«Rapport de la Linguistique et du *traitement automatique du langage à travers la notion d'entité nommée*»

Du TAL à la linguistique et de la linguistique au TAL

Mémoire

présenté par

Virgile CARTIER

MASTER 1

« Informatique et Langue »

dirigé par

Yoann Dupont

Gaël Lejeune

REMERCIEMENTS

Mes plus vifs remerciements à monsieur Dupont et monsieur Lejeune pour leur patience et leurs conseils, à l’équipe pédagogique du master de langue et informatique, à Martial Pastor pour son soutien, et bien sûr à mes parents.

LISTE DES TRADUCTIONS ET ABRÉVIATIONS UTILISÉES

Terme français	**Terme anglais**	**Abréviation française**	**Abréviation anglaise**
Nom propre	Proper name	NP	PN
Entité nommée	Named entity	EN	NE
Traitement automatique des langages naturels	Natural language processing	TAL ou TALN	NLP
Extraction d’information	Information extraction	EI	IE
Traduction automatique	Automatic translation	TA	AT
Intelligence artificielle	Artificial intelligence	I.A	A.I

Table des matières

INTRODUCTION

Avec l'essor d'internet s'est développé l'élan d'un partage rapide et, maintenant, quasi instantané de l'information. À ajouter à l'équation le développement de la numérisation de données multiples et variées : *textuelles, visuelles,* etc. – le tout stocké et partagé sur le réseau. De fait, quelque soit nos occupations, nous sommes tous dorénavant amenés à partager, échanger, reproduire, rechercher des informations quelconques par le biais de ces documents numériques. Du reste, l'internet est de ces inventions qui marquent une nette frontière entre l'*avant* et l'*après* : tant pratique a t'elle été, qu'elle apparaît maintenant comme indispensable. La tâche la plus triviale – rechercher l'horaire d'un film, le titre d'un roman quelconque, etc – conduit à son utilisation. Quelque part, internet a suffisamment su se montrer essentiel qu'il conditionne de plus en plus nos activités ; sont produits et exploités un nombre vertigineux de documents – et au regard de l'immensité de cette production, il s'est imposé l'idée de mettre au point des systèmes automatisés dont le *Traitement automatique du langage Naturel* pour traiter toutes ces données textuelles.

Aussi, il faut tout de même rappeler que le rapport de la langue à l'informatique remonte à ses prémices, dans les années 50, essentiellement dans le monde anglo-saxon et plus précisément aux États-Unis. Néanmoins cette explosion significative de l'accès à l'information ainsi qu'à son partage a été vécue comme un second souffle, insufflant une vigueur nouvelle à la recherche en traitement de données textuelles. Une vigueur d'autant plus forte que cette nouvelle discipline est apparue de plus en plus intéressante dans des perspectives autre qu'une compilation de savoir ; de fait, les applications possibles étaient – et sont toujours – bien réelles et ont intéressées des acteurs tout autres qu'académiques. Que ce soit dans l'industrie ou encore dans l'armée, ces prometteuses perspectives méritaient que l'on y investisse... ce qui fut fait... de nouveaux moyens, un nouvel engouement, tout était réuni pour que le TAL s'épanouisse. Enfin, la dimension syntaxique étant peu ou prou domptée, celui-ci s'est intéressé à

capter l'information portée par les textes dans le but d'atteindre leur sens. Aussi, la dimension sémantique est aujourd'hui prépondérante dans les travaux de TAL.

Or, si le sens n'est déjà pas chose aisée à traiter en linguistique, cette difficulté n'a pue que se transposer en TAL : les problèmes s'accumulent à tous les niveaux, en linguistique le sens repose sur chaque domaine : lexicale, sémantique, pragmatique et syntaxique – de plus, ce concept est présent dans plusieurs écoles de linguistique avec ses acceptions propres – enfin dans différents domaines connexes à la linguistique (psychologie, science cognitive, philosophie du langage, etc.).

Néanmoins, ces phénomènes qui font le sens en langue sont certes complexes, il n'en reste pas moins analysables et dûment analysés par les linguistes depuis la conception de cette science. Aussi, il s'agira en TAL de réussir à les modéliser dans un format compréhensible par l'ordinateur.

La composante sémantique correspondra alors aux traitements permettant de représenter le sens diffusé dans un texte, ceux-ci s'appliquent sur différentes unités dont fait parti les *entités nommées*.

Elles correspondent entre autres unités linguistiques à l'ensemble des *noms propres* que l'on peut extraire d'un texte, qui sont repérés puis catégorisés selon différents groupes sémantiques : personnes, lieux, organisations, compagnies, etc. Ce à quoi est adjoint des unités numériques comme les distances ou les dates, horaires, etc. Si certaines de ces catégories se retrouvent d'un système à l'autre, elles ne sont pas toutes canoniques ; à chaque système sa classification et sa méthode. Cette multitude de classification atteste pour le moins de la difficulté qu'il peut y avoir à les catégoriser – et si le TAL a hérité des délicatesses inhérentes au traitement des phénomènes de sens de la linguistique, il y a fort à parier que celui-ci a hérité des problèmes de catégorisation du *nom propre* en *linguistique* même. C'est à cette interrogation que ce mémoire tentera modestement de s'intéresser.

Objectif et problématique

Le TAL est une discipline doublement à la croisée des chemins : l'informatique et la linguistique d'un coté, imbriquées dans la recherche et l'applicabilité de l'autre.

Aussi, nous avons choisi de prioriser avant tout une base définitoire, d'essayer au mieux de déterminer ces entités que nous sommes amenés à manipuler – de comprendre ce à quoi elles sont tributaires et, en partant de la théorie linguistique, de mettre en lumière et d'expliquer au mieux d'où peuvent en venir d'éventuels problèmes.

En ce sens, nous nous demanderons en quoi la notion d'*entité nommée* est idéale pour illustrer les problèmes des rapports du TAL avec la linguistique.

Organisation du mémoire

Ce travail comporte trois parties, la première est consacrée à la présentation du *nom propre.*

La deuxième sera quant à elle consacrée à la notion d'*entité nommée* dans une perspective TAL comme linguistique.

La troisième enfin mettra en lumière les problèmes soulevés dans les deux premières grâce à trois systèmes de repérage d'*entités nommées.*

1 Notion linguistique du nom propre

1.1 Pour un tour d'horizon de la notion de nom propre

1.1.1 Naissance de l'écriture

A la naissance de l'écriture, le langage s'est inscrit comme objet, un instantané de cette capacité propre à l'homme s'est alors capturé en un état donné. De fait, l'écriture est une révolution intellectuelle pour l'humanité, les textes sont inaltérables, l'écriture fixe la langue et donc objective l'altérité. En ce sens, les savants constatèrent rapidement que celle-ci était effectivement malléable : le grammairien indien *Pāṇini*[1] a possiblement composé son traité *Astadhyayi*[2] en constatant que le *Sanskrit* était déjà un dialecte archaïque, quoique compréhensible – cette observation suscita l'interrogation sur la langue et par extension, une volonté de la décrire, de l'expliquer – pour soit, peut-être, mais surtout dans une perspective sociétale, il s'agissait de garantir l'accessibilité de textes religieux fondamentaux. À plus d'un titre, celui-ci était révolutionnaire : il y conceptualisait alors des notions (*morphèmes*, *phonèmes*, etc.) qui n'apparaîtront dans le savoir occidental qu'au XIXe siècle.

Par ailleurs, il nous faut préciser que l'écriture n'a pas de « foyer » définitif, à chaque civilisation son moment d'invention, de développement d'une écriture et, avec elle, la naissance d'une réflexion propre sur le langage, que celle-ci soit grammaticale, rhétorique, logique ou encore philosophique. Ainsi, les philosophes grecs aussi firent ce même constat de mouvance du langage, que ce soit par confrontation à la pluralité des dialectes helléniques ou encore par l'écart entre la langue poétique d'Homère et la leur.

Notons enfin que la naissance de l'écriture pose nécessairement des problèmes d'ordre sémiotique qui, au regard de nos acquis actuels, peuvent sembler triviaux – est-ce qu'il faut séparer la phrase en unité ? Si oui,

1Nous ne savons pas le siècle précis où il vécut, vraisemblablement entre le VIIe et le IIIe siècle avant Jésus-Christ

2 Aṣṭādhyāyī de Pāṇini [PAN-6]

comment ? Quel nom lui donné ? Qu'y a t-il derrière ? Etc. Aussi, cette seule réflexion sur la segmentation s'inscrit déjà dans une démarche métalinguistique. En ce sens, *Antoine Meillet*[3], historien de la langue et de la grammaire, faisait remonter la naissance de la linguistique à la naissance de l'écriture dans la mesure où un système d'écriture témoigne de connaissances de structures linguistiques, c'est-à-dire de phonogrammes et de logogrammes. Or, s'il y a là démarche métalinguistique, il est tout à fait possible d'affirmer qu'il s'agit, au préalable, d'une démarche linguistique. Ainsi, par l'écriture, la réflexion sur la langue naquit. L'homme la considéra comme un objet d'étude à part entière et s'intéressa alors à la descriptions des phénomènes qui lui sont propres. Cet intérêt culmina au XXe siècle avec la formalisation de la linguistique, une discipline scientifique s'intéressant à l'étude de ces-dits phénomènes langagiers selon une approche strictement descriptive – il s'agit pour le linguiste de s'intéresser et *a fortiori* d'étudier les mécaniques des langues.

1.1.2 Un système fonctionnel universel

Dés lors, il a été constaté que, bien que plurielles et diverses, les langues fonctionnent néanmoins toutes à travers un même processus, elles renvoient à la réalité qu'elles visent à décrire à travers un objet linguistique : le signe. Il s'agit d'une entité tri-dimensionnelle qui permet à la langue de construire un système de description du monde reposant sur trois éléments fondamentaux : le **signifiant**, soit l'objet concret, le mot – *symbol* (selon Ogden & Richards) – l'*objet acoustique* (selon Saussure). Celui-ci renvoie à un **signifié**, c'est-à-dire le *concept*, « l'*objet mental* » auquel celui-ci fait penser ; propre à chaque culture – ou plus précisément peut-être, à chaque individu. Enfin, le *signifiant*, employé en situation d'énonciation renvoie à un objet – qu'il soit concret ou abstrait – ou à une action précise. Cet objet ou action du monde qu'il dénote est le **référent**. [SAU10]

Le processus de référence est fondamentalement lié au contexte. À titre d'exemple, le signe « chien » a le potentiel de renvoyer à l'intégralité des objets du monde descriptibles par celui-ci (un labrador comme un chihuahua), la distinction ne se fera alors que par l'à coté que ce soit dans la langue –

3 Linguistique historique et linguistique générale [MEI21]

cotexte – ou dans la situation d'énonciation – *contexte*. Aussi, la très large majorité du lexique, et ce en français comme dans les autres langues du globe, fonctionne ainsi.

Or, il existe aussi des signes plus particuliers – encore une fois retrouvés dans chaque système langagier, censés renvoyer à un objet unique. Nous les appelons communément les noms *propres*.

1.2 Le nom propre

Il n'est pas si aisé que cela de définir précisément ce qu'est un *nom propre* – plusieurs problèmes rentrent en ligne de compte ici. À première vue, cette notion semble acquise, il n'y aurait pas grand-chose à en dire sinon qu'elle réfère à un objet du monde. « *En règle général, l'enquête se clôt sur leur rôle référentiel, lorsqu'après avoir souligné que leur seule tâche est de référer à des individus particuliers le linguiste pense avoir tout dit ».[KLE99]*

1.3 Une notion « primaire »

Dans un premier temps, il s'agit généralement d'un concept « acquis » dés l'école primaire : le nom *propre* est à comprendre en quelque sorte comme en négatif du nom *commun*. Il fonctionnerait de manière similaire – c'est à dire qu'il aurait la capacité de désigner un objet du monde ; à la différence prêt, apparemment simple, que contrairement au nom commun il renvoie à un objet *unique :* en général, une personne, un individu ou encore une ville. Or, comme la plupart des concepts grammaticaux traités dés l'école primaire ou le collège, ils ne sont essentiellement évoqués que dans l'idée d'offrir une base à la compréhension orthographique. Ceci acquis, il n'est pas courant de revenir dessus ; d'autant plus lorsque l'on ne se destine pas à la vocation grammaticale ou linguistique.

1.3.1 Une notion pour plusieurs disciplines

1.3.1.1 Le *nom propre* dans les premières grammaires

La notion de *nom propre* est systématiquement présenté dans toutes les premières grandes grammaires, et ce dés qu'il faut présenter les noms à l'intérieur des parties du discours. Il s'agit du reste d'une notion très ancienne que l'on retrouve notamment dans la grammaire grecque : Ὄνομα κυριων[4] qui sera alors transposé *tel quel* en latin par nomen proprium qui se traduit littéralement comme « non à proprement parlé » soit le nom qui fait « acte de nommer ». Cela dit, cette notion a fini par glissé peu à peu de ce sens à celui de « *nom qui appartient en propre à un individu*. » [GAR91]

Qui plus est, M-N, Gary-Prieur remarque que la notion de *nom propre* ne fait que rarement l'objet d'une étude spécifique et ce dans plusieurs grammaires notoires :

- la *Grammaire Larousse du Français Contemporain* (Chevalier, J-C., Arrivé, M., Benveniste, B., Peytard, C.)
- le *Code du français Courant* (Bonnard, H.)
- la *Grammaire du français classique et moderne* (Wagner, R-L. Pinchon, J.)

Généralement, cette notion se comprend de la même façon de ce que nous sommes amenés à apprendre en CE1, c'est-à-dire une notion qui est ce que n'est pas un nom *commun*. Ainsi, le nom a deux catégories et le classement est vite réalisé. Néanmoins, la nomenclature que les auteurs dressent dans ces grammaires pour catégoriser les substantifs ne fonctionnent tout simplement pas : « *si variés qu'ils soient, les substantifs n'en constituent pas moins une espèce bien définie. Son unité est de nature morphologique. Appartiennent à l'espèce des substantifs tous les mots :*

a) qui ne tiennent que d'eux-mêmes leurs marques de genre et de nombres

b) qui s'appuient sur des déterminants spécifiques. »[W&P92]

4 onoma kurion

Pourtant, certains noms n'ont pas de genre spécifié : le prénom *Camille* peut tout autant être féminin que masculin par exemple. En outre, tous cas contrevenants à leurs observations seront rangés comme exceptions – ils tentent par exemple de dresser une liste exhaustive des noms propres pluriels.

Ce n'est qu'à partir de la *Grammaire d'aujourd'hui [AGD89]* qui se voit développée un chapitre sur le *nom propre* seul, en insistant spécifiquement sur des unicités syntaxiques : l'article devant certains noms propres : l'*Allemagne*, *la Meuse* – qui sont stabilisés alors dans une catégorie précise, les *noms propres* géographiques incluant les noms de régions, pays, rivières, etc. Ils notent cela dit quelques cas inhabituels comme *le grand César*, en définitive, ils dressent un inventaire à porté exhaustive des possibles usages et occurrences de cette notion, relevant certains cas inhabituels et soulignant la difficulté que l'on peut avoir à les classer.

1.3.1.2 Le *nom propre* dans les dictionnaires

Globalement, nous trouvons assez peu d'occurrences de noms propres dans les dictionnaires classiques ou alors, ceux-ci font l'objet d'une partie spécifique. En tout cas, les deux ensembles ne semblent mêlés que dans les encyclopédies. Néanmoins, nous les trouverons essentiellement sous la forme d'antonomase dans les dictionnaires classiques, donc de *nom commun* : *un casanova*, *une poubelle*, etc. Par ailleurs, il faut souligner que l'existence des dictionnaire de noms propres est censé palier à ce manque ; au passage, cela officie en quelque sorte la séparation de ces entités du reste du lexique. L'idée étant qu'un terme issu d'une antonomase incomprise, ou des adjectifs construits par dérivation substantives : *mccarthysme*, *macronisme*, *saussurien*, etc. puisse être trouvée par le lecteur d'un dictionnaire l'éclairant ainsi sur l'histoire d'un lexème donné. Par ailleurs, ces dictionnaires devraient vraisemblablement être qualifiés comme encyclopédies en tant que la compréhension d'un *nom propre* se fait – dans le cas de ces dictionnaires – à travers une connaissance de monde. Aussi, l'objectif de ces dictionnaires est précisément d'apporter cette *connaissance du monde* potentiellement absente chez le lecteur – celui-ci aurait été confronté à une entité qu'il n'aurait jamais vu, ou plutôt dont il n'aurait jamais entendu parler. Enfin, il nous faut noter

que les dictionnaires de nom propre format papier ont perdus de leur superbe à l'avènement d'internet. En effet, qui est confronté à une entité inconnue peut tout simplement l'entrer sur *google*[5] et une page wikipedia indiquera *a priori* tout ce que celui-ci à besoin de savoir.

