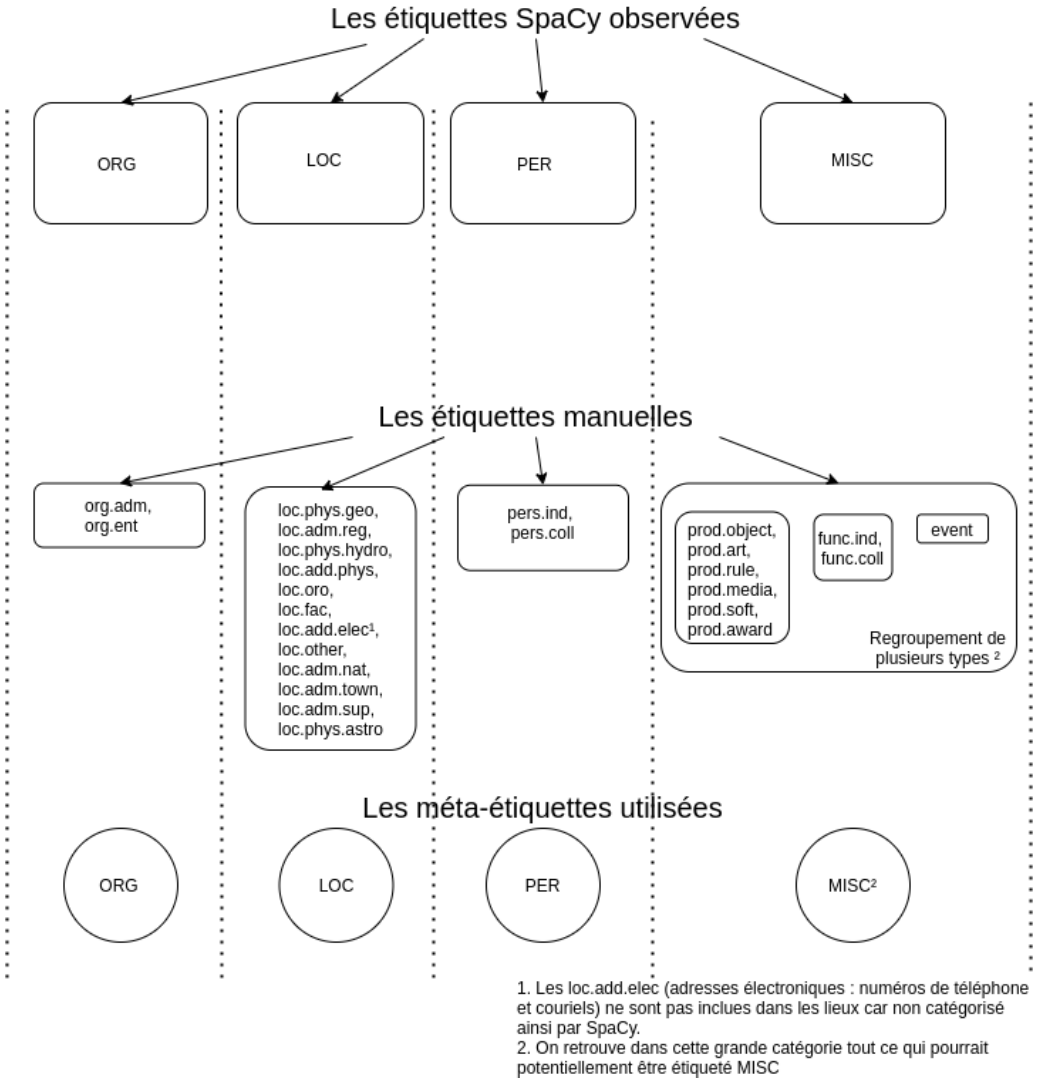

FIGURE 2.5 – Regroupements des étiquettes SpaCy et manuelles effectués pour ce mémoire.

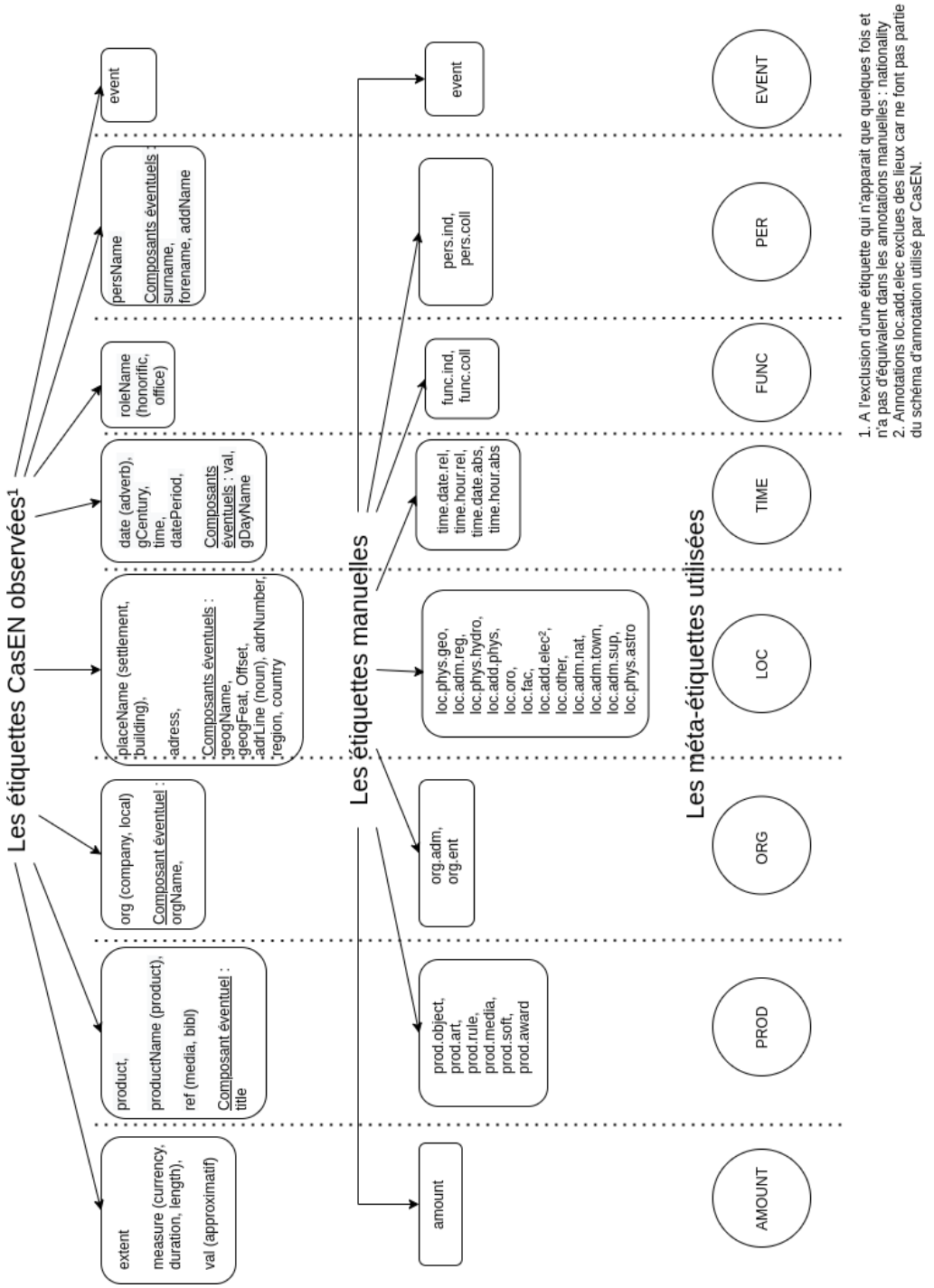

FIGURE 2.6 – Regroupements des étiquettes CasEN et manuelles effectués pour ce mémoire.