1.3.1.3 Le nom propre comme entité transdisciplinaire

Dans un second temps, le *nom propre* est une entité *transdisciplinaire* : on le trouve notamment en *sociologie*, en *psychologie* ou encore en *philosophie du langage*. Aussi, nous sommes amenés à nous demander si le sens de ce concept au sein de chacune de ces disciplines est tout à fait similaire. Qui plus est, la linguistique est une discipline relativement jeune – malgré le fait que, comme nous l'avons évoqué, la réflexion sur le langage date de l'écriture, la formalisation de cette réflexion est quant à elle plus récente (XIXe siècle) – en cela la linguistique est tributaire de ces disciplines connexes tant dans sa méthodologie que dans sa terminologie. Seulement, puisant dans ces métalangues qu'elle réunit dans un tronc commun « la terminologie linguistique est un ensemble de métalangues caractérisées par des méthodologies différentes. » [LER06], la polysémie risque immanquablement de s'y glisser. En cela, si un terme « vise à réunir les conditions maximales de transparence sémantique, et à établir un rapport de référence directe et univoque avec son domaine. » [LER06], la monosémie et l'univocité dans un métalangage terminologique ne restent bien souvent que des idéaux et, malheureusement, la réalité y déroge. En somme, ce concept peut avoir des acceptions différentes autant en dedans qu'en dehors de la réflexion linguistique.

Pour tenter d'y voir plus clair, il nous faut s'intéresser à ce qui fait traditionnellement la différence entre nom *commun* et nom *propre*, aussi Jean-Louis Vaxelaire [VAX16] relève 4 critères discriminants traditionnels.

- Premier critère discriminant : la **majuscule** :

  Selon *A. Frontier* « il suffit d'une majuscule pour transformer un nom commun en nom propre ». Il s'agit d'un critère acceptable pour les alphabets européens, *pierre* vs. *Pierre* – et encore que l'Allemand fait

5 si tant est que leur clavier le permet – du reste, il y a toujours une possibilité de contourner cette problématique relativement facilement

usage de majuscule *intraphrastique* pour des mots qui ne rentrent pas nécessairement dans la catégorie du *nom propre*. Qui plus est, les alphabets Arabes ne connaissent pas la majuscule. Enfin, cette notion est absolument absurde concernant les *idéogrammes* ou les *kanji*.

- Deuxième critère discriminant : l'**intraduisibilité** :

  Selon *Witold Mańczak [MAN91] l'intraduisibilité* serait le critère définitoire du nom *propre*. Or, cela dépend des cultures et des époques, les noms des partis politiques se traduisent par exemple en français – *CDU* : *Union des Chrétiens Démocrates*. Du reste, certains noms de villes ou de pays ne sont pas à proprement parlés traduit mais adaptés à une langue – voir une culture – donnée ; il s'agit néanmoins de phénomènes compliqués à déterminer car certainement motivés par des faits extra-langagiers (*historique* notamment). *Aachen* (en *allemand*), *Aix-la-Chapelle* (en *français*). Notons également les appellations de pays : *Deutschland* (en *allemand*), *Allemagne* (en *français*), *Niemcy* (dans les parlers *slaves*), *Tyskland* (dans les parlers *scandinaves*), *Németország* (en *hongrois*). En l'occurrence, c'est ici motivé par les contacts historiques qu'avaient ces populations avec les tribus germaniques qui leurs étaient limitrophes, à l'ouest : les *allamans*, au nord les *tudesques* (sous la forme : *tysk* en Suède, Danemark et en Norvège), etc. Par ailleurs, *Mańczak* soulève que certaines traduction de *nom propre* d'une langue à l'autre serait plutôt lié à une plus prosaïque adaptation phonétique : « *Par exemple, un nom commun comme* ville *est traduit en italien par* città*, en anglais par* town*, en allemand par* Stadt, *tandis qu'un nom propre comme* Paris *ne l'est pas, cf. it. Parigi, angl. Paris ou all. Paris. Alors que le nom commun* ville *est traduit dans les langues en question, le nom propre* Paris *ne l'est pas ; il ne fait que subir une adaptation phonétique et morphologique.* » *[MAN91]*

- Troisième critère discriminant : l'**absence** ou **présence** de **déterminants** :

  Celui-ci revient régulièrement, il semble particulièrement efficace en langue *anglaise* et *a priori* également en *français*. Or, il existe des cas où

cela ne se vérifie pas, quelques locutions plus archaïques comme *la Marie est parti chercher le pain*, ou encore les noms de périodes historiques qui s'emploient pour la très grande majorité avec un article : la *Renaissance*. Encore une fois, ce critère dépend fortement des langues : à titre d'exemple, le Russe n'a pas d'article ; le Grec (ancien comme moderne) utilise quant à lui des articles aussi bien apposés aux noms *propres* qu'aux noms *communs* Ὁ Σωκράτης et ὁ ἄνθρωπος[6] – celui-ci pourra alors avoir une valeur *outil* en participant à la construction d'un c*omplément attributif*, et son absence sera également porteuse de sens : elle donnera au nom une valeur indéterminée.

- Quatrième critère discriminant : l'***asémentisme*** ou *l'**absence de sens***

  Qu'est-ce qui est *dans* un nom ? Un nom propre ne signifierait rien, en ce sens la seule dénomination d'une personne – *Jacques* par exemple – n'indiquerait en rien ce à quoi elle ressemblerait. *Vendryes* disait « les noms propres n'ont d'autre signification que de désigner une personne ou un lieu. Aussi sont-ils en principe intraduisibles d'une langue à l'autre. Que veut dire *Alexandre*, *César* ou *Napoléon* ? Le contenu de ces noms a pour limite la personnalité qu'ils désignent. Pour ceux qui n'en ont jamais entendu parler, ils sont littéralement vides de sens. » [M&V63] Cette idée suggère que le sens d'un nom *propre* serait strictement lié au *référent* qu'il désigne ; or cela renvoie plutôt à la théorie de l'*arbitraire du signe linguistique* – celui-ci peut être motivé pour essayer de copier, de reproduire une caractéristique de l'objet qu'il désigne (les *onomatopées* par exemple) mais il peut tout autant être immotivé. Qui plus est, il faut également prendre en considération l'aspect culturel d'un patronyme – chaque culture à ses appellations qui leur sont typiques et qui peuvent parfois poser des problèmes de transcription – on se perd souvent à la lecture d'un roman russe tant les patronymes peuvent différer selon le lien que les personnages exercent entre eux, la formalité de la situation, etc. De plus, certains noms évoquent instantanément une « aire culturelle » en quelque sorte, bien que cette tendance semble fluctuer, du reste, il semble admissible d'affirmer qu'il ne s'agit en rien d'une tendance langagière et qu'en

6 *O sokratos et o antropos : « Le »* Socrate et *L'*homme

conséquence nous quittons vraisemblablement là le domaine de la linguistique. Néanmoins, il faut rappeler qu'un nom se comprend en relation à l'entité à laquelle il réfère, en ce sens, pour qu'un locuteur d'une langue donnée utilise un nom propre, cela présuppose qu'il ait une connaissance du référent en question. En ce sens, Kerstin Jonasson [JON93] souligne le caractère déictique des noms propres : leur compréhension passe systématiquement par un renvoie à une situation ou aux connaissances partagées entre les locuteurs. De fait, nous employons très régulièrement le nom de nos amis, de nos proches, des villes et quartiers – connus ou encore inconnus, pour le cas d'une banlieue par exemple – et ce sans avoir aucunement besoin d'un recours quelconque à un recueil compilant l'intégralité des *noms propres* possibles et imaginables. En définitive, le contexte sera généralement toujours évoqué pour permettre la compréhension d'un *nom propre* potentiellement inconnu.

Ces critères, même associés, ne semblent pas suffire pour définir ce qu'est un *nom propre* – aucun d'eux ne saurait être utilisés en tant que critère universel référentiel.

1.3.1.4 Le *nom propre* en *linguistique*

Dés les premiers pas de la linguistique saussurienne, il n'y a presque aucune mention de la notion de *nom propre* mis à part une rapide allusion concernant l'analogie « *les seuls formes sur lesquelles l'analogie n'ait aucune prise sont naturellement les mots isolés tels que les les noms propres, spécialement les noms de lieux (cf. Paris, Genève, Agen, etc.) qui ne permettent aucune analyse et par conséquent aucune interprétation de leurs éléments »[SAU16].* Du reste, le système tripartie – que nous avons vu ci-dessus – proposé par Saussure rencontre un problème de taille concernant la notion de *nom propre*. De fait, celle-ci ne peut techniquement pas être intégré au système de la langue dans la mesure où un *nom propre* n'a pas de signifié. Conséquemment, il sera traité comme simplement isolé et inanalysable. À l'inverse, Jakobson intègre lui les noms propres au système de langue qu'il théorise : selon sa thèse, trois types d'expressions linguistiques possèdent une structure double :

- le *discours cité* : le **message** renvoie au **message**,
- les *embrayeurs* : le **code** renvoie au **message**,
    - ou à l'inverse : le **message** renvoie au **code**.

Ainsi, le *nom propre* fonctionnerait comme des signes qui ne peuvent se définir en dehors d'un renvoie au code « dans le code de l'anglais 'Jerry' signifie une personne nommée Jerry. » [JAK63].

Enfin, quant aux thèses plus récentes, la distinction entre le nom *propre* et le nom *commun* procède d'abord du domaine de la *logique* avant d'avoir été repris par la *grammaire*.

Un problème émerge alors car une influence importante[7] des théories *logiques* sur le nom *propre* en obscurcit les spécificités *linguistiques*. En ce sens, cette origine extérieure pourrait potentiellement expliquer l'absence d'une définition unanime chez les linguistes ; de fait, le transfert d'un concept d'une discipline à une autre n'est généralement pas chose aisée – peu sont fructueux. Quoi qu'il en soit, certaines observations liées à cette approche logique peuvent éclairer et apporter des indications intéressantes : *Socrate* renvoie par exemple à un individu par le biais d'une *chaîne causale*. Cela indique que la désignation du *nom propre* en tant que tel ne repose pas uniquement sur des critères *morphologiques*, ni sur des critères entendus de *syntaxe*. En ce sens, l'approche *aristotélicienne* – aussi appelée « *approche classique* » ou encore « *approche des conditions nécessaires et suffisantes* », déclare que les propriétés qui sont partagées par les entités en question sont la base du regroupement. Conséquemment, pour estimer si une entité quelconque fait parti d'une catégorie, il est nécessaire qu'elle possède les propriétés de celle-ci – conditions *nécessaires* – et il est suffisant qu'elle les possède pour y appartenir – conditions *suffisantes*. Une fois le processus terminé, l'élément appartient ou non à la catégorie en question. Par ailleurs, il est possible par une expression type : '*X* est *Y*' d'inférer que '*X*' possède les propriétés correspondantes à '*Y*'.

7 De Kripke notamment

1.3.1.5 Le nom propre en sémantique

Deux thèses majeurs s'oppose concernant la notion de *nom propre* en sémantique :

- de nombreux sémanticiens partent du postulat que le nom propre n'a pas de sens et qu'en conséquence il n'appartient pas à la sémantique.

  « *Les noms propres n'ont pas de* sens *et, par conséquent, la notion de signification ne s'applique pas à eux. La fonction d'un nom propre est l'identification pure : distinguer et individualiser une personne ou une chose à l'aide d'une étiquette spéciale.* »[ULL52]

- d'autres stipulent que le *nom propre* est au contraire à exclure de la sémantique dans la mesure où celui-ci aurait *trop* de sens.

  « *Si l'on classait les noms d'après la quantité d'idée qu'ils éveillent, les noms propres devraient être en tête, car ils sont les plus significatifs de tous, étant les plus individuels. Il suffit de rapprocher le mot* César*, entendu de l'adversaire de* Pompée*, et le mot allemand* Kaiser*, qui signifie « empereur » pour voir ce qu'un nom propre perd en compréhension à devenir un nom commun. »* [BRE24].En ce sens, cette thèse rapproche le *nom propre* de ce qu'il était dans la grammaire grecque, la dénomination absolue.

Aussi, Jesperssen précise la thèse de *Bréal* en affirmant que le sens d'un nom propre dépend de son emploi « *Dans un cas, le mot* pipe *désigne une* pipe *que l'on fume, dans un autre un tuyau de canalisation, ou bien un porte-voix, ou encore un tuyau d'orgue. De même, le mot* John *prend un sens différent chaque fois qu'il est employé, et seul le contexte permet de le découvrir : le fait que ce sens soit plus spécialisé dans chacun de ces cas de celui de* pipe *tient à ce qu'un nom propre évoque un plus grand nombre de traits particuliers qu'un nom commun* ». [JES05]

Ces deux thèses conduisent à deux acceptions possibles du *nom propre* concernant l'approche *sémantique* : dans un premier temps, le *nom propre* a effectivement du sens et, visant un individu, celui-ci a plus de propriété sémantique qu'un nom commun ; en somme, la compréhension augmente alors que l'extension diminue. Par ailleurs, cela nécessite de l'adjoindre à la

catégorie du *nom commun* – donnant un sens plus fonctionnel à cette traditionnel opposition.

Aussi, si la notion de *nom propre* peut être considéré d'un point de vue sémantique comme une étiquette apposée à un individu dans le but de le différencier d'une multitude d'autre, il existe des cas où l'interprétation des *noms propres* peut s'avérer plus complexe que ce que nous pourrions supposer. Aussi, celui-ci peut dans certains cas exprimer le *thème* du *discours*, sa signification sera alors bien plus importante que celui de simple prédicat de dénomination : *Belleville, c'est terminé* – ici, *Belleville* n'est pas à comprendre comme la dénomination d'un quartier de *Paris* par opposition aux autres mais à ce que ce référent a pu être, a pu incarner, spécifiquement pour le locuteur. Dans un même genre mais quelque peu différent, *Tu la vois encore cette Jeanne* – notons que pour référer à l'individu *Jeanne*, l'utilisation du déterminant démonstratif n'est pas nécessaire, cela permet par contre au syntagme nominal d'apporter une signification en plus. Dans cet esprit, Damourette et Pichon disait que « *sans que l'extension sémantique du nom propre soit aucunement changée, il est distingué, à l'intérieur même de sa personnalité, plusieurs espèces en raison des points de vue divers auxquels on peut l'envisager* » [D&P39]. Du reste, G. Kleiber suggérait qu'il serait possible de chercher ce qui fait la différence entre nom *commun* et *propre* au niveau du sens des déterminants – dans les cas où, bien sûr, ceux-ci s'utiliseraint avec un *nom propre*.

Enfin, d'après F. Rastier dans son ouvrage *L'analyse thématique des données textuelles* [RAS95] le *nom propre* relèverait, au même titre que le reste du lexique, de la *Norme* tel que défini par E. Coseriu c'est-à-dire « *un espace intermédiaire entre la Langue et la Parole* » [COS52]. La *Norme* contient ce qui dans la *Parole* est répétition de faits déjà réalisés, mais elle est aussi façonnée et modifiée par les normes au sens large, c'est-à-dire sociales, religieuses, etc. En ce sens, les habitudes culturelles d'appellation créeraient des « dictionnaires » de *nom propres*. Ceux-ci peuvent évoluer selon l'époque, les habitudes de ceux qui les forment et les entretiennent – à titre d'exemple, la liste des noms les plus courants a significativement évolué d'il y a 50 ans à nos jours.

1.3.1.6 Observation du *nom propre* dans la syntaxe

Quel statut pouvons-nous donner au *nom propre* ? Dans la grammaire, le *nom propre* a le statut d'un nom mais en considérant qu'il peut seul saturer l'argument d'un verbe, il peut donc avoir le statut de syntagme nominal. Aussi, la démarche des grammairiens et linguistes consiste à définir l'usage de l'objet linguistique « nom » selon des couples d'oppositions :

- noms communs vs. noms propres,
- noms comptables vs. noms non-comptables,
- etc.

Ainsi, certains linguistiques se sont évertués à poursuivre cette méthodologie en soulevant plusieurs sous-catégories à l'intérieur même de la catégorie du *nom propre* :

- *nom propre* animé (**Jean**) vs. *nom propre* non-animé (**Paris**)

Aussi, il a été souligné que certains usages peuvent faire se comporter le *nom propre* comme un nom abstrait *Ah, c'est signé Martin ça !*.

Par ailleurs, un des éléments *discriminant* que nous avons vu ci-dessus est la *majuscule*, celle-ci se retrouve dans quatre éléments en langue *française* : le *titre*, la *phrase*, le *vers*, le *nom propre* ; d'un certain point de vue ils constituent un tout, aussi bien sur le plan de la forme que du sens. Du reste, ils peuvent constituer un tout sur le plan de la forme seule (le *vers* poétique par exemple) autant que du sens seul, comme un nom *commun* qui avec une majuscule peut aisément faire référence à un *nom propre* : l'*Empereur* pour *Napoléon* par exemple. Du reste, si ce critère formel de *majuscule* peut sembler être à étudier dans un cadre syntaxique, il semble que l'analyse qu'on peut en tirer est d'ordre sémantique.

1.3.1.7 La communisation

Il s'agit du terme employé par les linguistiques pour référer à un *nom propre* dont la construction le rapproche d'un *nom commun* ; aussi, l'absence d'un statut syntaxique clair pour définir l'opposition entre ces deux catégories s'explique certainement par ce phénomène : « *linguistiquement parlant, il est impossible de tracer une ligne de démarcation rigoureuse entre les noms*

propres et les noms communs. Nous avons vu les cas où l'on passe insensiblement des uns aux autres, mais le cas inverse est tout aussi fréquent. » [JES05]

Cette idée peut se justifier en la considérant selon une perspective synchronique : *Jean Charbonnier* pourrait s'appeler ainsi car un ancêtre de cet individu devait vraisemblablement occuper la profession de charbonnier lors d'un recensement quelconque. Jesperssen fait reposer cela sur une propriété inhérente à la catégorie du nom, car il est effectivement possible de faire basculer un nom d'une catégorie lexicale à l'autre : un nom *massif* peut devenir *comptable* et inversement,

Je vais au supermarché m'acheter une eau, je reviens

Eh bien, il y a du policier par ici !

Il est également possible pour un nom *abstrait* de devenir *concret* et ce grâce au contexte syntaxique, *ils ont fait du sale*. Enfin, ce changement peut s'opérer d'une catégorie lexicale à l'autre : un nom peut notamment s'adjectiver, *il fait très étudiant assis comme ça avec son livre et ses lunettes*, et à l'inverse, un adjectif peut se substantiver, *le grand – l'inénarrable.*

Quoi qu'il en soit, la transition de la catégorie de *nom propre* à *nom commun* ne peut s'expliquer par la détermination, nous l'avons vu, certaines expressions archaïques autorisent déjà ce genre d'usage ; en somme, deux phénomènes différents seraient considérés ici sur le même plan.

Aussi, quand bien même le phénomène de *communisation* permet une explications de nombreux cas exceptionnels du *nom propre*, il ne s'agirait pas de risquer d'en faire une catégorie par défaut où chaque cas problématiques serait consignés sans autre forme de procès. À noter que ce phénomène peut lui-même être sous-catégorisé, comme le soulève les exemples suivant :

- (i) *Je montais tristement la garde sur une Méditerranée houleuse*
- (ii) *Un Auguste peut faire des Virgiles*

Dans l'exemple (i) malgré la *communisation*, le nom renvoie toujours à un référent unique, qui sera systématiquement la mer Méditerranée - . Par contre dans une formule comme (ii), ces deux *noms propres* renvoient à des référents plurielles et la *communisation* opère d'une certaine façon comme

une comparaison en interne, un *Auguste* comme un *dirigeant influent et mécène à ses heures perdues* et des *Virgiles* comme *de potentiels artistes talentueux à qui il ne manquerait que le financement*, aussi : ceux-ci ne renvoient évidemment pas au référent communément admis d'*Auguste* ou de *Virgile*, cela chaque locuteur le comprend – du moins ceux ayant cette connaissance du monde même si cet acquis n'autorise qu'une interprétation de cette énoncé.

Ainsi, deux problèmes potentiels se révèlent ici qui peuvent devenir problématique dans une perspective de formalisation des méthodes – un problème qui pourrait du reste dépasser la simple frontière de la théorie *linguistique*, notamment dans une perspective d'application *TAL* :

- Simplification excessive qui cantonne la syntaxe du nom propre à un nom commun sans déterminant ; chaque usage allant à l'encontre de cette observation sera traité comme une exception.
- Le *nom propre* n'est qu'assez rarement étudié dans une perspective strictement linguistique – ainsi, la définition du *nom propre* se limiterait à un aspect référentiel, qui sera étudié dans ce devoir, ainsi le *nom propre* ne se comprendrait qu'à travers le lien référentiel.

Pour répondre à ces deux soucis, Marie-Noëlle Gary-Prieur propose que l'on « *[renouvelle] la problématique des noms propres, enlisée depuis trop longtemps dans les mêmes débats généraux (le* nom propre *a-t-il un sens ? Le* Nom Propre *est-il un signe?). Un renouvellement ne peut venir, à mon sens, que d'une reformulation des problèmes à partir d'une idée naïve mais féconde issue de la grammaire générative : puisque nous comprenons que nous employons des énoncés comportant des noms propres, la grammaire du français doit non seulement reconnaître le* nom propre *comme un objet de langue, mais aussi se fixer comme tâche de décrire les mécanismes d'interprétation qui lui sont associés en fonction de ses constructions. » [GAR91deux]*

Nous pouvons en effet considérer qu'une réflexion sur la méthodologie d'étude du *nom propre* pourrait sortir la notion de l'impasse dans laquelle elle semble s'être lové. Si la *logique*[8] a pu avoir une influence sur la *linguistique*,

8 notamment *kripkienne* – à travers le concept de *désignateur rigide*

c’est bien au travers de sa méthodologie d’observation du *nom propre* prit pour lui-même c’est-à-dire isolé de tout autre élément de phrase. Or, si cela est tout à fait cohérent dans une perspective *logicienne*, il se pourrait que cela s’avère contre-productif en *linguistique*. Une influence que nous pouvons voir à travers l’usage des exemples prototypes de *nom propre* dans les grammaires :

Jean aime Marie. Marie mange du chocolat, etc.

ceux-ci mettent l’accent sur un usage entendu du nom propre – très régulièrement le nom d’une personne ou d’un lieu, ne traitant alors jamais des cas qui bousculerait ce qu’elles affirment – ce qui est malheureusement souvent le propre des grammaires : beaucoup d’exemples jouets. Il est pourtant suffisamment commun de construire des *noms propres* avec détermination dans un langage moins policé, plus vernaculaire :

Les Jeans sont tous comme ça… Jean ? Mais de quel Jean tu parles ?… Ce jean ? Ah !.

Par ailleurs, des formes de *noms propres* peuvent fréquemment s’utiliser seules dans des positions non-référentielles, comme nous l’avons évoqué, dans une interprétation plus métaphorique, notamment dans des expressions que nous aurions qualifié de figées sans remarquer qu’elles peuvent susciter un patron de construction type en fonction d’un contexte donné :

c’est un vrai Taj mahal !

Une litote dont le modèle peut s’exporter à peu près partout, devant une tour peu impressionnante par exemple :

c’est une vraie Tour Eiffel !, etc.

Enfin, le nom propre s’utilise régulièrement à la manière d’un adjectif qualificatif et, en conséquence, en partage les propriétés sémantiques :

j’adore ton sac, il fait très chanel.

Ça c’est une description Balzac dis donc.

Du reste, il faut noter que l’étude syntaxique du *nom propre* peut occasionner des ambiguïtés en interne :

la critique de Paul : *Paul* a émis une critique vs. *Paul* a essuyé une critique

Il n'est pas tout à fait évident de saisir si Paul est victime ou émetteur d'une critique dans ce genre d'énoncés.

1.4 Limites et perspectives

1.4.1 limites

Les approches que nous venons d'étudier sont pour la très large majorité propres au français, aussi toute observation à caractère généralisant devrait être confrontée à d'autres langues pour vérifier si celles-ci ont une quelconque portée universelle. De fait, si l'opposition *nom commun* vs. *nom propre* peut sembler fonctionnelle en *français*, il n'est est rien dans une langue telle que le chinois puisque « *seul le contexte permet de séparer Np et Nc, les deux étant issus des mêmes stocks lexicaux.* »[VAX16]

Au demeurant, le concept du *nom propre* couvre de nombreuses sous-catégories comme le remarque Jean-Louis Vaxelaire « *d'un point de vue épistémologique, l'intérêt est également de démontrer que la classe est bien plus large qu'on ne le pense au premier abord : anthroponymes et toponymes y côtoient des tires de roman, des noms d'objets, de partis politiques, d'ouragans, d'événements historiques, etc. Cette hétérogénéité se répercute également sur la forme des Np qui peuvent aller de la simple lettre (le chanteur M) à un syntagme large (Mais qu'est-ce que j'ai fais au Bon Dieu pour avoir une femme qui boit dans les cafés avec les hommes?).* » [VAX16] mais les études semblent essentiellement s'intéresser aux cas prototypiques de la notion de *nom propre*. De fait, cela peut potentiellement poser un problème endémique, si au plus bas de l'étude le *nom propre* n'est réfléchi qu'à travers ses occurrences les plus évidentes (*patronyme* et *anthroponyme*), il est tout à fait possible que cette notion soit considérée comme un acquis sans plus d'interrogation « Dans les travaux sur les Np, il est extrêmement rare de définir son objet d'étude, comme si la question ne se posait pas même pas »[VAX16].

Du reste, ces études restent particulièrement *occidentalo-centrées* dans la mesure où elles postulent des règles à porté générale sur cette notion à partir de propriétés qui se trouvent essentiellement dans les noms occidentaux. Par exemple, l'une des propriétés du *nom propre* est de ne pouvoir être facilement modifié – ce qui n'est pas un fait universel, certaines cultures admettent une évolution du nom en fonction des stades de la vie d'un individu (*enfance, mariage,* etc.) qui plus est, si le *patronyme* occidental se caractérise par cette fixité, il ne semble pas raisonnable d'affirmer que cela soit la conséquence d'un quelconque critère linguistique, il s'agirait plus vraisemblablement d'une conséquence de l'état civile qui nous empêcherait de trop facilement en changer.

1.4.2 Perspectives conceptuelles

Ainsi, une approche linguistique précise de ce concept reste à être entreprise. D'un point de vue méthodologique, il apparaît nécessaire de l'observer en tant qu'objet d'étude linguistique, c'est-à-dire dans son emploi textuel « Encore la lexicologie et la syntaxe relèvent-elles de la linguistique, mais l'étude des textes se trouve généralement dévolue à d'autres disciplines, poétique, sémiotique, herméneutique, etc. Si la linguistique restreinte, centrée sur la morphosyntaxe, domine encore, nous entendons prouver le mouvement en marchant, montrer que le texte est irréductible à une suite de phrases ; mieux, qu'il constitue non seulement l'objet empirique, mais l'objet réel de la linguistique » [RAS89]. Enfin, il est nécessaire de faire l'inventaire des multiples formes qu'il peut prendre, nous avons vu que le *nom propre* peut à lui seul occuper la fonction de *syntagme nominal*, ainsi, nous pouvons nous interroger sur ses frontières, dans un énoncé comme

Ville de Paris

est-ce que celui-ci peut être perçu en *entier* comme un *nom propre* ? De fait, selon *wikipedia*, ce syntagme peut renvoyer à « *la « Ville de Paris », c'est-à-dire la collectivité à statut particulier »* ou encore à « *la mairie de Paris, c'est-à-dire l'administration dirigée par les élus de la Ville de Paris et sise à l'Hôtel de Ville de Paris* »[WIKIPARIS]. Or, celui-ci pourrait vraisemblablement tout aussi être traité comme *Paris* mettant le '*Ville de*' de coté.

Par ailleurs, les ambiguïtés sont monnaies courantes car si le *nom propre* peut devenir un *nom commun*, l'inverse est aussi vraie et bien plus systématique qu'on ne pourrait le penser – à travers les études du concept dont nous venons de proposer une synthèse, il s'agit du phénomène de *dénomination prédicative* mis en lumière dans une étude de Marc Wilmet, celle-ci « *assigne à un référent R une dénomination commune ou dénomination propre, explicitement (p. ex. L'animal domestique nyctalope digitigrade des zoologistes s'appelle un chat ou Ce philosophe s'appelle Socrate), ou implicitement (les titres : Papa, Maman, Maître, Excellence… les apostrophes : Officiers, Sous-officiers, Caporaux et Soldats… , les vocatifs en général : Chauffe, Marcel / Bonjour Facteur)* » [WIL91].

Ainsi, ce sont là d'autant plus de formes du *nom propre* suffisamment utilisées, mais insuffisamment étudiées. En conséquence, voici encore une occasion d'être confronté à des problèmes d'interprétations.

1.5 Pour une exportation du concept

S'il s'agit pour l'*entité nommées* de relever ce qui correspond dans une phrase aux *noms propres* puis de les catégoriser, nous venons de constater que cela n'est apparemment pas une mince affaire. À ce point, les problèmes sont plurielles et relèvent, du reste, de considération théorique, de problèmes de méthodologies pour une notion qui a apparemment souffert de sembler par trop évidente. Nous allons maintenant nous atteler à considérer la notion d'*entité nommée* en elle-même, et tâcherons de traiter celle-ci avec la même attention, et, au mieux, avec la même exhaustivité.

2 Notion linguistique de l’entité nommée

2.1 Contextualisation du terme

2.1.1 Une perspective historique

La perspective de pouvoir automatiquement traduire d’une langue à l’autre relevait plus du fantasme qu’autre chose. Du reste, l’homme a su contourner les problèmes d’incompréhension inhérents à la pluralité des parlers ; que ce soit la *koinè* du grec antique parlée aux quatre coins de la méditerranée – le français comme langue de la diplomatie du XVIIe siècle environ jusqu’au XXe siècle – le Grec, puis l’Arabe, puis le Latin comme langue des sciences – jusqu’à l’anglais qui, de nos jours, revêt cet aspect de langue par défaut, que ce soit à travers la diplomatie, le commerce ou encore la science. Par ailleurs, dans une perspective strictement belligérante, la barrière de la langue n’était qu’assez rarement un problème ; à titre d’exemple, la conquête du Mexique par Hernán Cortés a été facilité par *La Malinche* d’abord esclave des conquistadors puis interprète de ceux-ci auprès des Aztèques, traduisant alors de l’Espagnol au Nahualt.

En définitive, la traduction automatique n’aurait pas répondu à un besoin immédiat et ce n’est finalement qu’au XXe siècle que les premiers projets concrets de traducteurs automatiques se réalisent en France (en 1932) – *Georges Artsrouni* développe une première machine en 1932, une deuxième en 1935 – et en URSS, en 1933, *Petr Smirnov-Trojanskij* conceptualise une machine qui n’a peut-être jamais été réalisée dans la mesure où les autorités soviétiques responsables n’en voyait alors pas l’intérêt – il faut noter que la perspective applicative de ces deux projets n’étaient pas seulement de l’ordre de la traduction automatique ; le brevet déposé par Georges Artsrouni en 1935 stipule « un appareil rendant mécanique et automatique l’emploi d’horaires de chemins de fer, d’annuaires téléphoniques, de dictionnaires, etc. ». [DAU65]

2.1.2 La langue et l’informatique naissante

Néanmoins, aux début du développement des techniques informatiques, les premiers ingénieurs investissent la langue d’un intérêt tout particulier. En effet, la fameux *test de Turing* décrit dans sa publication *Computing Machinery and Intelligence* d’Octobre 1950 [TUR50] – un test pensé pour déterminer si une machine est intelligente ou non, ou plutôt si elle serait en mesure de tromper l’intelligence humaine. Du reste, il précise que cette notion de ‘penser’ associer à la machine a quelque chose d’absurde « *If the meaning of the words ‘machine’ and ‘think’ are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, ‘Can machines think ?’ is to be sought in statistical survey such as a Gallup poll. But this is absurd.’* » [TUR50] pour des raisons qui nous semblent dorénavant évidentes : la pensée est un concept dont nous avons tous une idée qu’approximative, nul n’en connaît les tenants et aboutissants. Quant à la machine, elle n’en était alors qu’à ses balbutiements et, au-delà de toutes pratiques, les réponses qu’il pouvait apporter étaient pour l’essentiel théoriques, voir même prophétiques – stipulant ainsi sur ce qu’il voyait être la technologie informatique des années à venir, ces estimations étaient parfois étonnamment précises. Néanmoins, cette article inscrit une idée qui aura sa part d’influence pour le développement des techniques informatique et, surtout, pour le domaine du traitement automatique d’une langue. En effet, il y affirme que pour qu’une machine puisse être considérée comme intelligente, elle doit pouvoir tromper une personne dans une situation de communication ; c’est-à-dire posséder des notions conversationnelles suffisamment pointues lui permettant au moins d‘émuler une conversation et, au mieux, de la produire. Aussi, cette perspective entraîna un intérêt croissant, une amorce en quelques sortes, pour cette idée : *traiter automatiquement une langue*. Aussi, dans un contexte de guerre froide, cette perspective glissa très vite dans une visée applicative, l’idée de traduire automatiquement les messages de l’ennemi était par trop tentante.