Nous présentons en figure 2.3.2 et en figure 2.8 les pourcentages de représentation de chaque type (donné par les méta-étiquettes) dans le cas des annotations manuelles réorganisées respectivement pour CasEN et SpaCy.

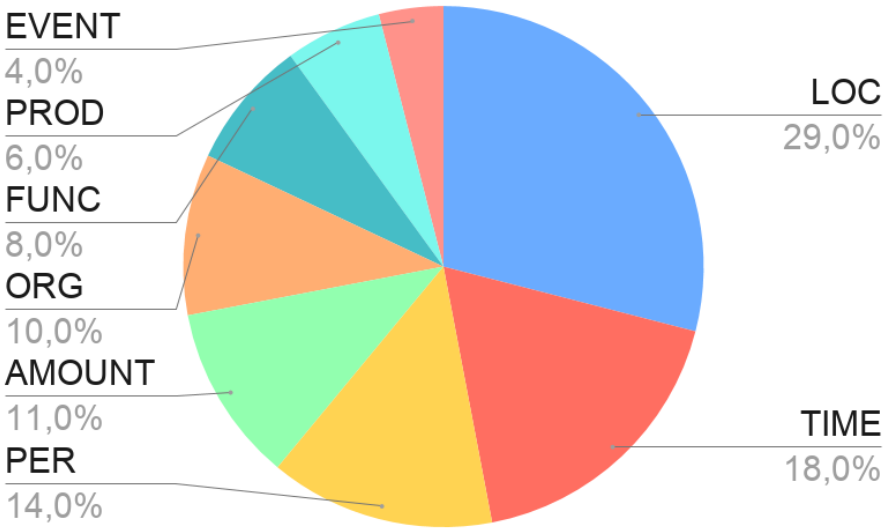

FIGURE 2.7 – Pourcentage de représentation de chaque catégorie dans les annotations manuelles réorganisées pour CasEN.

Nous avons utilisé les étiquettes SpaCy telles quelles : PER, ORG, LOC et MISC. Selon que l'on utilise les modèles *sm*, *md* ou *lg* évoqués précédemment, nous n'avons pas la même quantité ni la même qualité d'étiquette. Nous prendrons en compte ici les entités nommées et étiquettes obtenues avec le modèle *lg* qui permet une meilleure performance. Cela représente, après exécution, 369 entités nommées annotées.

2.4 Proposition d'unification

De manière générale, les annotations manuelles sont plus nombreuses et plus précises. D'un autre côté, l'annotation manuelle est laborieuse et source d'erreurs (il est parfois difficile de savoir si x est une entité nommée, ou de savoir pour une entité donnée quelle étiquette lui donner). Cela dit, l'annotation manuelle laisse ressortir une particularité essentielle des entités nommées. Cette particularité réside dans leur appréhension. La notion d'entité

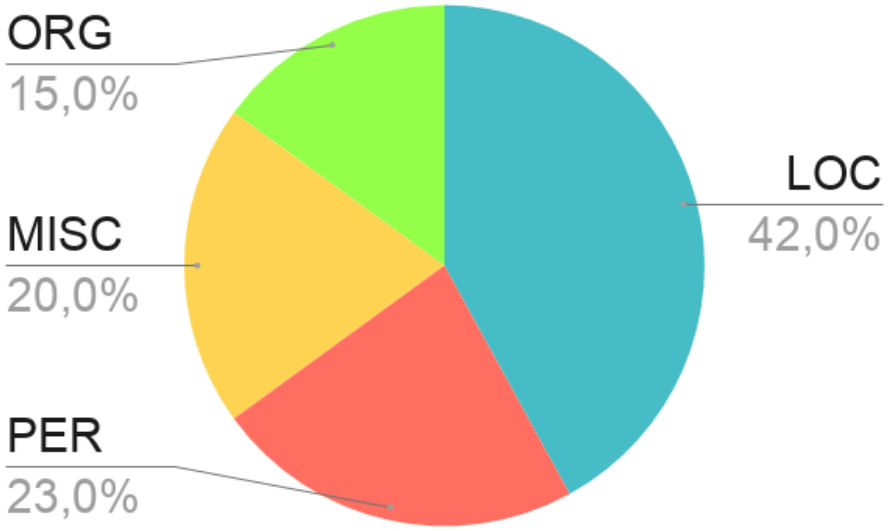

FIGURE 2.8 – Pourcentage de représentation de chaque catégories dans les annotations manuelles réorganisées pour SpaCy.

nommée est très variable d'un guide à l'autre, et donc d'un outil à l'autre. C'est là une caractéristique qui avait été évoquée dans l'état de l'art et qui faut à tout prix garder en tête avant de se lancer à corps perdu dans la comparaison et l'interprétation des résultats. Il serait peut-être judicieux de revoir les catégories d'annotation et d'unifier le tout. Cela a été tenté plusieurs fois et ce mémoire n'a pas la prétention de le faire mais nous proposons tout de même de catégoriser les outils de reconnaissance d'entités nommées de la manière suivante :

— ceux qui reconnaissent ENAMEX, NUMEX, et TIMEX
— ceux qui ne reconnaissent qu'ENAMEX

Cette catégorisation apparaîtrait dans la fiche des caractéristiques de chaque outil.

Par ailleurs, nous pourrions décider qu'au sein même de chaque type-d'entité (ENAMEX, NUMEX et TIMEX) il y aurait un certain nombre de champs à reconnaître et en fonction du nombre de champs reconnus, l'outil serait assigné à une classe. Par exemple pour les ENAMEX :

— classe A : reconnaissance minimum (PER, LOC, ORG)
— classe B : reconnaissance intermédiaire (PER, ORG, LOC, PROD, TIME, FUNC, EVENT AMOUNT) mais les étiquettes ne sont pas

détaillées
— classe C : reconnaissance complète. Cela inclut toutes les étiquettes de la classe B ainsi que les détails de chaque étiquette (par exemple pour FUNC : s'agit-il d'une fonction individuelle ou collective ? pour PROD : s'agit-il d'un objet ? d'une oeuvre ?, etc.)
— classe D : reconnaissance avancée. Cela inclut tout ce qui est reconnu par les outils de la classe C, mais également les composants de chaque étiquette.

Enfin, il convient d'ajouter que pour chaque type d'entité (LOC, ORG, etc.), une liste précise des éléments composant le type devrait être prescrite. Nous pourrions proposer de faire cette liste la plus exhaustive possible afin d'éviter les indécisions. Par exemple, prenons le cas de LOC et FUNC :

— LOC comprendrait toute sorte d'adresse : électronique ou géographique. Les numéros de téléphone sont considérés comme des adresses éléctroniques.
— FUNC comprendrait tant des fonctions officielles que non officielles et parmi ces dernières on considère fonction tout ce qui relève d'un travail : pompier est autant une fonction que danseur tant que dans le contexte donné désigne la danse comme le travail de l'individu.

Dans chacune des entités, les articles définis, les pronoms démonstratifs et autres tokens grammaticaux n'apparaîtraient pas, mis à part pour les TIMEX qui ont besoin d'être précisées.