2.1.3 Le commencement des premières tâches de traitement automatique

Véritablement, le **NLP** *natural language processing* ou **TAL** *traitement automatique du langage* naît dans les années 1946 et 1947 quand *Andrew Booth* et *Warren Weaver* se rencontrent et évoquent l'idée d'utiliser les premiers ordinateurs pour la traduction de langues naturelles. Aussi les recherches en la matière se développèrent aux États-Unis au *Massachussets Institute of Technology* au tout début des années 1950 sous la direction de *Yehoshua Bar-Hillel*. Cela culmina avec l'expérience de *Georgetown* sous la direction de *Léon* Dostert en collaboration avec IBM en 1964 : un échantillon contrôlé de 49 phrases en russe ont été traduites en anglais. L'expérience utilisait alors un vocabulaire très limité de 250 mots et seulement 6 règles grammaticales. Quoi qu'il en soit, les résultats suscitèrent l'engouement de la presse étasunienne, « *Although the system had little scientific value, its output was sufficiently impressive to stimulate the large-scale funding of MT research in the USA and to inspire the initiation of MT projects elsewhere in the wolrd, notably in the USSR.* »[HUT07] et, conséquemment, stimula les ressources allouées à la recherche en traduction automatique aussi bien aux États-Unis qu'en URSS – conséquence direct des premiers temps de la guerre froide – à l'image de la conquête spatiale, cette technologie jouissait d'une certaine aura de prestige car moderne, complexe et technique – en ce sens, les subsides étaient particulièrement conséquents.

À l'origine, les recherches se scindèrent en deux approches, l'une plutôt empirique de « trial-and-error » dite « brut-force » utilisait des approches statistiques pour découvrir des régularités lexicales et grammaticales qui pourraient être exploitées informatiquement. Celle-ci avait pour but de développer des systèmes de traduction basique mais pratique. L'autre approche était quant à elle plus théorique et exploitait des recherches en linguistique fondamentale, elle marqua par ailleurs les début de la *linguistique computationnelle* – son but était de poser les bases de système de traduction précis qui aurait requis une supervision humaine minime sinon nulle, « *In most cases, the 'empiricists' adopted the 'direct translation' approach, often using statistical analyzes of actual texts to derives dictionary rules – often of an ad hoc nature, with litttle or no theoretical foundation. The*

'perfectionists' were explicitly theory-driven, undertaking basic linguistic research, with particular attention given to methods of syntactic analysis. » [HUT07]

Aussi, les espérances étaient particulièrement audacieuses : les progrès en informatique ainsi qu'en linguistique formelle – notamment en syntaxe – laissèrent envisager de grands progrès en terme de qualité. Néanmoins, l'optimise laissa vite place à un certain désenchantement quand les problèmes linguistiques apparurent de plus en plus importants et que la recherche en traduction automatique était confrontée à un obstacle d'ordre sémantique apparemment insurmontable – du moins avec les moyens dont ils disposaient.

Yoshua Bar Hillel commente dés 1953 que « *the task of instructing a mahcine how to translate from one language it does not and will not understand into another language il does not and will not understand presents a real challenge for structural linguists, in that their thesis that language can be exhaustively described in non-referential terms undergoes here an experimentum crucis. If, in a translation program, some step has to be taken which directly or indirectly depends upon the machine ability to understand the text on which it operates, the machine will simply be unable to make this step, and the whole operation will come to a full stop* » [BHI53]

Nous sommes en 1965, et déjà deux méthodologies bien différentes – critères linguistiques vs. une approche informatique pure – se distinguent.

2.1.4 Les rapports ALPAC

En 1966 sonna le glas, tous les avis convergent en cela, que ce soit Yeshoshua Bar Hillel – c'est également ce qu'atteste Marcel Cori et Jacqueline Léon dans une étude historique de la constitution du TAL « l'idée d'une traduction entièrement automatisée est déjà publiquement mise en doute, et dans un contexte culturelle et scientifique très différent : importance moins accordée à l'informatique, attitude spécifique des linguistiques et des institutions par rapport à la TA [traduction automatique]. » [COL02]. La *traduction automatique* souffre alors d'un manque de résultat satisfaisant – au mieux celle-ci à produit un système de *traduction* pour l'aviation américaine « *a system installed for the US Air Force produced 'translations' until the early 1970s. By any standards the output was crude and sometimes barely*

intelligible, but no excessive claims were made for the system, which with all its deficiencies was able to satisfy basic information needs of its users. »[HUT07], au-delà de cela, les développements ne rencontraient pas les objectifs visés malgré leurs coûts substantiels – rappelons à ce titre que ces coûts sont également fortement liés à des problèmes matériels ; l'informatique de l'époque supposait des installations particulièrement conséquentes.

Aussi, soucieux des doutes expriméss par les experts – et, surtout, des sommes investies jusqu'à présent, le gouvernement américain décida d'évaluer la rentabilité de l'affaire « *the imminent prospect of good-quality MT was receding, and in 1964 the government sponsors of MT in the United States (mainly military and intelligence agencies) asked the National Science Foundation to set up the Automatic Language Processing Advisory Committee (ALPAC) to examine the situation.* »[HUT07], aussi, la conclusion de celui-ci est sans appel, « *it concluded that MT was slower, less accurate and twice as expensive as human translation and 'there is no immediate or predicatale prospect of usefull machine translation' (ALPAC 1966)* »[HUT07] et ses effets sont dévastateurs, « *The best know event in the history of machine translation is without doubt the publication thirty years ago in November 1966 of the report by the Automatic Language Processing Advisory Committee (ALPAC 1966). Its effect was to bring to and end the substantial funding of MT research in the United States for some twenty years. More significantly, perhaps, was the clear message to the general public and the rest of the scientific community that MT was hopeless.* »[HUT97]. En conséquence, les subventions cessent nettes et le *traitement automatique* d'une langue sommeillera – du moins dans des perspectives pratiques – quelques décennie avant qu'une nouvelle technologie ne change la donne.

2.1.5 Sommeil du traitement automatique, engouement de la computational linguistic

Aussi, malgré une réduction substantielle de la pratique du *traitement automatique* lié au langage, le développement théorique perdura quant à lui au travers la *computational linguistics.* Il nous faut préciser que celle-ci est tout de même antérieur à l'*ALPAC* et n'en est aucunement la résultante, aussi elle traitait le pendant *théorique* de la *traduction automatique*, on lui doit du

reste plusieurs innovations comme notamment les premiers *analyseurs syntaxiques* au début des années 1950.

Véritablement, c'est plutôt vers les années 1960 – autour du rapport de *Yeshoshua Bar Hillel* – que cette discipline devient plus lucide quant à l'avenir de la *traduction automatique*, aussi celle-ci se recentre sur un aspect plus théorique. Si les rapports *ALPAC* ont stoppé net la *traduction automatique*, ceux-ci ont au contraire contribué à la légitimation de la *computational linguistic* « *the committee agreed at the outset that support for research in this area 'could be justified on one of two bases : (1) research in an intellectually challenging field that is broadly relevant to the mission of the support agency and (2) research and development with a clear promise of effecting early cost reductions, or substantially improving performance, or meeting an operational need* »[HUT97]. C'est à ce moment que le terme *Natural Language Processing* (NLP) – *traitement automatique des langages naturels* – commence à être employé dans les années 1980 pour s'installer durablement dans les années 1990 ; celui-ci n'y est pas assimilé à *computational linguistic* ou inversement – la dissociations des deux domaines se fait autour de leur visé, *théorie* pure vs. *praticabilité*.

2.1.6 Conséquences de l'épopée de la TA

En définitive, ce premier élan n'a pas eu les retombés escomptées – aucun système de traduction en perspective. Cela dit, il a permis d'éclairer une méthodologie propre à ce domaine neuf de *traitement automatique des langues*, mettant en lumière l'intérêt productif de lié la *linguistique* à *informatique*. Néanmoins, cette entre-deux entraîne un problème d'ordre épistémologique, personne ne savait alors vraiment comment classer cette nouvelle discipline, de quelle obédience était-elle exactement ? Plus issue de l'informatique ou de la linguistique ? En ce sens, le *traitement automatique du langage naturelle* a eu quelque peine à être reconnu pleinement comme une discipline scientifique.

2.1.7 Le traitement automatique des langages naturels

C'est vers les années 1990 et le début des années 2000 que l'explosion de l'utilisation d'*internet* offrit au *traitement automatique du langage naturel*

un nouveau souffle. En effet, le contenu d'*internet* amené à une échelle presque mondiale ne cessait – et ne cesse toujours du reste – de grandir. Le contenu s'élargit, les données textuelles se multiplient, d'importantes ressources textuelles sont numérisées – en d'autres termes : les données à traiter sont légions, les techniques sont affinées par les systèmes développés en *TA* et perfectionnés à la suite de son âge d'or, les théories sont multiples, la visé est réaliste : les perspectives sont immenses. Qui plus est, la position du *traitement automatique du langage* reste singulière au regard des techniques informatiques, à la croisée des chemins entre deux domaines significativement différents quoique complémentaires. Aussi ce syncrétisme conduit le *TAL* à deux horizons : les recherches visant à résoudre les difficultés inhérentes au traitement automatique d'une langue naturelle – qui reposent en conséquence sur une approche théorique plus *linguistique* et *cognitive*, et les travaux occasionnant l'optimisation des techniques déjà développées dans ce domaine, qui trouveront quant à eux un intérêt beaucoup plus important dans le domaine industriel. En conséquence, le flou s'épaissit – le *« taliste »* chercheur ou ingénieur ? À cela s'ajoute les multiples disciplines connexes au *TAL* dont le lien n'est pas forcément évident : *mathématiques*, *intelligence artificielle*, *linguistique*, *informatique*, etc. à cela s'ajoute que ce champ disciplinaire est en permanent changement, les évolutions sont constantes tant et si bien que Cori et Léon parlent de l'« *inanité d'une impossible quête, celle de définir un champ unifié qui, tout en englobant les applications industrielles, soit scientifiquement fondé*. »[COL02].

Néanmoins, quelle que soit l'étiquette à lui apposer, le *traitement automatique du langage* existe en propre et le lien entre ces disciplines est tout à fait cohérent, en un sens la *linguistique* et l'*informatique* collaborent à travers le *TAL* comme le note P. Enjalbert « *si les réalisations technologiques posent de réels et souvent passionnants problèmes informatiques ou d'ingénierie de la connaissance, la dimension linguistique doit aujourd'hui être pleinement reconnue. Et si le recours à l'intuition du concepteur d'applications n'est certes pas à proscrire, il n'est pas possible de se replier sur une sorte de « sémantique naïve », partagée « naturellement » par tout locuteur : la langue est une affaire trop complexe pour cela. Ici comme*

ailleurs, le « détour » par la théorie est nécessaire et la sémantique linguistique est riche de modèles qui ne peuvent être ignorés. »[ENJ05].

Aussi, dans une perspective de technicien, l'idée est la même « *la mise en place d'applications concrètes, ne manque cependant pas de souligner la possibilité de « confronter la linguistique, qui pendant longtemps demeura descriptive, à des exigences d'opérationnalité, ou plus précisément ses modèles aux exigences opératoires de la modélisation informatique* »[PIE00].

Néanmoins, F. Rastier distingue ces deux pendants du *TAL* comme distincts mais semble en dire essentiellement la même chose, il y a une inter-communicabilité de ces deux domaines qui peuvent être considérés comme une face différente de la même pièce, sans l'un nous n'aurions pas l'autre et inversement : « *Même si elles abordent par des biais différents des problèmes analogues, il convient de distinguer la linguistique informatique et l'informatique linguistique, bien qu'elles restent généralement confondues sous l'étiquette commode de traitements automatiques du « langage naturel ». La première est une branche de la linguistique qui utilise les technologie de l'informatique pour valider expérimentalement ses hypothèses et pour mettre en œuvre ses résultats par des applications. La seconde est une branche de l'informatique qui utilise les connaissances et les techniques de la linguistique pour développer des applications.* »[RAS91]

2.1.8 Du TAL à l'entité nommées

Cette mise en perspective historique nous a permis de souligner un point qui trouve, selon-nous, un intérêt tout particulier au regard du sujet de ce mémoire. Le *traitement automatique du langage* est une discipline entre linguistique et *informatique*, entre *recherche* et *applicabilité*. Il est ainsi cohérent de s'interroger quant aux conséquences sur sa terminologie, sur la catégorisation des notions qu'elle manipule. Bien sûr, l'objet de ce travail n'est pas d'en dresser une liste exhaustive – du reste, nous ne prétendons pas arriver à ce résultat concernant l'*entité nommée* – néanmoins, nous nous intéresserons maintenant à observer ce qu'est une *entité nommée* à travers son histoire dans le *TAL* et ce à quoi cette notion correspond linguistiquement parlant pour enfin en observer des occurrences concrètes issues du *TAL*.

2.2 Les entités nommées, considérations linguistiques

2.2.1 Pour une présentation de l'entité nommée

La reconnaissance d'*entité nommée* s'intéresse à l'extraction d'unités lexicales précises : noms de personnes, d'organisations, de lieux – un ensemble auquel aura été progressivement ajoutés d'autres entités comme les unités monétaires, les pourcentages ou encore les dates. Aussi, l'objectif de la *reconnaissance d'entité nommée* est tout autant de repérer ces unités dans un texte que de les classer en fonction des types sémantiques précis que nous venons d'évoquer : ce processus s'appelle la *catégorisation* des entités, et elle se matérialisera – cela dépend des systèmes employés – en balise *lxml*, en *parenthèses*, etc. à titre d'exemple, la phrase suivante (un titre tiré de l'édition du monde du 29 août 2020) :

États-Unis : un mort à Portland lors d'affrontements entre manifestants antiracistes et pro-Trump, ce samedi 29 août.

Les systèmes repéreront a priori les entités États-Unis, Portland, etc. puis leur attribueront une étiquette correspondante. Ainsi, l'annotation pourrait se présenter ainsi :

<LOC>*États-Unis****</LOC>*** *: un mort à*

<LOC>*Portland****</LOC>*** *lors d'affrontements entre manifestants antiracistes et pro-****<PERS>****Trump****</PERS>****, ce* ***<DATE>****samedi 29 août****</DATE>****.*

Aussi, bien que la reconnaissance d'*entités nommées* se soit d'abord développée et affinée sur des corpus journalistiques, d'autres systèmes ont été développés dans d'autres domaines, dans le but de fonctionner avec des critères particuliers censés alors opérer sur des corpus spécifiques. À titre d'exemple, il y a une très forte demande de ce genre de systèmes en biologie et médecine dans la mesure où la production scientifique de ces communautés est particulièrement importante – ces systèmes de reconnaissance opèrent donc comme des aides pour repérer et annoter des noms de molécules, de protéines, des expressions de maladies, etc. En somme, ils peuvent être conçus pour une *métalangue* propre à une discipline donnée.

Quoi qu'il en soit, la tâche la plus courante de la *reconnaissance d'entité nommée* reste le repérage et la classification en type sémantique. En définitive, cette opération a pour objectif d'extraire et annoter certains éléments d'un texte – aussi, celle-ci a été inventée comme sous-tâche de l'*extraction d'informations*. Aussi, nous proposons ici de mettre en perspective historique la notion d'*entité nommée* pour bien cerner à quelle problème celle-ci est censée répondre et dans quel contexte elle a été conceptualisée.

2.2.2 Les conférences MUC

2.2.2.1 Contextualisation historique

Le concept d'entité nommée est né dans les années 90, à l'occasion des conférences d'évaluation MUC (*Message understanding Conference*). Celles-ci avaient pour objectifs d'encourager la recherche en extraction d'information, c'est-à-dire une tâche visant à extraire le sens d'un texte : il s'agissait de développer des systèmes capables de remplir des formulaires de façon automatique concernant des événements et, dans ce cadre, certains objets textuels précis ont été regroupés sous le nom d'*entité nommée*. Aussi, l'extraction d'information a succédé aux systèmes qui prédominaient alors qui avaient pour objectif de *comprendre* des texte dans leur intégrité – un objectif pour le moins ambitieu : « *Il ne s'agit donc pas simplement de sélectionner un fragment brut du texte, mais de mettre des éléments en relation pour restituer une information complète et structurée. Sauf dans les cas très simples, c'est une tâche difficile qui requiert une part de compréhension et nécessite des connaissances, des ressources lexicales, sémantiques et conceptuelles adaptées aux documents et au domaine à traiter.* »[BNN01].

Ainsi, l'*extraction d'informations* ne vise pas à comprendre l'ensemble du texte : elle cherchera plus simplement à en extraire des éléments d'informations *a priori* pertinents et ce selon des critères qui auront été précisés au préalable. Pour un texte donné, sera repérer et extrait des occurrences particulières qui seront alors restituées dans une représentation systématisée. Aussi, si ce principe d'*extraction d'informations* n'est pas nouveau, la tâche n'a cessé d'évoluer, pour s'affiner significativement, au fur et à mesure des conférences MUC (*Message Understanding Conferences*)

[MUC19] & [GSU96] – cycle de conférences qui se sont déroulées de 1987 à 1998, organisé par plusieurs institutions américaines et financé par le *DARPA* (*Defense Advanced Research Projects Agency*) – elles ont pour le moins stimulé la recherche en la matière. Financées par le *DARPA*, il va sans dire que le but premier de ces conférences était la compréhension automatique de messages militaires – cela semble indissociable concernant l'informatique et le langage comme nous avons pu le voir, aussi intéressons-nous aux retombés de ces conférences et voyons si elles se sont avérées plus fructueuses que leurs aînées de la *traduction automatique*.

Combien même celles-ci sont nommées « conférences », il s'agit véritablement d'évaluations de systèmes : il est confié à un certain nombre de participants un corpus d'entraînement et des instructions précises sur ce qu'il faudra automatiquement en extraire – après quoi, ils devront appliquer le système qu'ils ont développé sur un corpus de test donné. Enfin, les résultats seront réunis et évalués puis présentés lors de la conférence dont seul les participants à l'évaluation ont le droit d'assister, « *The Message Understanding Conferences were initiated by NOSC to assess and to foster research on the automated analysis of military messages containing textuel information. Although called 'conferences', the distinguishing characteristic of the MUCs are not the conferences themselves, but the evaluations to which participants must submit in order to be permitted to attend the conference*. »[GSU96] . Aussi, le contenu de ces conférences a été documenté, nous nous appuierons sur le travail de R. Grishman et B. Sundheim [GSU96] afin de nous intéresser aux évolutions et surtout mettre en avant le moment où est apparu la notion qui nous intéresse ici : l'*entité nommée*.

2.2.3 Conférences MUC, contenu

Les deux premières conférences de 1987 et 1989 étaient d'un intérêt assez succinct, réellement celles-ci ont surtout servies pour poser une nomenclature de ce qui serait attendu. Le premier corpus sur lequel les ingénieurs travaillèrent était un corpus de messages de la marine américaine dactylographiés. Aussi, ils n'avaient pas dressé de liste de ce qui était censé en être extrait, et n'avaient pas moyen en conséquence, d'évaluer les premiers résultats. C'est ainsi qu'à la deuxième conférence, ils établirent un patron de

ce qui était recherché – et conçurent également un système pour en évaluer l'efficacité : les calculs de précision et de rappel.

Après celles-ci, la tâche était plus clairement définie et les systèmes systématiquement évaluables ; aussi la tâche d'extraction d'informations s'affina, ces mêmes systèmes devinrent de plus en plus complexe et en conséquence, de plus en plus précis. Dans les conférences MUC-3 et 4 (1991 et 1992), les corpus étaient composés de dépêches journalistiques type :

> TST2-MUC3-O069 BOGOTA, 7 SEP 89 (INRAVISION TELEVISION CADENA 1) -- [REPORT] [MARIBEL OSORIO] [TEXT] MEDELLIN CONTINUES TO LIVE THROUGH A WAVE OF TERROR. FOLLOWING LAST NIGHT'S ATTACK ON A BANK, WHICH CAUSED A LOT OF DAMAGE, A LOAD OF DYNAMITE WAS HURLED AGAINST A POLICE STATION. FORTUNATELY NO ONE WAS HURT. HOWEVER, AT APPROXIMATELY 1700 TODAY A BOMB EXPLODED INSIDE A FAST-FOOD RESTAURANT. A MEDIUM-SIZED BOMB EXPLODED SHORTLY BEFORE 1700 AT THE PRESTO INSTALLATIONS LOCATED ON [WORDS INDISTINCT] AND PLAYA AVENUE. APPROXIMATELY 35 PEOPLE WER~! INSIDE THE RESTAURANT AT THE TIME. A WORKER NOTICED A SUSPICIOUS PACKAGE UNDER A TABLE WHERE MINUTES BEFORE TWO MEN HAD BEEN SEATED. AFTER AN INITIAL MINOR EXPLOSION, THE PACKAGE EXPLODED. THE 35 PEOPLE HAD ALREADY BEEN EVACUATED FROM THE BUILDING, AND ONLY 1 POLICEMAN WAS SLIGHTLY INJURED; HE WAS THROWN TO THE GROUND BY THE SHOCK WAVE. THE AREA WAS IMMEDIATELY CORDONED OFF BY THE AUTHORITIES WHILE THE OTHER BUSINESSES CLOSED THEIR DOORS. IT IS NOT KNOWN HOW MUCH DAMAGE WAS CAUSED; HOWEVER, MOST OF THE DAMAGE WAS OCCURRED INSIDE THE RESTAURANT. THE MEN WHO LEFT THE BOMB FLED AND THERE ARE NO CLUES AS TO THEIR WHEREABOUTS.
>
> *Texte 1: Exemple d'article du corpus MUC-3*

[CLH93]. Les patrons proposés par les organisateurs comportent de plus en plus de champs à détecter (allant jusqu'à 24). Aussi, le changement de nature de corpus impose un affinement des techniques dans la mesure où ceux-ci sont plus longs ; conséquemment les informations à en extraire s'avèrent plus compliquées à repérer. On voit apparaître là les premiers systèmes fonctionnant à base d'automates ou ceux basés sur des méthodes statistiques. Du reste, pour affiner l'évaluation, une nouvelle mesure est introduite : la *F-mesure* qui combine les taux de précision et de rappel afin de rendre la comparaison entre systèmes plus aisée.

À la MUC-5, ils compliquèrent la tâche en proposant cette fois plusieurs corpus (technologique et commercial) et ce, pour deux langues : l'anglais et le

japonais. Aussi, le temps de développement des systèmes aura été beaucoup plus long, 6 mois, et ceux-ci ne furent pas plus performants que leurs prédécesseurs. En conclusion de cette conférence, il apparaissait comme de plus en plus évident de devoir diviser en fonctionnalités autonomes les tâches d'*extraction d'informations* devenues alors trop complexe pour fonctionner en un bloc « *This raises the issue of portability--most systems spent approximately one person year preparing their systems to run for MUC-3. If porting takes 12 months of time for highly trained system developers, portability will be a serious stumbling block both to building real systems and to changing the evaluation paradigm. At the end of MUC-3, the participating system developers did not want to spend another year porting their system to yet another evaluation application. This underscores the need for serious research on portable systems.* » [CLH93]. et de poursuivre :

« *In order to evaluate the linguistic coverage, we de- vised a successful method for isolating specific linguistic phenomena and measuring system performance on them within the black-box framework, even though specifics of the performance of systems varied and made these tests more difficult to interpret. Development of a suite of such tests with adequate linguistic coverage would pro- vide insight into how the handling of certain common linguistic phenomena relates to overall system performance* » [CLH93].

Après celles-ci, deux dernières conférences eurent lieux. La MUC-6 (1995) qui introduisit trois nouvelle tâches dans le but de répondre aux nouveaux objectifs fixés : il s'agissait de tenter de concevoir des systèmes autonomes pouvant aisément être utilisés mais aussi d'inciter à créer des systèmes plus portables – de favoriser des tâches étant en mesure de s'atteler à une compréhension profonde. Il s'agissait du pôle « SemEval »[MUC19] regroupant des tâches comme : détection de structure prédicat-argument, désambiguïsation lexicale et résolution de coréférence. Par ailleurs, et c'est là le plus intéressant au regard de notre sujet, il est créé au travers de cette volonté de simplification du patron traditionnel conjointe à la volonté d'inciter à la création de fonctionnalités indépendantes d'analyse, une nouvelle tâche : celle de la *Named Entity Recognition* (NER) – *reconnaissance d'entité nommée*. Par ailleurs, le temps de distribution des corpus de test étant plus cours, cela incita les participants à développer des systèmes transposables.

Les niveaux de performances furent particulièrement bons : pour la tâche de repérage des *entités nommées*, plusieurs systèmes atteignent une F-mesure de 0,9. Les autres tâches obtinrent également des résultats plus que corrects. Une franche réussite. Aussi, ces résultats tout à fait encourageant étaient la preuve que segmenter la tâche d'*extraction d'informations* s'avérait sinon nécessaire du moins fructueuse : au regard des très bons résultats de certaines de ces sous-tâches mais aussi pour faciliter la conception de systèmes génériques autonomes. Enfin, la dernière conférence MUC-7 (1998) n'occasionne pas le même engouement que la précédente, les résultats ne sont pas particulièrement impressionnants et, du reste la leçon méthodologique en a déjà été tirée. Notons que parallèlement à la MUC-6 et 7, une conférence est organisé : la *MET* : *multilingual Entity Task*. Celle-ci permet de faire le pont entre les systèmes développés dans le monde anglophone à d'autres langues, au rang desquelles : l'espagnol, le japonais et le chinois. En cela, elle internationalise cette méthodologie anglo-saxonne.

L'entité nommée actuelle

2.2.4 Un problème terminologique et définitoire

à l'image des noms propres, les entités nommées ne sont pas particulièrement évidentes à définir. Dans un premier temps, l'intersectionnalité de toutes ces disciplines impose un questionnement : quelle méthodologie semblerait appropriée pour définir un objet de TAL ? La réalité de l'entité nommées est plurielle : il s'agit d'un besoin concret du traitement automatique de la langue dans la mesure où celles-ci fait partie intégrante de l'extraction d'information, elle est « *davantage concernée par des processus que par des objets proprement dits* »[EHR08]. Cela dit, leur matérialité textuelle comme orale relève tout autant de la linguistique. Aussi, définir les entités nommées suppose de prendre en considération ces deux dimensions. Du reste, il existe aussi des termes correspondant à des objets manipulés en TAL quoique ceux-ci sont en grandes parties incorporés à d'autres domaines, par exemple, la notion de *corpus* provient directement de la *linguistique* même si sa signification est remaniée pour s'intégrer pleinement au cadre du *TAL*. En ce sens, il n'est ni aisé de la définir

linguistiquement ni satisfaisant car objet du *TAL*. Par ailleurs, si la *linguistique* s'attache tout particulièrement, nous l'avons vu avec le nom propre précédemment, à définir ses notions avec le plus de réserve et de précision possible – pouvant occasionner d'ailleurs de longs, fastidieux mais passionnants débats – le TAL ne possèdait vraisemblablement pas une tradition définitoire, celle-ci est une science jeune ainsi qu'une science à visé applicative, il est conséquemment possible de penser que ses définitions ne sont pas aussi poussées qu'en *linguistique* et que ses éléments ne font pas nécessairement l'objet d'études approfondies – du moins pas encore, et pas à la même échelle.

Aussi, c'est lors de campagne MUC-6 que la tâche de reconnaissance d'*entité nommée* est conceptualisée et donc, plausiblement, définie. Quelques définitions tirées des MUC ou des conférences MET-1 et MET-2 (*Multilingual Entity Task Conference*) pour le chinois et le japonais :

- « *On the level of entity extraction, Named Entities (NE) 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* » [CHI98]
- « *The Named Entity task consists of three subtasks (entity names, temporal expressions, number expressions). The expressions to be annotated are « unique identifiers » of entities (organizations, persons, locations), times (dates, times), and quantities (monetary values, percentages). » précisant d'ailleurs : « This subtask is limited to proper names, acronymes, and perhaps miscellaneous other unique identifiers* »[CHI97]
- « *Names entities are phrases that contain the names of persons, organizations, locations, times and quantities* »[TKM03]

Pour des traductions hors-conférences, T. Poibeau définit ainsi l'*entité nommée* « *On appelle traditionnellement « entités nommées » (de l'anglais named entity) l'ensemble des noms de personnes, d'entreprises et de lieux présents dans un texte donné. On associe souvent à ces éléments d'autres syntagmes comme les dates, les unités monétaires ou les pourcentages repérable par les mêmes techniques à base de grammaires locales.* »[POI06]

Enfin, nous pouvons également relever : « *In this paper, term « Named Entity » includes names (which is narrow sens of Named Entity) and numeric expressions. The definition of the Named Entity is not simple, but, intuitively, this is a class that people are often willing to know in newspaper articles*. » [SSN02].

Aussi, ce qui semble faire œuvre de dénominateur commun dans chacune de ces définitions est que l'*entité nommée* renferme une information qui est susceptible d'intéresser son utilisateur[9], compréhensible uniquement à partir du contexte immédiat. Cette information, les systèmes *souhaitent* l'*extraire*. Du reste, nous pouvons remarquer que ces définitions semblent construire leur propre terminologie pour définir une notion, comme s'il fallait inventer de nouveaux critères définitoires pour un élément qui se trouve pourtant *de facto* dans la langue.N. Friburger le souligne en expliquant que « *les informaticiens qui travaillent dans le domaine de l'extraction d'information, ont abordé le problème de manière pragmatique. Ils ont défini la notion d'entités nommées pour regrouper tous les éléments du langage définis par référence : les noms propres au sens classique, les noms propres dans un sens élargi mais aussi les expressions de temps et de quantités* »[FRI02].

À cela s'ajoute que cette catégorisation de l'*entité nommée* a eu nécessité d'évoluer au fur et à mesure des conférences MUC, du fait que les critères existants alors ne permettaient pas d'englober une entité souhaitée. D'une certaine façon, cela peut sembler étonnant qu'un tournant plus linguistique n'ait pas été considéré plus tôt mais, dans un domaine d'entre-deux, comme nous avons pu le souligné, cela n'est pas tant surprenant : il s'agissait de définir au mieux, le plus efficacement possible, une catégorie à extraire de textes dont l'extraction sera du reste mesurée – quant à la réalité théorique que cela suppose, ce n'était pas l'objet de ces conférences.

9 à comprendre *lecteur* ou *machine*, tout ce qui est en mesure de traiter un texte

2.2.5 Tentative de définition selon des critères linguistique

Dans ces tentatives de définir les *entités nommées*, il ne sera plus question d'approximation mais de s'intéresser à une catégorisation proprement *linguistique* de ce terme, considéré pour lui-même et en tant qu'objet d'une discipline précise et distincte, le *TAL*.

- « *étiquetage sémantique des entités nommées : il faut détecter toutes les formes linguistiques qui, à l'instar des noms propres désignent de manière univoque une entité par leur pouvoir de sélectivité : noms de personnes, d'institutions et entreprises, de lieux, ainsi souvent que les dates, unités monétaires, etc. Il faut aussi leur affecter une étiquette sémantique choisie parmi une liste prédéfinie* » [ENJ05].
- « *Entité nommée est la notion utilisée en TAL pour désigner les éléments discursifs monoréférentiels qui coïncident en partie avec les noms propres et qui suivent des patrons syntaxiques déterminés. Du point de vue de leur structure syntaxique, les entités nommées sont fortes quand elles sont composées exclusivement de noms propres et faibles quand elles sont constituées par un nom propre (ou entité nommée forte) et une forme catégorisante* » [VIC05]

On remarque là une tendance plus prononcée des définitions à se diriger vers des critères relevant avant tout de la *linguistique* – et ce dans son intégralité, non plus seulement pour en évoquer le contenu de sens, c'est-à-dire selon une perspective non plus exclusivement *sémantique*. Par ailleurs, en nous intéressant à la définition de la *named entity recognition* sur *wikipedia*, nous pouvons remarquer que celle-ci adopte même des critères de logique[10] « *Named entity recognition (NER) (also known as entity identification (EI) and entity extraction) is a subtask of information extraction that seeks to locate and classify atomic elements in text into predefined categories such as the names of persons, organizations, locations, expressions of time, quantities, monetary values, percentages, etc. (…) In the expression named entity, the word named restricts the task to those entities for which one or many rigid designators, as defined by Kirpke, stand for the*

10 certes souvent repris tel quel en *linguisitique*

referent »[WIKI] - à noter du reste que la page *française* ne fait pas cette distinction et n'évoque pas le travail de *Kripke [WIKIFR].*

2.2.5.1 Une porté alpha-numérique

L'*entité nommée* a également cela de particulier qu'elle recouvre une multitude d'objets distincts. Aussi, si les premières conférences *MUC* lui avait donner l'objectif d'extraire ce qui correspondait à des *noms propres* purs tel que des noms de personnes, d'organisation, etc. les suivantes y ont adjoint une nouvelle liste d'objet. Aussi, la proximité entre les *noms* propres et les mesures de distance, les dates ou, plus prosaïquement, les expressions numériques peut sembler étonnante – quoi qu'il en soit, ce rapprochement s'est fait. Ainsi, s'ajoutèrent à la nomenclature de l'*entité nommée* les NUMEX et autres TIMEX [MUC19]. Par la suite, l'*entité nommée* s'étendra encore pour recouvrir d'autres éléments lesquels sont des objets relativement éloignés, encore, du concept strict du *nom propre*.