L'unification ne se ferait alors pas par l'imposition d'un seul standard auquel tous les outils devraient se conformer mais plutôt par la proposition d'une classification que les outils auraient intérêt à suivre et ce pour deux raisons. La première raison serait que les créateurs des outils sauraient plus précisément ce à quoi ils doivent tendre lors de la programmation de leur outil. La deuxième raison serait que les utilisateurs sauraient ce qui doit être trouvé avec exactitude. Cela impliquerait bien entendu que chaque outil mentionne, commé évoqué précédemment, la catégorie de reconnaissance à laquelle il appartient.

2.5 Méthodologie pour mener notre comparaison

La comparaison des annotations pour une entité correctement trouvée a fait l'objet d'une adapatation de la distinction réalisée par Maud Ehrmann

dans sa thèse [Ehrmann, 2008] : une exactitude stricte concerne un annotable correctement trouvé et une bonne étiquette ; une exactitude partielle concerne un annotable correctement trouvé et une mauvaise étiquette. Cela permet d'une part la diversité des approches puisqu'on offre deux manières d'évaluer l'exactitude d'une reconnaissance, d'autre part l'amélioration éventuelle des outils (pourquoi telle entité a bien été trouvée mais l'étiquette est erronée ?) et de corriger les erreurs de programmation. Si l'exactitude partielle désigne habituellement une entité correctement étiquetée mais dont les bornes sont fautives, cela n'a pas été repris pour ce mémoire parce qu'une étude attentive des entités obtenues a révélé que les entités anormalement bornées ne sont pas correctement étiquetées. Nous invitons le lecteur à prendre connaissance des deux tableaux 2.7 et 2.8 ci-dessous qui reprennent tout ce qui a été annoté par les outils de manière fautive. Nous donnons en plus dans le tableau 2.9 ce qui aurait pu faire l'objet d'une autre catégorisation.

AMOUNT	un instant, Une seconde, des soirées, premiers livres, des saisons, Une soirée, 5S, 2ème jour, cinq minutes
EVENT	mondial
FUNC	général, foreuse, père, chef, voyageur, mère, chefs
LOC	Saint-Jean,surface, sommet, ville de Bretagne
ORG	jardin, Cabaret, entreprise, boutique, élevage, jardins, armées, école, équipe, presse, université, entreprises, marines, Café
PER	Division, Michel, Juif
PROD	Chroniques, magazine, documentaires
TIME	pour ce moment, les journées, la journée, avril, pour l' instant, chaque année, tout an, le siècle, chaque mois, le jour,

TABLE 2.7 – Liste des faux positifs en CasEN.

Méta-étiquette	Eléments annotés
MISC	XXème, Méchoui, Scène ouverte, Concert d'orgue, Division 2, Chroniques, Stage, l'Atelier de Sylvie Vernageau, Boule Palisienne, Reconnaître 5, http ://villa-de-l-extra.skyrock.com/, sauvageonnes comestibles, Juif, Vide grenier, PUISEAUX, Cabaret itinérant, l'Être ultime, Sortie au Musée Colette, Ma mère tout, Animations musicales, Frégates de Taïwan, Exposition, Sortie au départ d'AIX-EN-OTHE, internet Stop USA, Café Surprise : pratique artistique et chansons
LOC	Couvent de l'Orée Contact, Rencontres, Promenade, le Japon, Ours, Budget, Soirée, Pays d'Armance, Tous, Les Salins, Uni, Largentier)à 19 h., Renseignements, Burkina, Chantier, Randonnée, Rando Voirloups, Saintes Boson, Soi, États, Faso, Octeville, Cherbourg, département du Morbihan, Aix, Ethiopie la Somalie, rivière des Perles
ORG	Distribution, EDD, Andragatos, Metro, Patrimoine mondial, Club des Anciens, Goldwyn, Moët Hennessy Louis Vuitton
PER	Maraye, les Besse, Saint Laurent, Bob Rock, V, Mayer, M. Patrice Lanini, M. Christian Langlois

TABLE 2.8 – Liste des faux positifs en SpaCy.

Nom de l'entité	**Méta-étiquette**
Octoville	LOC
Cherbourg	LOC
Patrimoine mondial	ORG
Club des Anciens	ORG
M. Patrice Lanini	PER
département du Morbihan	LOC
Aix	LOC
Ethiopie la Somalie	LOC
M. Christian Langlois	PER
rivière des Perles	LOC
Moët Hennessy Louis Vuitton	ORG

TABLE 2.9 – Problèmes d'empan pour certaines entités annotées avec SpaCy.

Par ailleurs, nous invitons le lecteur à prendre connaissance des trois tableaux ci-dessous. Ces tableaux permettent de visualiser les différences d'annotation entre les sorties CasEN, SpaCy et les annotations manuelles.

	Extraits Aijwikiner
CasEN	L' <placeName>Inde</placeName>a trois périodes de vacances nationales. On retrouve l' influence de Ballard dans la musique de groupes de post-punk comme <persName>Joy Division</persName>, <orgName>The Normal </orgName> et <persName>John Foxx</persName>.
SpaCy	"Inde" : "LOC", "Joy Division" : "ORG", "The Normal" : "PER", "John Foxx" : "PER"
Manuel	"Inde" : "loc.adm.nat", "trois périodes" : "amount", "Ballard" : "pers.ind", "Joy Division" : "pers.coll", "The Normal" : "pers.coll", "John Foxx" : "pers.ind"

TABLE 2.10 – Exemples de sorties en CasEN, SpaCy vis-à-vis des annotations manuelles sur le corpus Aijwikiner.

	Extraits APIL
CasEN	<date when="-07-16">16 juillet</date> : BERULLE - Randonnée de <measure type="duration" quantity="9" unit="h">9 h</measure> 0 à 18 h (prévoir pique-nique) " <measure type="duration" quantity="Une" unit="seconde">Une seconde</measure> vie pour des arbres " Départ et retour à Bérulle. Participation <measure type="currency" quantity="5" unit="€">5 €</measure> et <measure type="currency" quantity="3" unit="€">3 €</measure> (adhérents).
SpaCy	"BERULLE" : "LOC", "Randonnée" : "LOC", "Bérulle" : "PER"
Manuel	"16 juillet" : "time.date.abs", "BERULLE" : "loc.adm.town", "de 9 h 30 à 18 h" : "time.hour.abs", "Bérulle" : "loc.adm.town", "5 €" : "amount", "3 €" : "amount"

TABLE 2.11 – Exemples de sorties en CasEN, SpaCy vis-à-vis des annotations manuelles sur le corpus APIL.

	Extraits GSD
CasEN	Puis elle rejoint <placeName>Hollywood</placeName> où, <date from="1930" to="1942">de 1930 à 1942</date>, elle participe à 175 films américains au sein de la Metro-Goldwyn-Mayer, principalement comme <roleName type="office">conceptrice </roleName> de robes, entre autres aux côtés d'Adrian.
SpaCy	"Hollywood" : "LOC", "Metro" : "ORG", "Goldwyn" : "ORG", "Mayer" : "PER", "Adrian" : "PER"
Manuel	"Hollywood" : "loc.adm.town", "de 1930 à 1942" : "time.date.abs", "175 films" : "amount", "Metro-Goldwyn-Mayer" : "org.ent", "Adrian" : "pers.ind"

TABLE 2.12 – Exemples de sorties en CasEN, SpaCy vis-à-vis des annotations manuelles sur le corpus GSD.

Chapitre 3

Résultats

Dans cette partie, nous nous efforçons de mener une comparaison des sorties obtenues et d'en donner des interprétations. Ensuite, nous proposons une discussion sur ce travail et les apports éventuels à la tâche de reconnaissance des entités nommées.

3.1 Comparaison et évaluation

La comparaison dont fait l'objet cette partie se situe à plusieurs niveaux. Nous les évoquons puis nous présentons les chiffres obtenus lors de notre comparaison.