2.2.6 critère formel vs. critère référentiel

Il nous faut désormais évoquer un point fondamental concernant la notion d'*entité nommée* et de *nom propre*. Leur point commun concerne leur référentialité – comme nous l'avons-vu, celle-ci n'est pas nécessairement systématique pour le *nom propre*, néanmoins elle l'est pour l'*entité nommée* ; or le *nom propre* n'est pas le seul objet *linguistique* qui est référentiel. Considérons par exemple le syntagme suivant, *le maire de la commune*, ici, cela fait référence *a priori* à un individu unique. En effet, il faudrait préciser le nom de la commune pour que ce soit le cas, mais en admettant que ce syntagme apparaisse dans un article du quotidien de *p. ex Joinville-le-Pont*, nous pourrions logiquement conclure que celui-ci réfère à un individu unique – certes dans la limite de l'exercice de ses fonctions par rapport à l'instant d'édition du périodique. Néanmoins, ce syntagme *réfère* et, en conséquence, celui-ci peut-être considéré comme une *entité nommée*. Ainsi, Dubois et d'autres linguistes parlent de la notion de *référence* comme de la « *propriété d'un signe linguistique lui permettant de renvoyer à un objet du monde extra-linguistique, réel ou imaginaire* »[DUB94]. Une propriété que ce syntagme

nominal possède clairement. Néanmoins, il faudrait selon M. Ehrmann préciser que « *d'une part l'élément du réel dont il est question se trouve dans un réel conceptualisé sur la base d'une intersubjectivité stable, et, d'autre part, cet élément existe en dehors du langage.* »[EHR08] suggérant par là qu'il existe un flottement sur ce processus de référence qui n'est pas si évident, direct et consensuel – d'un locuteur à l'autre, d'un sujet-parlant à l'autre, tel ou tel syntagme va avoir un degré de signification légèrement différent. Aussi, nous verrons que cela peut poser un problème non nécessairement pas pour repérer une *entité nommée* mais pour la *délimiter*. D'autant que dans le cadre d'un processus de *TAL*, M. Ehrmann précise « *que la référence ne peut être la référence au monde dans son entier mais seulement à une représentation partielle de ce monde, ou modèle du monde*. »[EHR08]

S'il est entendu que les *noms propres* sont, comme nous l'avons vu, le prototype de l'unité référentielle en *linguistique*, comment procèdent par contre ces syntagmes référentiels ? À quoi correspondent-ils exactement ? En quoi sont-ils aptes à référer ? Peut-on déterminer clairement leur fonctionnement ? Nous allons tâcher d'y répondre.

2.2.7 Syntagme référentiel, théorie de *Kleiber*

Kleiber s'appuie sur les théories des *actualisateurs : virtuel et actuel* des distributionnalistes qui affirment que tous les syntagmes nominaux renvoient à des êtres ou des objets particuliers, ainsi que sur celles de la *description définie* des logiciens *Russel, Frege* et *Quine* pour comprendre ce qui fait référence dans la langue. Dans son ouvrage *Problèmes de référence : descriptions définies et noms propres* – celui-ci s'attache à ré-actualiser ces théories et à répondre à leurs problèmes méthodologiques – les *actualisateurs* ont tiré des conclusions trop globalisantes « *L'erreur des distributionalistes et des tenants de la théorie de l'actualisation a été d'étendre la fonction référentielle que remplissent les SN L'homme et Un homme dans les énoncés L'homme est venu et Un homme est venu à tous les SN.* »[KLE81]. quant aux *logiciens* leurs thèses souffre simplement du fait « *qu'une langue naturelle n'est pas affaire mathématique* ».

Aussi, celui-ci pose plusieurs points conditionnels sur la référence d'un syntagme nominal qui se construit sur la notion de référence étendue aux lexique dans son ensemble, établissant ainsi un *parcours* du *référer* :

- 1 :« *tout item lexical réfère, parce qu'il présuppose l'existence d'une référent conceptuel* ; »[KLE81]
    - c'est-à-dire présuppose pour chaque « item » un quelque chose flou et très général mais réel.
- 2 :« *On peut dire des noms qu'ils réfèrent par rapport aux autres items lexicaux comme les verbes ou les adjectifs, parce que la majeure partie d'entre eux, la catégorématiques, présupposent l'existence d'une catégorie référentielle stable, autonome, continue pour les globalisants, discontinue pour les individuants, catégorie qui passe pour être plus réelle* ; »[KLE81]
    - c'est-à-dire qu'est donné ici une réalité plus grande aux noms qui sont séparés des autres items lexicaux. *Individuants* d'un coté, soit ceux qui présupposent une catégorie référentielle que l'on peut découper en individu, *peinture, peintre, artiste,* etc. Les *globalisants* vont quant à eux « *réifiés en discontinu* » soit, réduire à l'état d'objet une notion abstraite.
- 3 :« *Les syntagmes nominaux réfèrent, quelle que soit leur position dans l'énoncé parce qu'ils présupposent, du fait de la quantification restreinte, c'est-à-dire du fait de l'absence de variables totalement indéterminées, l'existence d'une classe référentielle non vide.* »[KLE81]
    - c'est-à-dire, la référence clarifiée grâce au processus syntaxique, l'apposition de complément – adjectifs, etc.

Ainsi, « *Si l'on définit la référence comme le lien entre le langage et le réel (ou monde extérieur), on comprendra pourquoi on parlera plus facilement de référence à propos de 3°, c'est-à-dire pour les syntagmes nominaux, qu'à propos de 1°. On comprendra également pourquoi ce sont les syntagmes nominaux qui sont prédestinés à servir d'expression référentielle pour la référence aux particuliers*. »[KLE81]

Ainsi, à partir de ce constat, Kleiber propose de définir les *descriptions définies* ainsi « *D'un point de vue référentiel, notre définition équivaut à dire que les descriptions définis sont des expressions qui peuvent servir à référer à un objet particulier. Aussi peut-on réserver l'appellation descriptions définies aux seuls SN définis singuliers qui, lorsqu'ils sont utilisés en position référentielle, renvoient à un particulier. Étant donné notre propre définition des particuliers, on pourra également qu'une description définie est un SN défini singulier qui, en position de sujet, ne peut jamais, de par son seul sens, indiquer totalement quel est le référent. Il faut obligatoirement des points ou des indices référentiels pour repérer le référent visé.* »[KLE81], en somme, autant de critères formels qui nous permettent d'identifier précisément ce qui peut, ou non, correspondre à une *entité nommée* – singulier, position sujet, points ou indices référentiels. En d'autres termes, ce sont là autant de critères pertinents pour en délimiter la frontière théorique.

2.2.8 Tentative d'une définition linguistico-applicative

En ce sens, la définition de *Serkine* soulevait intuitivement ce qui justifie l'ensemble des *entités nommées*, c'est-à-dire le rapport essentiel avec la notion de fonctionnement référentiel. Aussi, en essayant de rattacher cela à des catégories syntaxiques précises, nous pouvons constater que cet ensemble est majoritairement composé de *noms propres*, de syntagmes nominaux obéissants aux critères de la *description définie* – et ce qui paraît être commun à ces unités est que leur fonction principale est de viser un élément du réel. Pour elles, la référence est fondamentale – celles-ci sont parfois proprement et strictement dénominative, leurs interprétations ne pouvant se faire que par des indices référentiels. Or, c'est là une spécificité particulièrement complexe à analyser déjà d'un point de vue *linguistique*, d'autant plus en *TAL* à partir du moment où cela doit, d'une façon ou d'une autre, être modélisé pour la machine.

Ainsi, M, Ehrmann en propose la définition suivante « *étant donnés un modèle applicatif et un corpus, on appelle entité nommée toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus*. » [EHR08].

Dés lors, l'un des problèmes majeurs concernant les *entités nommées* réside dans le fait que celles-ci doivent répondre à une applicabilité et, en ce sens, à une catégorisation prédéfinie. En cela, nous entendons que pour un système de détection donnée, celui-ci doit obéir à une définition des catégories d'*entités nommées* qu'il souhaite extraire d'un texte – orientée selon le type de texte dans lequel celui-ci se spécialise, si tant est que ce soit le cas. Or, cela n'est pas une tâche triviale dans la mesure où il est nécessaire de déterminer les catégories mais aussi les différents constituants de chacune d'elles. Aussi, comme nous l'avons vu, les conférences *MUC* avaient rajoutées de plus de plus de sous-catégories dans le but de préciser au mieux les choses. Aussi, la clef ne serait pas de démultiplier les catégories pour couvrir chaque potentiel objet du monde mais plus simplement de « *prendre en compte le domaine applicatif, de considérer la catégorisation comme un véritable enjeu et d'adopter une démarche méthodologique pour sa réalisation (…) il n'existe aucune catégorisation idéale, ni de solution pour y parvenir ; le mieux semble être de suivre la proposition de S. Sekine « We believe that there is no ultimate solution, so we seek rather empirical solution », et de multiplier les sources d'inspiration.* » [FEH12].

2.3 Conclusion

Ainsi, le concept d'*entité nommée* n'a pas de définition apparemment consensuelle. Ce qui est du reste compréhensible en considérant la jeunesse du *TAL* en tant que discipline, de l'*extraction* d'*information* en tant que domaine de celle-ci et de l'extraction de l'*entité nommée* comme sa sous-tâche. Qui plus est, le lien entre le *nom propre* et l'*entité nommée* suppose que celle-ci hérite tout autant de ses soucis de catégorisation, et de la problématique, plus large, de définir un objet sémantique complexe tel que celui-ci. Un rapprochement sans doute facilité par le fait qu'à ses débuts, l'*entité nommée* n'avait pour but que d'extraire des noms de personnes, organisation, etc. assimilables à des *noms propres*.

Aussi, l'*entité nommée* recouvra à terme bien plus de données que les seuls *noms propres ;* elles extraient chaque unité référentielle, admettant alors cette notion de *description définie* – à travers des syntagmes qui ont été somme toute assez peu traiter dans la théorie *linguistique* – de fait, *Saussure*

posait l'objet d'étude de la *linguistique* sur le *signifié* et le *signifiant* balayant le *référent* de l'équation car relevant de la parole et non du langage[SAU10]. La tradition saussurienne influença grandement la *linguistique* française jusqu'à récemment où des linguistes comme Kleiber s'évertuèrent à réintégrer le référent à sa place due « *cette sémantique que nous avons appelée à cause de (i), et uniquement à cause de (i), référentielle (cf. le sous-titre « Essais de sémantique référentielle » de Nominales [réf. À son ouvrage de 1994]), est résolument ouverte au cognitif en ce qu'elle s'ancre dans l'expérience humaine sous toutes ses dimensions interactives : perceptuelle, sociale, culturelle, etc. Cette vocation cognitive s'exprime par une volonté de justifier le langage par notre conceptualisation et notre représentation du monde*. »[KLE99deux] une approche qu'il justifie comme « *le réel est en dehors du linguistique et n'a donc logiquement rien à faire dans les affaire du langage. D'un autre coté, si l'on accepte que parler, c'est dire quelque chose, le quelque chose en question, que l'on ne peut éviter, nous pousse à répondre positivement : oui, le réel et partie prenante dans le commerce linguistique, puisque c'est sur lui que s'exerce notre dire*. »[KLE99].

À cela s'adjoint que cette tâche se fait selon des critères précis et standardisés par les ingénieurs réalisant ces-mêmes systèmes ; au vu de la tâche qui les incombe, il n'est pas nécessairement évident de rendre toute les subtilités de ces notions applicables.

En ce sens, la dimension applicative du TAL n'autorise qu'assez difficilement de définition normative classique – il s'agirait plutôt de déterminer assez globalement un ensemble d'objets à récupérer, une *guideline* en quelque sorte, pour guider le développement des systèmes proposés pour ces conférences comme l'atteste ce genre de précisions : « In some cases, multi-word strings that are proper names will contain entity name substrings; such strings are not decomposable; therefore, the substrings are not to be tagged » [CHI98]. De fait, l'*entité nommée* s'est développée d'elle même, par les ingénieurs qui avaient nécessités de concevoir des systèmes fonctionnels avant de s'intéresser à la qualité *linguistique* de leurs projets.

Par ailleurs, l'*imbroglio* de la sémantique lors de la conceptualisation de ces unités ne facilita la tâche à personne, comme le disait Kleiber « *Tout bouge en sémantique à l'heure actuelle. Finie l'époque (heureuse?) du*

structuralisme et de sa sécurisante analyse en sèmes ou traits de significations oppositifs. »[KLE99deux] et à Rastier d'ajouter sur *les échecs définitoires « toutefois cet échec est surtout imputable à la linguistique puisque, faute d'avoir développé correctement la sémantique fondamentale, elle n'a pu produire de théorie cohérente de la traduction (même non automatique)* »[RAS91] et de louer par ailleurs la capacité des informaticiens « *à l'exception de quelques applications somme toute marginales, tous ces domaines de l'ingénierie linguistique se heurtent à des problèmes sémantiques non résolus, voire mal posés. À l'instar de bien des linguistes, les informaticiens ont déployé des ruses de Sioux pour les contourner*. »[RAS94]

3 Expérimentations sur trois systèmes de reconnaissance d'entités nommées

3.1 Traitement de données

3.1.1 Corpus utilisé, l'est républicain

Ce corpus se compose des données textuelles réunies sur deux années de l'intégralité des éditions du quotidien *L'est républicain*. Celui-ci est principalement diffusé dans les régions de *Lorraine* et de *Franche-Comté*, lu par quelques 114 000 personnes environ chaque année. En ce sens, il traitera tout autant d'informations internationales, nationales ou régionales – cela nous donne trois tiers possibles de reconnaissances potentielles des entités nommées, notons que ce classement est fait selon des critères strictement personnels :

- Les plus facilement repérées : en ce sens que les personnalités, les grands pays, ou encore les lieux-dits réputés (pour raisons architecturales, artistiques, etc.) sont susceptibles de faire l'objet de plus de représentation dans certains articles de presses internationales anglo-saxon, français, etc. qui auraient servi à nourrir les bases de données utilisées pour concevoir ces systèmes de reconnaissance d'*entités nommées*. Nous pourrions citer à titre d'exemple : le *Mont Saint-Michel*, la *cathédrale de Reims*, *Calais, Emmanuel Macron*, etc.
- En dessous : nous aurions des villes, personnalités, lieux-dits connus à une échelle nationale – renommés, certes, mais à un niveau qui n'occasionne pas nécessairement de représentation dans une presse générale internationale. Par « générale », nous entendons que certains de ces monuments peuvent, par des caractéristiques qui leurs sont propres (églises *cathares* par exemple), être représentés dans une presse internationale plus spécialisée – périodiques sur l'architecture dans ce cas. Nous pourrions citer à titre d'exemple : le *viaduc de Millot*, *Le Havre*.

- Enfin, nous aurions des villes, personnalités, lieux-dits qui n'auraient résonance qu'à échelle régionale, celles-ci seraient des villages – communes de ces régions couvertes par ce journal, et donc ce corpus. Nous pensons que cet ensemble sera le moins susceptible d'être repéré par ces systèmes.

Du reste, ce corpus est le fruit d'une collaboration entre le journal et *ORTOLANG* (*Outils et Ressources pour un Traitement Optimisé de la LANGue*) ; un organisme validé par l'état dans le cadre du programme étatique des *investissements d'Avenir*, des subventions destinées à l'enseignement supérieur et la recherche. En cela, il s'agit d'un corpus stable, sérieux, surveillé et déjà utilisé dans le cadre du *traitement automatique du langage* – aussi, c'est là la visée de cet organisme : regrouper des données utilisables comme des corpus ou encore des dictionnaires clairement documentés mais aussi libre d'accès afin de valoriser la recherche en langue française en *TAL*.

Ce corpus comporte la période de publication à partir de l'année 1999 jusqu'à l'année 2003 – nous sélectionnerons des articles pertinents – selon des critères que nous expliquerons – sur l'ensemble du corpus.

Il se compose respectivement de trois fichiers[11] :

- fichier 1999 : les articles parus entre le 17 mai et le 30 septembre 1999, comporte environ 36 millions de mots, soit 151 Mo de données.
- fichier 2002 : articles parus pendant toute l'année 2002, comporte environ 98 millions de mots, soit 408 Mo de données.
- fichier 2003 : articles parus en janvier et février 2003, comporte environ 15 millions de mots, soit 62 Mo de données.

3.1.1.1 Structure du corpus

Ce corpus se présente sous la forme d'une multitude de documents *XML* pour une année donnée (trois documents sont à notre disposition, pour les années 1999, 2002 et 2003). Nous en avons extrait les articles grâce au module *BeautifulSoup* une librairie développée pour faciliter le *parsing* d'*HTML* ou d'*XML*. Nous récupérerons les données aux balises

11 Téléchargeable ici : http://redac.univ-tlse2.fr/corpus/estRepublicain.html

* * avec lesquelles nous extrairons plusieurs articles que nous traiterons via *python*.

3.1.1.2 Critères de sélection

Notre objectif sera d'illustrer les problématiques de catégorisation de *nom propre* et d'*entité nommée* dont nous avons parlé dans les chapitres précédents. Il s'agira de les illustrer et d'en proposer une analyse selon les résultats que nous avons obtenu.

Nous utiliserons le module *pyplot* de *matplotlib* pour proposer une illustration schématiques desdits résultats – aussi, nous nous plongerons dans le texte pour proposer une représentation direct des résultats obtenus, et ce pour chaque système que nous utiliserons.