3.1.1 Les niveaux de comparaison

Commençons par parler des niveaux de comparaison possibles des différentes annotations obtenues. Il faut prendre tout d'abord en considération la quantité des annotations : aucun des deux outils ne parvient à annoter autant d'entités que ne le fait une personne. Pour rappel (voir chapitre 2 pour plus d'informations), l'annotation manuelle représente 636 annotations. Les annotations manuelles adaptées à ce qui est annoté par CasEN sont au nombre de 622, celles adaptées aux annotations SpaCy sont au nombre de 389. Tandis que les annotations automatiques (en CasEN et SpaCy) comprennent respectivement 349 et 369 entités. En ne prenant en compte que les annotations automatiques, nous remarquons déjà qu'il y a presque autant d'annotations réalisées en CasEN qu'en SpaCy or nous aurions pu imaginer que cela ne serait pas le cas. En effet, CasEN ayant plus d'étiquettes disponibles, nous aurions pu penser obtenir plus d'annotations en CasEN qu'en SpaCy. Ceci sera confirmé plus tard par les chiffres de précision et de rappel.

Un deuxième niveau de comparaison serait la qualité des annotations. Dans ce but, deux types de comparaison sont faites : des comparaisons avec exactitude partielle, et des comparaisons avec exactitude stricte. Ces distinctions ont fait l'objet d'une explication dans la partie précédente. Nous la reprenons rapidement. Une exactitude partielle désigne dans ce travail une même entité annotée par l'humain et par l'outil sans prendre en compte l'étiquette qui lui a été assignée. Une exactitude stricte répond aux questions suivantes : est-ce que toutes les entités nommées sont annotées ? Est-ce qu'elles sont correctement annotées ? Dans ce cas, on prend en compte une même entité annotée par l'humain et par l'outil avec cette fois la prise en compte de l'étiquette ; cette dernière doit être la même dans les deux cas. Une autre interrogation survient : A-t-on trouvé le bon empan de texte et ce dernier est-il bien typé ? Nous faisons apparaître ci-dessous, tableaux 3.1 et 3.2, des bruits pour six types d'étiquettes CasEN puis des exemples de bruits obtenus avec SpaCy.

les journées : TIME
élevage : ORG
Café : ORG
jardin : ORG
des soirées : AMOUNT
5S : AMOUNT
magazine : PROD
Chroniques : PROD
voyageur : FUNC
Juif : PER
Division : PER

TABLE 3.1 – Bruit en CasEN.

Budget : LOC
Renseignements : ORG
Sortie au départ d'AIX-EN-OTHE : MISC
V : PER

TABLE 3.2 – Bruit en SpaCy

Exactitude	Stricte			Partielle		
	VP	FN (silence)	FP (bruit)	VP	FN	FP
CasEN	281	359	50	299	341	50
SpaCy	245	145	105	284	106	67

TABLE 3.3 – Répartition des vrais/faux positifs/négatifs pour CasEN et SpaCy vis-à-vis des annotations manuelles (nombre absolu).

Les exemples donnés sont plutôt représentatifs de l'ensemble du bruit dont nous pouvons retrouver une liste complète dans le tableau 2.8.

Nous désignons par l'empan d'une entité sa longueur, c'est-à-dire là où elle commence et là où elle s'arrête. Les empans n'ont pas été pris en compte. Pour la comparaison avec CasEN, ils ne présentent aucun problème. En SpaCy, certaines entités comme dans l'exemple (1) présentent un problème d'empan. Parmi ces entités, deux cas. Un premier cas où ces problèmes d'empan incluent presque toujours un problème d'étiquetage :

(1) « Sortie au départ d'AIX-EN-OTHE » étiqueté MISC

Nous voyons bien qu'il y a une entité au sein de l'exemple (1) mais elle aurait dû être annotée LOC et non MISC.

Dans un deuxième cas, il y a un problème d'empan et l'entité est correctement annotée. Cela concerne très exactement dix entités. Nous en donnons également la liste dans le chapitre 2, (tableau 2.9).

Nous avons donc fait le choix de ne pas prendre en compte les empans mais il aurait pu en être autrement.

3.1.2 Evaluation

Éléments vrais ou faux, positifs ou négatifs

Nous évoquons dans cette partie les vrais positifs, vrais négatifs, faux positifs et faux négatifs et plus spécifiquement le bruit et silence observé. Les définitions que nous en donnons sont utiles pour comprendre les formules de précision, de rappel, et de f-mesure rappelées plus tard. Les chiffres sont donnés dans le tableau 3.3.

Nous considérons un élément vrai positif comme une entité qui est correctement trouvée.

Nous appelons bruit (faux positif) un élément qui est trouvé et qui n'aurait pas dû être trouvé. Les bruits n'ont pas un seul profil. On en a de deux types : des noms communs ou signes de ponctuation (pour le cas de SpaCy) qui n'ont absolument rien à voir avec les entités nommées, puis des morceaux d'entités nommées contenues dans un empan plus grand dont l'étiquette est erronée. Par exemple pour le cas de CasEN « Michel » a été trouvé et annoté comme PER mais il s'agissait en fait du Mont-Saint-Michel, qui aurait dû être annoté LOC. Avec SpaCy, un exemple de bruit serait « Saintes Boson » alors que l'on parle ici de l'évêque de Saintes et qu'il s'appelle Boson. On a aussi « Goldwyn » et « Mayer » annotés en PER alors qu'il s'agit de l'entité « Metro-Goldwyn-mayer » et que c'est une ORG. Encore un autre exemple : « ERVY-LE-CHATEL – Stage » dont l'annotation est MISC (le programme aurait juste dû trouver « ERVY-LE-CHATEL » sans « Stage » et avec l'étiquette LOC).

Nous considérons enfin comme étant un vrai négatif quelque chose qui n'est pas trouvé et qui n'aurait en effet pas dû être trouvé.

Nous considérons un élément faux négatif (silence) quelque chose qui n'est pas trouvé mais qui aurait dû être trouvé. Dans le cas de CasEN, les faux négatifs concernent toutes les catégories d'entité sans exception. On a jusqu'à 359 (exactitude stricte) éléments silences. Dans le cas de SpaCy, on en a 145.

Nous présentons en figure 3.4 et 3.5 les tableaux des catégories auxquelles appartiennent les faux négatifs avec leur pourcentage de représentation.

Catégorie	Pourcentage
LOC	29
PER	14
ORG	13
PROD	11
TIME	10
AMOUNT	10
FUNC	7
EVENT	6

TABLE 3.4 – Catégories auxquelles appartiennent les faux négatifs_CasEN.

Catégorie	Pourcentage
LOC	36
MISC	35
ORG	18
PER	11

TABLE 3.5 – Catégories auxquelles appartiennent les faux négatifs_SpaCy.

Toutes les catégories sont représentées bien entendu. Mais il faudrait également voir si ce sont les catégories les plus représentées de base. Nous présentons en figure 3.6 et en figure 3.7 les catégories telles que représentées de manière générale.

Catégorie	Pourcentage
TIME	26
LOC	25
PER	14
AMOUNT	12
FUNC	10
ORG	10
EVENT	2
PROD	1

TABLE 3.6 – Catégories représentées avec CasEN parmi les annotables reconnus.

Catégorie	Pourcentage
LOC	43
PER	24
MISC	18
ORG	15

TABLE 3.7 – Catégories représentées avec SpaCy parmi les annotables reconnus.

L'observation du tableau 3.6 nous permet d'indiquer que l'étiquette PROD représente 1 % des annotations en temps normal mais 11 % parmi celles qui n'ont pas été reconnues, et que l'étiquette TIME représente 26 % des annotations en temps normal mais 10 % parmi celles qui n'ont pas été reconnues.

Cela signifie, qu'en règle général, CasEN reconnaît bien ce qui se cache sous l'étiquette TIME (à savoir : les dates, les heures, les années, les moments) ; en

revanche il ne reconnaît pas très bien ce qui se cache sous la méta-étiquette PROD (à savoir les productions artistiques, scientifiques, les prix, etc.).

L'observation du tableau 3.7 nous permet de noter que l'étiquette MISC représente 18 % des annotations en temps normal mais 35 % parmi les annotables qui n'ont pas été reconnus et que l'étiquette PER représente 24 % des annotations en temps normal mais 11 % parmi les annotables non reconnus.

On peut donc dire que SpaCy reconnaît plutôt bien les noms de personne.

Précision, Rappel et F-mesure

Nous faisons apparaître la classification des résultats en CasEN et SpaCy. Afin d'évaluer les annotations obtenues avec l'un et l'autre des outils, nous faisons appel aux mesures de précision (annotations correctes de l'outil au sein de l'ensemble des annotations fournies par l'outil), de rappel (annotations correctes de l'outil au sein de l'ensemble des annotations de référence) et de f-mesure.

Nous rappelons le calcul de chaque mesure :

$$P^{1} = \frac{VP^{2}}{VP + FP^{3}}$$

$$R^{4} = \frac{VP}{VP + FN^{5}}$$

$$F^{6} = \frac{2PR}{P + R}$$

Ces mesures ont été calculées avec les données fournies précédemment (voir tableau 3.3) et nous en donnons les résultats ci-dessous.

Globalement, SpaCy a un rappel plus important que CasEN (voir tableaux 3.9 et 3.8). Nous pourrions alors dire que SpaCy annote plus d'entités par catégorie que ne le fait CasEN. Ceci dit, CasEN est meilleur en précision ; cela signifie que la proportion d'entités annotées pertinentes est plus importante.

1. Précision
4. Rappel
6. F-mesure

exactitude	strict			partiel		
	P	R	F1	P	R	F1
APIL	**0,84**	**0,33**	0,47	**0,89**	**0,35**	0,51
Wikinews	**0,81**	0,48	0,60	**0,87**	0,52	0,65
Sequoia	**0,91**	0,51	0,66	**0,94**	0,53	0,68
Aijwikiner	**0,81**	0,49	0,61	**0,90**	0,53	0,67
GSD	0,74	0,42	0,53	0,80	0,45	0,58
Spoken	**0,61**	0,46	0,52	**0,63**	0,48	0,55
pg6470	0,72	0,44	0,55	0,72	0,44	0,55
tous	0,81	0,44	0,57	0,83	0,45	0,58

TABLE 3.8 – Évaluation des annotations faites avec CasEN.

exactitude	strict			partiel		
	P	R	F1	P	R	F1
APIL	**0,45**	**0,46**	**0,46**	0,61	0,61	0,61
Wikinews	0,78	0,65	0,71	0,86	0,71	0,78
Sequoia	0,73	0,62	0,67	0,82	0,70	0,77
Aijwikiner	**0,80**	0,76	0,78	**0,91**	0,86	0,88
GSD	0,73	0,69	0,71	0,84	0,79	0,81
Spoken	0,78	0,55	0,64	**0,87**	0,61	0,71
pg6470	**0,86**	**0,75**	**0,80**	**0,93**	**0,81**	**0,87**
tous	0,70	0,63	0,66	0,81	0,73	0,77

TABLE 3.9 – Évaluation des annotations faites avec SpaCy.

Par ailleurs, nous remarquons que CasEN est meilleur avec des documents de type informatif qu'avec des documents didactiques, dialogiques et narratifs. En effet, avec Sequoia, Wikinews et APIL, il obtient de relativement bonnes mesures de précision (qui vont de 0,81 à 0,94). Une exception est à noter avec le document Aijwikiner qui obtient également de bons résultats en précision (0,81 et 0,90 en précision). Le document posant le plus de problème à CasEN est Spoken. Comme c'est un document de type dialogique, cela peut s'expliquer et nous reviendrons sur ce point ultérieurement.

SpaCy est meilleur avec les documents de type narratif (précision de 0,86 en exactitude stricte et de 0,93 en exactitude partielle, rappel de 0,75 en exactitude stricte et de 0,81 en partielle, F1-mesure de 0,80 en exactitude stricte et de 0,87 en partielle).

SpaCy génère beaucoup de bruit. On retrouve beaucoup de choses annotées qui ne sont pas même des bouts d'entités mais tout autre chose (pour voir ces le tableau 2.8). Il apparaît un peu meilleur que ne l'est CasEN sur l'oral (jusqu'à 0,87 de précision, 0,64 de rappel et 0,71 de f-mesure). Il obtient des scores relativement corrects avec les documents de type dialogique et informatif. Nous nous serions attendus à des scores plus élevés avec ces derniers puisque l'outil a été entraîné sur Sequoia et Wikiner. Enfin, si l'exactitude partielle permet d'obtenir en moyenne les mêmes scores que ceux indiqués par les constructeurs de l'outil (voir tableau 2.6), il n'en est rien en exactitude stricte.

3.2 Interprétations

3.2.1 Interprétations générales

Si CasEN est meilleur en précision, cela s'explique par la faible quantité de bruit présent parmi les annotations obtenues. En effet, l'outil obtient 50 éléments faux positifs tant en exactitude partielle que stricte contre 105 et 67 pour SpaCy qui n'annote pourtant pas autant d'entités que CasEN. Cela dit, CasEN a une très grande quantité de faux négatifs (359 en exactitude stricte et 341 en exactitude partielle) ; pour être exacte, CasEN a plus de faux négatifs que de vrais positifs alors que SpaCy a moins de faux négatifs que de vrais positifs. Cela rend SpaCy meilleur en rappel : proportionnellement (et sachant qu'il ne fait dispose pas de beaucoup de types différentes) il annote beaucoup plus, même si cela entraîne plus d'erreurs.

Les résultats sont pour SpaCy comme pour CasEN moins bons à l'oral (corpus Spoken). À l'oral, nous avons en effet tendance à moins enrober le message. Bien sûr, il y a des marques de réflexion ou d'hésitation comme « euh » qui peuvent apparaître, des tics de langage « genre », « du coup », « en fait » qui se glissent par-ci par-là. Mais, à l'oral, à moins que l'on ne se trouve dans une situation très

codifiée (discours présidentiel ; dans ce cas, le président lit ou a appris son discours par coeur, etc.), on délivre le message assez directement. Voilà pourquoi les résultats sont un peu meilleurs. À l'écrit, on pense son énoncé, on le réfléchit, on le tourne dans un sens ou dans un autre, on le reformule à volonté jusqu'à trouver la tournure qui nous plait.

3.2.2 Etude de cas d'un élément appartenant au bruit observé

On trouve dans le corpus Spoken un exemple d'entité nommée temporelle :

(2) depuis oui cinq minutes

Nous avons là un exemple de « bafouillage », phénomène qui a été décrit par [Blanche-Benveniste, 1987]. Lié à la dimension paradigmatique du langage, le bafouillage comprend entre autres les constructions interrompues. Ainsi, quand bien même un « oui » se soit incrusté en (2), et conformément au guide Quaero, l'annotation manuelle tient compte du fragment entier. En CasEN, on retrouve l'annotation suivante :

(3) cinq minutes | AMOUNT

Cette étude de cas est importante puisque les dysfluences propres à l'oral sont monnaie courante et il convient de pouvoir reconnaître les entités nommées tant dans un texte écrit qu'un corpus oral. On remarque dans ce cas particulier que l'outil utilisé (CasEN) manque de précision pour l'oral. Cela s'explique par la nature même de l'outil qui fonctionne à base de transducteurs. Le bafouillage de l'oral entrecoupe les constructions syntaxiques qui ne sont pas alors très bien analysées.