3.2 Mesures potentielles supplémentaires

De façon à établir une mesure plus exhaustive de ces systèmes, nous aurions pu appliquer le calcul de *F-mesure* à l'échelle du corpus entier. Pour ce faire, il nous aurait fallu définir des critères de mesure précis c'est-à-dire récupérer *x* annotations présentes pour chacun des systèmes.

L'objectif aurait alors été ensuite de les faire s'exécuter sur l'intégralité du corpus et de croiser les résultats pour définir une mesure *globale* sur laquelle nous aurions pu mesurer individuellement l'intégralité des systèmes en nous intéressant à la *catégorisation* comme à la *délimitation* des *entités nommées* repérées. Aussi, nous avons choisi de nous concentrer sur une approche plus illustrative.

3.3 Systèmes utilisés

Nous avons utilisé trois systèmes de repérage d'entité nommée, sélectionnés pour leurs méthodes ainsi que pour leurs utilisations. Bien sûr, il est important de préciser que ceux-ci ne représentent pas la majorité des systèmes disponibles en ce domaine. Cela dit, ces choix ont été motivé par le fonctionnement différent de ces systèmes.

3.4 Annotation faite main

3.4.1 méthodologie de l'annotation

Pour que l'annotation d'un *corpus* en *entité nommée* puisse être considérée comme standard, il faut définir précisément ses attentes et ses objectifs. Comme nous l'avons vu, les *entités nommées* n'ont pas de frontières définies ni de définition consensuelle – en l'occurrence, il revient à l'étiqueteur de justifier ses choix au travers une définition précise et nette des catégories qu'il souhaite extraire. Une fois fait, celui-ci doit procéder à son étiquetage manuellement puis le confronter à de tiers étiquetages (vraisemblablement réalisés par son équipe) de façon à vérifier si les résultats se chevauchent ou non, le cas échéant : l'annotation est validée ; autrement, il faut revoir et affiner les définitions.

De fait, nous l'avons mentionné ci-dessus, <ORG>Ville de Paris</ORG> pourrait se voir annoté comme Ville de <LIEU>Paris</LIEU>. En somme, il s'agit d'un travail minutieux et sérieux qu'il est nécessaire de respecter dans la mesure où nombre de corpus ont été étiquetés sans préambule pour en expliquer la nomenclature, les rendant en conséquence inutilisables, dans la mesure où ceux-ci ne répondent finalement à aucun critère. De fait, les *entités nommées* ont toujours un but applicatif, il s'agit avant tout de calculer le réel pour le rendre traitable et cela impose de devoir adopter un standard subjectif.

3.4.2 Nos critères d'annotation

Dans la balise <lieu> nous souhaitons réunir tous les noms *toponymiques* c'est-à-dire les *noms propres* désignant un lieu, que ce soit une ville, une région, un pays, un continent, etc.

Dans la balise <date> nous souhaitons réunir les dates précises, c'est-à-dire n'incluant que des journées – moment précises type *7 juillet 2018*.

Dans la balise <personne> nous souhaitons réunir tous les noms de personnes, selon que cela se composent d'un anthroponyme accompagné d'un patronyme, d'un anthroponyme seul, etc. Nous excluons de cette catégorie ce qui touche au titre comme suit : le Président <pers>Emmanuel

Macron</pers> ou encore Le Tsar <pers>Nicholas</pers> II de Russie, nous aurions du reste accepté le Tsar <pers>Nicholas Romanov</pers> de Russie.

Dans la balise <mesure> nous souhaitons réunir les mesures de distance précise à l'image de <date> comme *1 km 50*.

Dans la balise <org> nous incluons toute organisation que ce soit d'état ou de l'industrie tel que *la république française* ou encore, selon le contexte, *la France*.

Aussi, il nous faut préciser que cette annotation sert de guide pour trouver des comportements types que nous voulions mettre ne lumière et étudier – par ailleurs, celle-ci n'a pas fait objet d'un quelconque accord inter-annotateur.

Le système CasEn

CasEN est une cascade de transducteurs conçue pour la reconnaissance d'*entité nommée* implémentée sous le logiciel CasSys de la plate-forme Unitex, plus précisément, il s'agit d'un : « *système de création et de mise en œuvre de cascades de transducteurs, aujourd'hui intégré à la plateforme Unitex* »[FRI13].

Unitex est un logiciel librement mis à disposition des utilisateurs sous licence LGPL-LR.

Ces cascades de transducteurs réfèrent à des cascades de graphes au sens d'Unitex – définis comme suit « *Unitex peut manipuler plusieurs types de graphes qui correspondent aux utilisations suivantes : flexions automatique de dictionnaires, prétraitement des textes, normalisation des automates de texte, graphes dictionnaires, recherche de motifs, levée d'ambiguïtés et génération automatique de graphes.* »[PAU16].

Pour ce qui est de la reconnaissance d'*entité nommée*, le système les repère grâce à cette cascade qui « *utilise des ressources lexicales et des descriptions locales de motifs, des transducteurs qui agissent sur le texte par des insertions, remplacements ou suppressions*. »[CasEN11].

Le principe de la cascade réside dans la possibilité d'au choix « *utiliser dans des les descriptions suivantes les motifs déjà détectés ou, au contraire, d'éviter un étiquetage non souhaité pour un motif déjà reconnu*. »[CasEN11]

Tout d'abord, CasEN s'appuie sur le découpage en phrase d'*Unitex* – la cascade utilise un *graphe-dictionnaire*, et autres dictionnaires qui contient un inventaire de noms propres, nombres écrits en toutes lettres, prénoms, etc.

Il y a en tout et pour tout cinq catégories de graphes CasEN :

- « graphes de reconnaissance » qui étiquettent les catégories d'*entités nommées* repérées.
- « graphes outils » qui aide au traitement du texte en supprimant, par exemple, des caractères gênants l'analyse (des « . » intempestifs, etc.)
- « graphes de liste » qui vont faire le lien entre la cascade et des listes de mots, chiffres, etc.
- « graphes de masques » pour reconnaître des suites un peu particulières, comme les chiffres romains par exemple.
- « graphes étiqueteurs » qui ajoutent des informations sur des éléments internes à une *entité nommée*.

CasEN fera appel a chacun de ces graphes en fonction des besoins et produira en *output* un texte annoté avec les *entités nommées* catégorisées. Les critères de catégorisation ont été calqués d'abord sur ceux de la conférence MUC-7, puis ceux-ci ont été modifié en fonction des campagnes Ester et Ester 2.

3.4.2.1 Annotation CasEN

- **<persName>** : personne
- **<placeName>** : lieu administratif
- **<geogName>** : lieu géographique
- **<orgName>** : organisation
- **<measure>** : mesure
- **<date>** : date
- **<time>** : heure

Ces annotations sont extraites de la Text Encoding Initiative (*TEI*) [CasENsite]

3.4.3 Le système SEM

Ce système est un « *étiqueteur syntaxique du français dont le but est de proposer le parenthésage en chunks d'une suite de tokens. Le chunking se fait sur la base d'un étiquetage morpho-syntaxique préalable et sous la forme d'une couche d'étiquettes supplémentaire.* »[SEMsite]

Un « *chunk* » correspond à des « *séquences à la fois continues et non-récursives* »[SEMsite]. Aussi, il ne s'agit pas là d'une notion qui correspond à un objet *linguistique* précis, cela dit ces deux propriétés suscitées s'avèrent « *particulièrement pratique dans le domaine du traitement automatique des langues*. »[SEMsite].

Ce système fonctionne en succession d'étapes : tout d'abord, une segmentation (*tokenization*) qui sera suivi d'un étiquetage POS (celui-ci peut être personnalisé jusqu'à un certain point) puis un *chunking* construit sur la base de cet étiquetage.

Aussi, cet outil fonctionne avec des modèles, « *c'est-à-dire des données qui sont le résultat d'un apprentissage automatique et qui servent à paramétrer le programme.* »[SEMsite2] Par ailleurs, l'apprentissage de ce système à été fait sur le *French Treebank* – un corpus arboré de phrase du français tiré du journal *Le Monde*.

La liste des *entités nommées* a été définie selon l'annotation référentielle de ce corpus.

3.4.3.1 Annotation SEM

- **Company** : les entreprises
- **FictionCharacter** : personnages fictifs
- **Location** : lieux (ville, pays, etc.)
- **Organization** : les associations ou organisations (à but non lucratif)
- **POI** : lieux d'intérêts
- **Person** : personnes
- **Product** : produits

3.4.4 Le système Spacy

SpaCy est une librairie pour python gratuite et open-source pour faire des opérations de *traitement automatique des langages naturels*, notamment dans le cas qui nous intéresse, de la reconnaissance d'*entité nommée* et catégorisation de celle-ci. Cette librairie est codée en Cython ce qui lui garantie une vitesse d'exécution particulièrement rapide.

À savoir, cette librairie offre une multitude de processus de *traitement automatique* comme de la tokenisation, POS tags, Words vectors, etc.

Chacun de ces modules à des fonctionnements spécifiques pour toutes langues, censés répondre aux particularités propres à chacune.

Son système de reconnaissance d'*entités nommées* correspond aux « *real-world object* » [SpaCyDOC].

3.4.4.1 Annotation SpaCy

Il existe plusieurs annotations SpaCy dépendantes du corpus sur lesquelles les modèles ont été entraînés.

- Via *corpus wikipedia* :
  - **PER** : nom de personne ou de famille
  - **LOC** : nom d'un lieu politiquement ou géographiquement défini incluant : *villes*, *provinces*, *pays*, *régions internationales*, *mers – océans – rivières, etc.*, *montagnes*.
  - **ORG** : nom d'une entité *étatique*, *industrielle*, ou plus généralement *organisée* (*association*, etc.)
  - **MISC** : entités diverses : *événements*, *œuvres d'arts*, etc.

Nous utiliserons ces annotations-ci dans la mesure où ce corpus *Wikipedia* a été formée dans le but d'entraîner les systèmes de reconnaissance d'*entités nommées* multilingues « *We automatically create enormous, free and multilingual silver-standard training annotations for named entity recognition (NER) by exploiting text and structure of Wikipedia*. » [NRRMC10]. Ainsi, en *français*, ce sont les annotations par défaut.

À Savoir que SpaCy autorise de créer de nouvelles annotations offrant à l'utilisateur la possibilité d'entraîner ses systèmes avec ces nouveaux critères.

3.5 Premier article, présentation, traitement et résultats

Dans le cadre du 20e anniversaire du jumelage des deux villes, <lieu>Avril</lieu>, en <lieu>Meurthe-et-Moselle</lieu> et <lieu>Poissons</lieu>, en <lieu>Haute-Marne</lieu>, <pers>Gérard Gaudry</pers>, artiste, poète et marcheur, né un <date>1er avril</date>, a entrepris de relier les deux villes à la marche, soit un parcours de <mesure>151 km 500</mesure> dont l'itinéraire part d'<lieu>Avril (54)</lieu>, passe par <lieu>Saint-Benoit (55)</lieu>, <lieu>Void Vacon</lieu>, <lieu>Chassey-Beaupre</lieu> pour terminer à <lieu>Poissons</lieu>.

3.5.1 Particularités du premier article

Cet article à cela de particulier de proposer des noms de ville, soit des *noms propres*, construit par *antonomase*. Ceux-ci, *Poissons* et *Avril* pourraient s'avérer complexes à repérer pour des systèmes de reconnaissance d'*entité nommée* dans la mesure où ils pourraient être aisément confondu avec des *noms communs*. À noter par ailleurs que le seul *critère discriminant* est ici la *majuscule* – un critère régulièrement utilisé par les ces systèmes quand la langue à traiter est le français mais qui, nous l'avons vu, peut parfois s'avérer limité. Du reste, les entités nommées de ce systèmes appartiennent à notre *troisième catégorie*, en tant que le nom de ces villes et régions ainsi que de l'artiste n'a, à notre connaissance, pas de résonance au-delà de l'échelle régionale. En ce sens, nous estimons que ceux-ci peineront à être repérés ou précisément catégoriser par ces systèmes.

3.5.2 Résultats statistiques

personne	lieu	mesure	date
1	9	1	1
75.0 %	8.3 %	8.3 %	8.3 %

Tableau i: Entités nommées annotées à la main - compte & pourcentage

personne	lieu	mesure	date
0	0	1	3
0 %	25 %	75 %	0 %

Tableau ii: Entités nommées annotées avec CasEN - compte & pourcentage

personne	lieu	compagnie
1	4	2
57 %	28.6 %	14.3 %

Tableau iii: Entités nommées annotées avec SEM - compte & pourcentage

personne	lieu	organisation-compagnie
3	6	2
18.3 %	54.5 %	27.3 %

Tableau iv: Entités nommées annotées avec SpaCy - compte & pourcentage

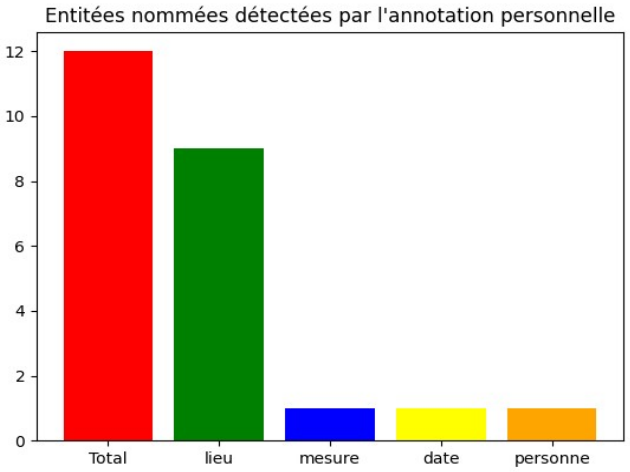
3.5.3 échantillons pertinents

- **CasEN** :
  - <date when="-04-">Avril</date>
- **SEM** :
  - Avril/NC – les deux occurrences du texte
  - (Location Meurthe-et-Moselle/NPP)
  - (Location Poissons/NC) - les deux occurrences
  - Saint-Benoit/NPP
  - (Company Void/NPP Vacon/NPP)

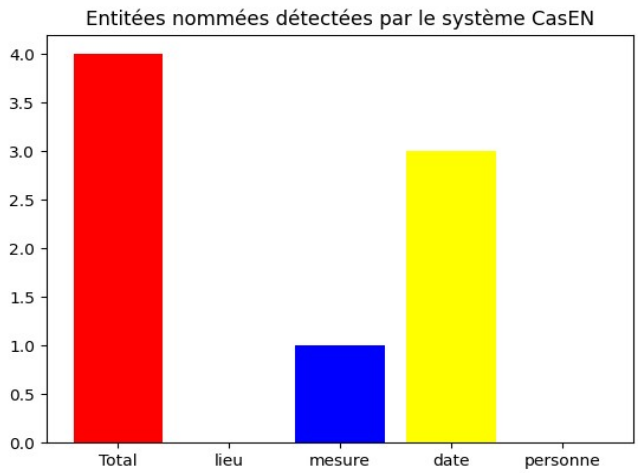
- (Company Chassey-Beaupre/NPP)

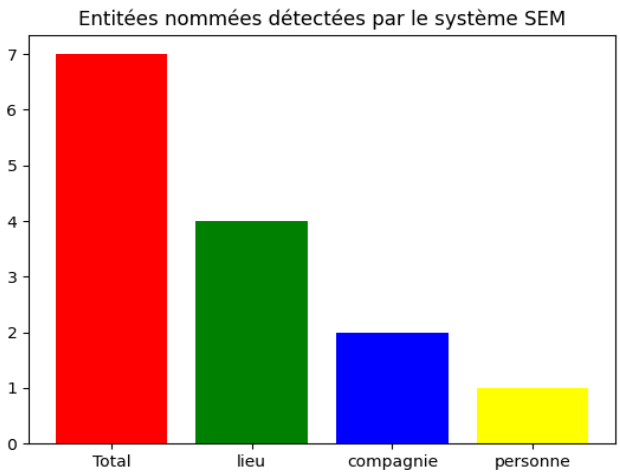
- **SpaCy** :
  - ['Avril', 'ORG'] – pour la première occurrence seulement
  - ['Haute', 'LOC'], ['Marne', 'LOC'] – reconnus comme deux entités nommées
  - ['Void Vacon', 'PER']
  - ['Chassey', 'PER'], ['Beaupre', 'LOC'] – reconnus comme deux entités nommées et catégorisés différemment.