3.3 Discussion

Notons concernant l'évaluation des outils utilisés qu'elle est le fruit de choix. Nous avons fait le choix d'adapter les annotations manuelles d'une part et d'appauvrir les annotations CasEN d'autre part pour que les deux jeux d'étiquettes coïncident ; ainsi, comme indiqué précédemment, n'apparaissent pas les composants (*forename* par exemple). Pour SpaCy, seules les annotations manuelles ont été adaptées et les annotations automatiques ont été utilisées telles quelles. Cela aurait pu sûrement être fait différemment. Les annotations manuelles auraient pu éventuellement faire figurer les composants de chacune de ces entités, ce qui aurait été à l'avantage de CasEN, l'on aurait pu n'évaluer qu'une partie des annotations automatiques (celles qui coïncident le plus), etc. Dans tous les cas, ces choix révèlent une chose très importante et qui doit faire l'objet d'une grande attention : les chiffres qui sont donnés dans ce mémoire ne sont donnés qu'à titre indicatif et ne permettent en aucun cas de comparer les deux outils immédiatement. Par ailleurs,

le cadre expérimental de ce travail n'est pas avantageux pour les outils puisqu'il concerne un corpus relativement bien équilibré et de sources diverses, ce qui n'est pas le cas des textes utilisés pour construire ces outils.

3.4 Apports

Bien que ces outils ne soient pas parfaits, il n'est pas nécessaire d'en exclure tout de go l'utilité.

Puisque CasEN semble être meilleur en précision et SpaCy en rappel, nous pourrions très bien imaginer une utilisation combinée de ces deux outils. Le second passerait en premier pour annoter un maximum d'entités ENAMEX. Il suffirait ensuite de passer le premier sur le même texte. En complément, il suffirait de passer manuellement sur chaque annotation pour vérifier leur cohérence. Cela permettrait sans aucun doute une annotation plus rapide. En revanche, il est sans doute utile de préciser qu'une telle utilisation ne peut se faire à grande échelle sans entraîner une perte de temps considérable et un taux de précision décroissant puisque les tâches répétitives, nous le savons, ne sont pas adaptées au cerveau humain qui se fatigue facilement.

Nous pouvons également proposer une utilisation combinée de ces deux systèmes au sein d'un même outil. C'est-à-dire un outil qui mélangerait l'apprentissage machine et la transduction.

Conclusion

Ce que l'on désigne comme entité nommée au sein de la communauté scientifique ne fait pas l'objet d'un large consensus. Il existe pourtant, au sein du discours, des éléments qui, sémantiquement parlant, sont plus importants et qui sont consensuellement regroupés sous l'expression d'entité nommée. En conséquence, les personnes élaborant un outil de reconnaissance d'entités nommées décident de deux choses. Ils décident de ce qui doit être reconnu par leur outil comme un annotable et ils décident d'adopter un schéma d'étiquetage. Dans ce mémoire nous avons discuté de deux outils très différents : CasEN (fonctionnant à base de règles) et SpaCy (fonctionnant à base d'apprentissage machine). Ces deux outils réalisent a priori la même tâche ; on appelle cette tâche la reconnaissance d'entités nommées. Pourtant, leur comparaison s'avère compliquée. Les résultats de leur exécution ne sont pas immédiatement comparables et les chiffres que nous avons pu faire apparaître dans ce mémoire ne rendent pas très clairement compte de leurs capacités respectives. Si les résultats ne sont pas immédiatement comparables cela s'explique d'une part par les choix réalisés par les constructeurs concernant ce qui se cache derrière la notion d'entité nommée et la catégorisation utilisée. D'un côté nous avons un outil qui utilise de larges types relevant des ENAMEX, TIMEX et NUMEX et, au sein de certains de ces types, des composants apparaissent ; il s'agit de CasEN. De l'autre côté, nous avons un outil qui n'utilise que quatre larges types ne relevant que des ENAMEX et, au sein de ces types, aucun sous-type ni composant n'apparait ; il s'agit de SpaCy. D'autre part, les chiffres ne rendent pas clairement compte de leurs capacités respectives à cause de choix qui ont été faits pour les obtenir. Les sorties CasEN ont dû faire l'objet d'une réorganisation de leurs étiquettes au sein de méta-étiquettes afin de pouvoir mener une comparaison avec les annotations manuelles préalablement obtenues (ces annotations manuelles ont été faites sur la base du guide Quaero). Cela joue, bien entendu, en la défaveur de CasEN puisque nous n'avons pas su faire figurer le détail des étiquettes renvoyées par l'outil. Par ailleurs, les chiffres ne sont pas comparables entre eux car malgré cette réorganisation, nous avons utilisé huit méta-étiquettes pour évaluer CasEN alors que SpaCy ne dispose que de quatre étiquettes qui n'ont pas eu à être remaniées ; cette différence de quantité rend périlleuse une quelconque comparaison. Il est utile de noter que chacun de ces deux programmes a malgré tout des avantages, des qualités, mais aussi des défauts. CasEN est un outil précis qui souffre pourtant d'un faible

rappel. SpaCy est un outil qui annote en masse mais dispose de peu d'étiquettes (type ou sous-type) et souffre d'un manque de précision. Nottons malgré tout que ces conclusions ne dépendent que des résultats obtenus ici et ne sont discutables que dans le cadre de l'expérience réalisée. Ce cadre dépend des choix qui ont été faits et qui ont été précédemment rappelés et des mesures qui ont été utilisées pour évaluer les outils. Nous noterons ainsi que malgré les tentatives d'unification d'annotation réalisées depuis l'apparition des entités nommées, aucune n'a encore véritablement abouti et que l'hétérogénéité des conceptions au sein de la tâche de reconnaissance d'entités nommées constitue un obstacle à surmonter lorsqu'il s'agit de mener une comparaison d'outils. Malgré tout, il semble évident que cet obstacle demeurera aussi longtemps que la langue nous restera atteignable que par ses manifestations. Ainsi, si les outils de reconnaissance d'entités nommées présentent des qualités indéniables, nous pouvons sans prendre trop de risque émettre l'idée que leur utilisation doit être couplée à une intervention manuelle en fin de traitement pour démontrer tout l'étendue de leur capacité.

Annexes

Extraits pour chaque document du corpus

aijwikiner

Deux temps structurent l' œuvre, les cinquante premiers livres sont dédiés aux 723 premières années de Rome, de sa fondation à la bataille d' Actium en -31 et les trente derniers aux 250 années impériales. L' Inde a trois périodes de vacances nationales. On retrouve l' influence de Ballard dans la musique de groupes de post-punk comme Joy Division, The Normal et John Foxx. Jiang Qing, la femme de Mao Zedong, et la bande des Quatre agite le mouvement contre les chaînes culturelles du passé : de nombreuses œuvres anciennes, livres, sculptures, bâtiments, etc. L' explication d' Arrhenius est que lors de sa dissolution, le sel se dissocie en particules chargées (que Michael Faraday avait nommé " ions " quelques années avant). Plus récemment, la série Fringe se déroule également à Boston. Les populaires séries télévisées québécoises Lance et compte (1986 –) ont également le hockey sur glace comme thème central.

APIL

ERVY-LE-CHATEL – 2ème jour du Marché du Livre à la salle des fêtes de 10 h à 18 h. Rencontres avec les auteurs, lectures entre 12 h et 14 h. Organisée par « Ervy-le-Châtel, village du livre et des arts » 03 25 76 88 78. 23 juillet : LIGNIERES – Reconnaître 5 sauvageonnes comestibles. Savoir les utiliser et les cuisiner. Organisé par Les Ombelles 03 25 43 92 26 ou www.lesombelles.com Octobre 2011 15 septembre : Sortie au départ d'AIX-EN-OTHE – Sortie au Musée Colette et Château de Guédelon. Organisé par le Club des Anciens et de amis d'Aix 06 75 15 06 14. 10 et 11 décembre : BOUILLY – Marché de Noël au foyer familial, organisé par le Comité des fêtes. ERVY-LE-CHATEL – Récital de Piano 4 mains avec Matthieu Normand et Nicolas Collinet à la Halle Circulaire à 20 h 30. Organisé par Villa de l'Extra 03 25 43 22 10 ou 06 79 07 79 25 ou http ://villa-de-l-extra.skyrock.com/

GSD

Outre ces îles, la ville est composée de nombreux quartiers tels que La Capte, Giens, L'Almanarre, L'Ayguade, Le Pyanet, Costebelle, Les Salins-d'Hyères ou Les Borrels. D'un côté El Niño est responsable d'une importante sécheresse et de l'autre une vague de froid traverse le pays. Shusterman poursuit cette réflexion dans plusieurs essais subséquents tels Practicing Philosophy (1997), Performing Live (2000) et Surface and Depth (2002). Le ‘ ndranghetiste, du mot grec Andragatos, qui signifie homme valeureux et courageux, a des origines lointaines. Cet antisémitisme ne constitue pas l'ensemble de la réflexion d'un penseur qui est l'un des plus grands analystes du XXème siècle en France. Et puis, ma chaîne est tombée. Puis elle rejoint Hollywood où, de 1930 à 1942, elle participe à 175 films américains au sein de la Metro-Goldwyn-Mayer, principalement comme conceptrice de robes, entre autres aux côtés d'Adrian. La Centrale féline Belge (CFB) est un registre d'élevage belge.

sequoia

Expression parue dans la revue n° 33 "Verdun la vie" et envoyée à tous les Verdunois, ne correspond pas à la réalité, ce que soulignent M. Christian Langlois et M. Patrice Lanini, délégués syndicaux CGT des Rapides de la Meuse. Les multiples auditions conduites par le juge Armand Riberolles ont essentiellement porté sur le rôle du promoteur Jean-Claude Méry, dont les révélations avaient relancé l'enquête. Plusieurs hommes auraient, pour le compte de Francis Poullain, fait plusieurs aller et retour entre la France et l'Afrique pour transporter des fonds. François Mitterrand était alors ministre des Colonies (à partir de 1950) et René Bousquet directeur de la Banque d'Indochine. - Le 17 mai 2004, l'ancien ministre socialiste du Budget Michel Charasse a reconnu avoir signé des commissions "légales" pour les frégates de Taïwan. 5 x 1 flacon (conditionnement unitaire) 2 flacons Quelques toiles, de l'acrylique, de l'huile, de l'encre de Chine, mais aussi de la sculpture et bien d'autres techniques. Blanche Maupas, décédée en 1962, avait aussi écrit un livre, paru en 1933, Le Fusillé, dont la réédition en 1994 comporte des illustrations de Tardi.

spoken

Expression parue dans la revue n° 33 "Verdun la vie" et envoyée à tous les Verdunois, ne correspond pas à la réalité, ce que soulignent M. Christian Langlois et M. Patrice Lanini, délégués syndicaux CGT des Rapides de la Meuse. Les multiples auditions conduites par le juge Armand Riberolles ont essentiellement porté sur le rôle du promoteur Jean-Claude Méry, dont les révélations avaient relancé l'enquête. Plusieurs hommes auraient, pour le compte de Francis Poullain, fait plusieurs aller et retour entre la France et l'Afrique pour transporter des fonds. François Mitterrand était alors ministre des Colonies (à partir de 1950) et René Bousquet directeur de la Banque d'Indochine. - Le 17 mai 2004, l'ancien ministre socialiste du Budget

Michel Charasse a reconnu avoir signé des commissions "légales" pour les frégates de Taïwan. 5 x 1 flacon (conditionnement unitaire) 2 flacons Quelques toiles, de l'acrylique, de l'huile, de l'encre de Chine, mais aussi de la sculpture et bien d'autres techniques. Blanche Maupas, décédée en 1962, avait aussi écrit un livre, paru en 1933, Le Fusillé, dont la réédition en 1994 comporte des illustrations de Tardi. Nous devons considérer les personnes qui votent pour ces partis et comprendre pourquoi cette situation voit le jour.

pg6470

Les mauvaises pensées me dérangent trop. – Toi, mon petit, je t'en.... et, tiens ! Ce fut Alexandre qui paya. grand poëte que vous êtes ! demanda la vieille, toute frétillante, enchantée d'apprendre que les deux femmes s'étaient disputées. Elle était tout près de lui. Les paroles de Lisa retentissaient, comme s'il eût déjà entendu les fortes bottes des gendarmes, à la porte de la chambre. J'y arriverai, à l'histoire du monsieur... Je te raconte l'histoire tout entière. Elle voulut deux tranches de galantine ; elle aimait ça. Mais elle dit que non, qu'il fallait savoir où ils en étaient auparavant. Tous trois s'en allaient, traînant les talons sur les trottoirs, tenant la largeur, forçant les gens à descendre. Alors, mademoiselle Saget lui répondait que, dame ! Si, d'ailleurs, on ne l'avait pas arrête, au 2 décembre, ce n'était pas sa faute. Tenez, je ne vous aime plus. Cette brutalité jeta Florent hors de lui. – Merci, gronda la portière, cinquante francs, pour l'avoir dorloté avec de la tisane et du bouillon ! C'était barbare et superbe, quelque chose comme un ventre aperçu dans une gloire, mais avec une cruauté de touche, un emportement de raillerie tels, que la foule s'attroupa devant la vitrine, inquiétée par cet étalage qui flambait si rudement... Quand ma tante Lisa revint de la cuisine, elle eut peur, s'imaginant que j'avais mis le feu aux graisses de la boutique. Là, dans l'air renfermé, dans le demi-jour des quelques becs de gaz, il retrouvait la fraîcheur de l'eau pure. – Monsieur Gavard est tout an fond, dit le jeune homme, qui marchait toujours.

wikinews

Valve : de nouveaux jeux annoncés après plusieurs années sans sorties 9 mars 2018 . Le M5S se présente généralement comme anti-système et œuvre pour la mise en place d'un revenu universel , la baisse de l'impôt sur le revenu, la couverture des coûts de garde des enfants et la mise en place de traités bilatéraux pour rapatrier les clandestins. Le mandat d'arrêt européen lancé à son encontre le 2 novembre 2017 alors qu'il était en Belgique, avait été retiré le 5 décembre, car la justice espagnole craignait que les juges belges refusent de retenir les chefs d'inculpation, affaiblissant le dossier. En 1952, il fonde sa maison de couture. Puis, il s'est agressé à un groupe de CRS qui étaient en train de terminer leur footing. La semaine dernière, le Royaume-Uni a déjà expulsé 23 diplomates russes. – Quelques mois après le réseau social Facebook , Google renforce le contrôle des publicités diffusées sur AdWords ,

son service de régie publicitaire. Les forces de défense et de sécurité burkinabés ont riposté, abattant les 8 assaillants et en capturant 2. Valve recommence à publier des jeux. Plus de cinquante wikimédien.ne.s provenant de plusieurs pays africains, Algérie , Égypte , Cameroun , Nigéria etc. – Le Glacier Perito Moreno classé au Patrimoine mondial de l'UNESCO [1] est sur le point de s'effondrer. Il s'agit d'un homme marocain de 26 ans qui aurait déclaré agir au nom de Daesh . La rentrée atmosphérique de Tiangong-1 intéresse particulièrement l'agence européenne, mais aussi les autres agences spatiales.

Bibliographie

[Artstein and Poesio, 2008] Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4) :555–596.

[Blanche-Benveniste, 1987] Blanche-Benveniste, C. (1987). Syntaxe, choix de lexique, et lieux de bafouillage. *DRLAV. Documentation et Recherche en Linguistique Allemande Vincennes*, 36(1) :123–157.

[Candito and Seddah, 2012] Candito, M. and Seddah, D. (2012). Le corpus sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *TALN 2012-19e conférence sur le Traitement Automatique des Langues Naturelles*.

[Chinchor and Robinson, 1997] Chinchor, N. and Robinson, P. (1997). Muc-7 named entity task definition. In *Proceedings of the 7th Conference on Message Understanding*, volume 29, pages 1–21.

[Chomsky, 1999] Chomsky, N. (1999). Language and freedom. *Resonance*, 4(3) :86–104.

[Church, 2011] Church, K. (2011). A pendulum swung too far. *Linguistic Issues in Language Technology*, 6(5) :1–27.

[Collins and Singer, 1999] Collins, M. and Singer, Y. (1999). Unsupervised models for named entity classification. In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

[De Castilho et al., 2016] De Castilho, R. E., Mujdricza-Maydt, E., Yimam, S. M., Hartmann, S., Gurevych, I., Frank, A., and Biemann, C. (2016). A web-based tool for the integrated annotation of semantic and syntactic structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 76–84.

[Ehrmann, 2008] Ehrmann, M. (2008). *Les Entitées Nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation*. PhD thesis, Paris Diderot University.

[Fort, 2012] 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.

[Francis and Kucera, 1979] Francis, W. N. and Kucera, H. (1979). Brown corpus manual. *Letters to the Editor*, 5(2) :7.

[Friburger and Maurel, 2004] Friburger, N. and Maurel, D. (2004). Finite-state transducer cascades to extract named entities in texts. *Theoretical Computer Science*, 313(1) :93–104.

[Gerdes and Kahane, 2017] Gerdes, K. and Kahane, S. (2017). Trois schémas d'annotation syntaxique en dépendance pour un même corpus de français oral : le cas de la macrosyntaxe. In *Atelier sur les corpus annotés du français (ACor4French)*.

[Grave, 2014] Grave, E. (2014). Weakly supervised named entity classification. In *Workshop on Automated Knowledge Base Construction (AKBC)*.

[Grishman and Sundheim, 1996] Grishman, R. and Sundheim, B. M. (1996). Message understanding conference-6 : A brief history. In *COLING 1996 Volume 1 : The 16th International Conference on Computational Linguistics*.

[Grouin, 2018] Grouin, C. (2018). Simplification de schémas d'annotation : un aller sans retour ? In *Actes de TALN*.

[Grouin et al., 2011] Grouin, C., Galibert, O., Rosset, S., Quintard, L., and Zweigenbaum, P. (2011). Mesures d'évaluation pour entités nommées structurées. *Évaluation des méthodes d'Extraction de Connaissances dans les Données, Brest, France*.

[Honnibal et al., 2020] Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020). spaCy : Industrial-strength Natural Language Processing in Python.

[Le Pevedic and Maurel, 2016] Le Pevedic, S. and Maurel, D. (2016). Retour sur les annotations des entités nommées dans les campagnes d'évaluation françaises et comparaison avec la tei. *Corela. Cognition, représentation, langage*, 14(14-2).

[Maurel et al., 2011] Maurel, D., Friburger, N., Antoine, J.-Y., Eshkol, I., and Nouvel, D. (2011). Cascades de transducteurs autour de la reconnaissance des entités nommées. *Traitement automatique des langues*, 52(1) :69–96.

[McDonald et al., 2013] McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., et al. (2013). Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, pages 92–97.

[Nadeau and Sekine, 2007] Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1) :3–26.

[Nivre et al., 2016] Nivre, J., De Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., et al. (2016). Universal dependencies v1 : A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

[Nothman et al., 2013] Nothman, J., Ringland, N., Radford, W., Murphy, T., and Curran, J. R. (2013). Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194 :151–175.

[Omrane et al., 2010] Omrane, N., Nazarenko, A., and Szulman, S. (2010). Les entités nommées : éléments pour la conceptualisation. In *21es Journées francophones d'Ingénierie des Connaissances*, pages http–www.

[Phase, 2005] Phase, I. (2005). Evaluation campaign for the rich transcription of french broadcast news, in. In *European Conference*.

[Pustejovsky et al., 2003] Pustejovsky, J., Castano, J. M., Ingria, R., Sauri, R., Gaizauskas, R. J., Setzer, A., Katz, G., and Radev, D. R. (2003). Timeml : Robust specification of event and temporal expressions in text. *New directions in question answering*, 3 :28–34.

[Rau, 1991] Rau, L. F. (1991). Extracting company names from text. In *Proceedings the Seventh IEEE Conference on Artificial Intelligence Application*, pages 29–30. IEEE Computer Society.

[Rosset et al., 2012] Rosset, S., Grouin, C., Fort, K., Galibert, O., Kahn, J., and Zweigenbaum, P. (2012). Structured named entities in two distinct press corpora : Contemporary broadcast news and old newspapers. In *6th Linguistics Annotation Workshop (The LAW VI)*, pages 40–48.

[Sang and De Meulder, 2003] Sang, E. F. and De Meulder, F. (2003). Introduction to the conll-2003 shared task : Language-independent named entity recognition. *arXiv preprint cs/0306050*.

[Sekine and Nobata, 2004] Sekine, S. and Nobata, C. (2004). Definition, dictionaries and tagger for extended named entity hierarchy. In *LREC*. Lisbon, Portugal.

[Shen et al., 2004] Shen, D., Zhang, J., Su, J., Zhou, G., and Tan, C. L. (2004). Multi-criteria-based active learning for named entity recognition. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 589–596.

[Stanislawek et al., 2019] Stanislawek, T., Wróblewska, A., Wójcicka, A., Ziembicki, D., and Biecek, P. (2019). Named entity recognition–is there a glass ceiling ? *arXiv preprint arXiv :1910.02403*.

[Stenetorp et al., 2012] Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and 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*, pages 102–107.

[Tjong Kim Sang, 2002] Tjong Kim Sang, E. F. (2002). Introduction to the conll-2002 shared task : Language-independent named entity recognition. In *Proceedings of CoNLL-2002*, pages 155–158. Taipei, Taiwan.

[Tran and Maurel, 2006] Tran, M. and Maurel, D. (2006). Prolexbase : Un dictionnaire relationnel multilingue de noms propres. *Traitement automatique des langues*, 47(3) :115–139.

[Yimam et al., 2013] Yimam, S. M., Gurevych, I., de Castilho, R. E., and Biemann, C. (2013). Webanno : A flexible, web-based and visually supported system

for distributed annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, pages 1–6.