3.5.4 Rapport sur le premier article

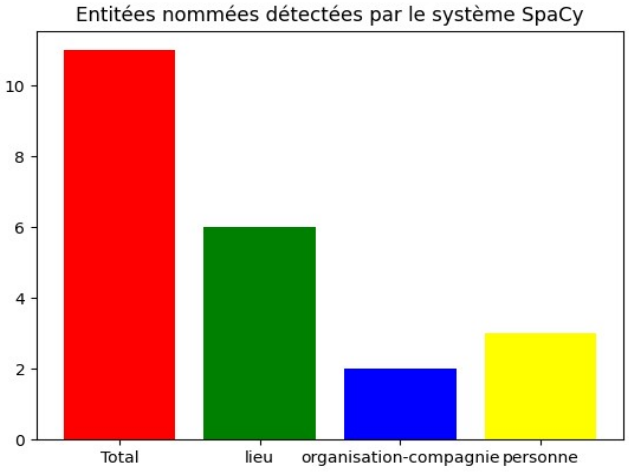


En regardant ces résultats de prêt, nous pouvons remarquer que la ville d'*Avril* n'a jamais été correctement catégorisée. Du reste, selon les systèmes, celle-ci est belle et bien repérée. Par exemple, CasEN repère **Avril** en tant que date *<date when="-04-">Avril</date>* là où SEM considérera ce même *token* comme un *nom commun* (*Avril/NC*) – SpaCy repérera bien la ville qu'il catégorisera par contre comme une organisation (*['Avril', 'ORG']*). Aussi, il est tout à fait probable que cet erreur d'*annotation* soit tout simplement liée au phénomène d'*antonomase* suscitée. De fait, les grammaires ont surtout tendance à traiter ou plus simplement à référer à ces phénomènes que dans le cas où un *nom propre* devient un *nom commun* (phénomène de *communisation*), du reste, celui-ci concerne pour l'essentiel des *anthroponymes* et qu'assez rarement des *toponymes* bien que des cas soit attestés (*un vrai Taj Mahal dites moi !*). L'inverse par contre est suffisamment succinct dans la langue française pour n'avoir occasionné d'étude systématique. Aussi, cet article était prototypique de notre





troisième catégorie, ces villes, régions sont potentiellement trop rares en occurrences dans ces corpus pour que les modèles se calibrent selon celles-ci ou qu'elles soient inclues dans des listes de communes ou régions. En ce sens, nous avons remarqué que CasEN n'a relevé ni les noms de départements ni des villes mentionnées ni le nom de l'artiste contenus dans cet article. Aussi, ce système est une base sur lequel l'utilisateur peut, en fonction de ses besoins, rajouter des graphes pour justement répondre à ce genre d'erreur d'annotation – aussi, rajouter une liste incluant ces noms de communes adresseraient le problème. SEM a bien repéré le nom de l'artiste, par ailleurs, la ville de *Poissons* et le département de *Meurthe-et-Moselle* ont bien été repérés. Par contre, *Void Vacon* ainsi que *Chassey-Beaupre* ont mal été catégorisés. SpaCy semble avoir tendance à séparer les mots composés – du moins lorsqu'il s'agit de nom propre.



Aussi, cet article a permis d'illustrer le fait que le phénomène d'*antonomase* s'avère potentiellement problématique dans le sens *nom commun* à *nom propre* dans la mesure où ce phénomène n'est pas systématique, *anthroponyme* excepté à la rigueur – sachant que la liste de ceux-ci s'actualisent en fonction de la société qui les emplois. De fait, si tant est qu'une nouvelle personnalité s'illustre pour les *français* selon un critère bien particulier, une *antonomase* en sera potentiellement dérivée – cela est inhérent à la nature mouvante du langage.

3.6 Deuxième article, présentation, traitement et résultats

Les rapports épineux entre <pers>Jacques Chirac</pers> et <pers>Jean-Marie Le Pen</pers> reviennent sur le tapis à trois mois de la présidentielle, avec l'évocation d'une rencontre clandestine en <date>1988</date>, qui a obligé l'entourage présidentiel à démentir «toutes relations» entre les deux hommes. Dans son livre «L'homme qui ne s'aimait pas» (<org>Balland<org>), le journaliste <pers>Eric Zemmour</pers> affirme que <pers>Jacques Chirac</pers> a rencontré secrètement <pers>Jean-Marie Le Pen</pers> entre les deux tours de l'élection présidentielle de <date>1988</date> pour lui demander son aide dans son duel contre <pers>François Mitterrand</pers>. «Décidément, rien ne nous sera épargné», soulignait hier un proche du chef de <org>l'Etat</org>.

3.6.1 Particularités du deuxième article

Cet article a cela de particulier de comporter en *entité nommée* des personnalités politiques particulièrement connues dont les noms ont défrayés la chronique, aussi bien en France qu'à l'étranger. Aussi, nous estimons *a priori* que les systèmes n'auront aucun problème à les détecter.

3.6.2 Résultats statistiques

personne	date	organisation
6	2	2
60 %	20 %	20 %

Tableau v: Entités nommées annotées à la main - compte & pourcentage

personne	date	mesure	rôle	organisation
0	2	1	1	1
0 %	40 %	20 %	20 %	20 %

Tableau vi: Entités nommées annotées avec CasEN - compte & pourcentage

personne	organisation-compagnie
6	1
85.6 %	14.3 %

Tableau vii: Entités nommées annotées avec SEM - compte & pourcentage

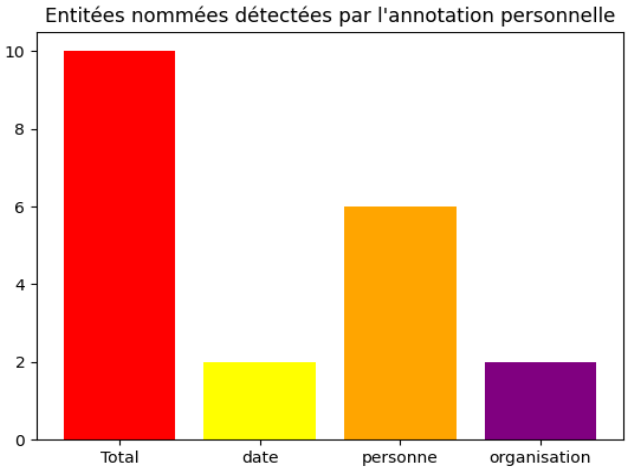
personne	lieu	organisation	divers
9	4	0	2
60.0 %	26.6 %	0.0 %	13.3 %

Tableau viii: Entités nommées annotées avec Spacy - compte & pourcentage

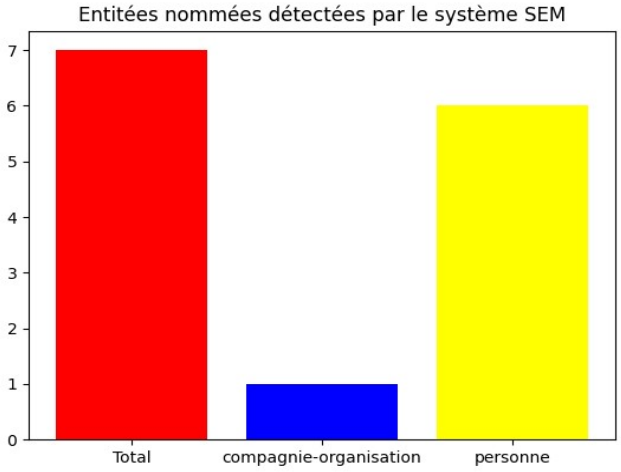
3.6.3 échantillons pertinents

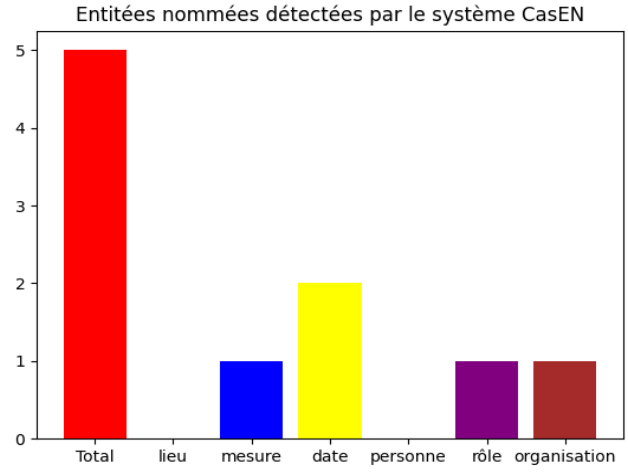
- **CasEN**
  - Aucune personne repéré. *A priori*, nous avons mal utilisé le système
  - <extent>à <measure type="duration" quantity="trois" unit="mois">trois mois</measure></extent>
  - <date when="1988--">en 1988</date>
  - <date type="adverb">hier</date>
- **SEM**
  - relève toutes les personnes de l’article
  - Etat/NPP
  - (Company Balland/NPP)
- **SpaCy**
  - relève toutes les personnes mais délimite mal les frontières
  - ['PER', 'Jacques Chirac'] mais ['PER', 'Jean'] & ['PER', 'Marie Le Pen']
  - ['PER', 'Balland']
  - ['LOC', 'Etat'] – erreur de catégoristation
  - ['LOC', '»'] & ['LOC', '»']
  - ['PER', 'Balland'] – erreur de catégorisation
  - ['MISC', 'Décidément']

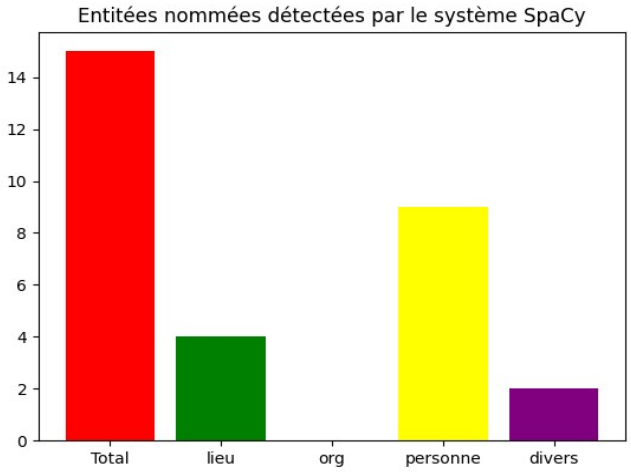
3.6.4 Rapport sur le deuxième article



Nous estimons que notre utilisation de CasEN est déficiente, de fait, ce système a tout repéré avec précision, pourtant aucune des personnes n'a été détectées. Chose d'autant plus étonnante dans la mesure où celles-ci sont relativement connues ; en ce sens, tous les autres systèmes n'ont apparemment pas eu de problème pour les détecter – même si SpaCy les a mal découpées. Quoi qu'il en soit, nous pouvons remarquer que SpaCy et SEM ont effectivement détectés toutes les balises avec une relative efficacitée. En revanche, SpaCy semble peiner à détecter précisément les frontières des *entités nommées* – aussi *Jean* est séparé de *Marie Le Pen*, autant nous ne pouvons que conjecturer sur ce qui en serait la raison, mais il semble que le tiret n'est pas pris en considération lors du processus de reconnaissance des *entités nommées*. Du reste, certaines d'entre elles détectées sont pour le moins surprenantes ; '»' et '«' sont considérés comme telles. Aussi, il est possible que ce soit dû à un caractère particulier propre à ce corpus qui ne soit pas reconnu par le système de SpaCy. Cela dit, il ne serait pas particulièrement compliqué de traiter le texte de façon à les en extraire. Par ailleurs, il y a quelques erreurs de catégorisation chez SpaCy, *État* comme un lieu, *Balland* comme une personne mais surtout *Divertissement* comme un élément divers. Là où SEM n'a pas considéré *État*







comme une organisation mais *Balland* comme une compagnie.

Néanmoins, cet article démontre qu'effectivement plus une personne est est célèbre, et donc susceptible d'apparaître fréquemment dans de nombreux articles – donc corpus, plus leur repérage semble acquis.

3.7 Troisième article, présentation, traitement et résultats

<lieu>MOSCOU</lieu>. Les douaniers russes ont saisi des calendriers japonais décrivant comme territoires japonais <lieu>les îles Kouriles</lieu>, annexées par l'<org>URSS</org> à la fin de la deuxième guerre mondiale et revendiquées par le <org>Japon</org>. Depuis soixante ans, la question territoriale des <lieu>Kouriles</lieu> empêche <org>Moscou</org> et <org>Tokyo</org> de signer un traité de paix.

3.7.1 Particularités du troisième article

Cet article vise à réaliser cette même observation mais cette fois avec des noms de pays. Ceux-ci ont d'intéressant de pouvoir correspondre tout autant à un *lieu* qu'à une *organisation* selon le contexte. Aussi, il s'agira plutôt de s'intéresser à la *catégorisation* proposée par ces systèmes.

De plus, cet article utilise plusieurs mesure de date (qui ne rentre pas dans nos critères d'exactitude) ainsi que des noms d'événements et souhaitons voir comment les systèmes les extrairont, s'ils le font bien entendu.

3.7.2 Résultats statistiques

lieu	organisation
5	2
71.4 %	28.6 %

Tableau ix: Entités nommées annotées à la main - compte & pourcentage

mesure	lieu	spécificité géographique
1	2	1
25 %	50 %	25 %

Tableau x: Entités nommées annotés avec CasEN - compte & pourcentage

lieu	compagnie-organisation
5	0
100 %	0 %

Tableau xi: Entités nommées annotées avec SEM - compte & pourcentage

lieu	organisation
5	2
71.4 %	28.6 %

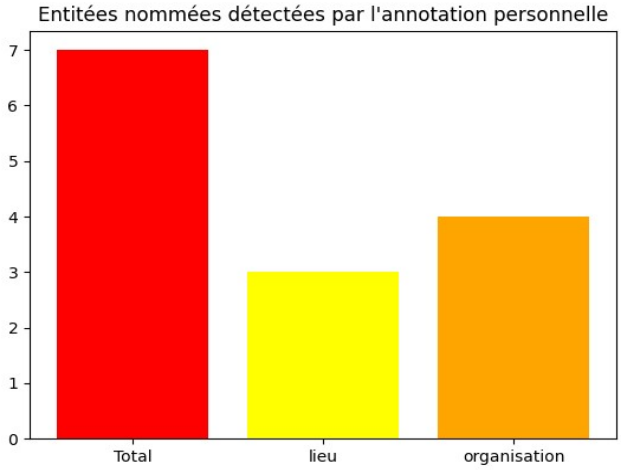
Tableau xii: Entités nommées annotés avec SpaCy - compte & pourcentage

3.7.3 échantillons pertinents

- **CasEN**
  - Les noms de pays « connus » ne sont pas reconnus ce qui nous conforte dans notre idée que notre utilisation de CasEN n’est pas correct.
  - <extent>Depuis <measure type="duration" quantity="soixante" unit="ans">soixante ans</measure></extent>
  - <geogName><geogFeat>îles</geogFeat> Kouriles</geogName>
  - et : <geogName>Kouriles</geogName>

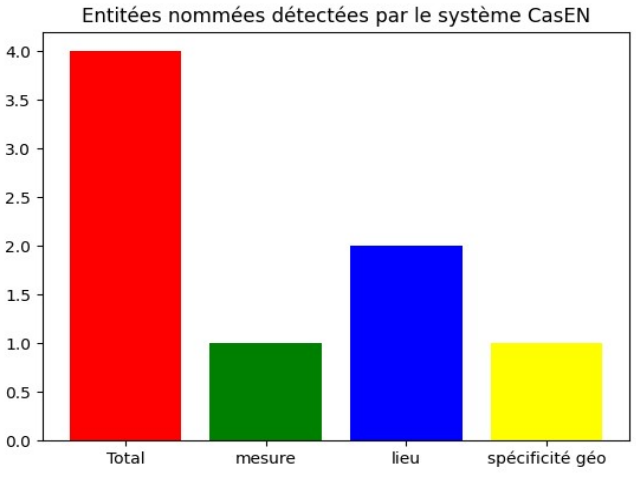
- **SEM**
  - (Location URSS/NPP) selon l'annotation, on ne peut faire la distinction entre un pays qui exerce un pouvoir souverain (soit l'État au pouvoir dans ce dit pays) et le pays géographiquement parlant
  - Kouriles/NPP détectées comme étant un *nom propre* mais pas relevées ni catégorisées
- **SpaCy**
  - toutes les *entités nommées* à détecter l'ont été, découpage parfait
  - ['LOC', 'Moscou'] & ['LOC', 'Tokyo'] & ['LOC', 'le Japon'] comme *lieu*
  - alors que ['ORG', 'URSS']
  - ['LOC', 'îles Kouriles'] et ['LOC', 'Kouriles']

3.7.4 Rapport sur le troisième article



Manifestement, la version de base de CasEN semble nécessiter quelques ajouts de la part de l'utilisateur. Néanmoins, la finesse de son annotation, vis-à-vis de ce qui est effectivement récupéré est remarquablement précise. Aussi, nous pouvons remarqué que les deux autres systèmes ont également réussi à récupérer la majorité des données à traiter – exception faite de SEM qui n'a pas détecté les *kouriles*. Néanmoins, il s'agit d'îles assez peu connues aussi bien d'un point de vue géographique qu'historique, du moins en France – d'autant que son orthographe anglais en diffère, *kurils*.

Du reste, l'intérêt de cet article résidait essentiellement sur la référence plurielle que peut renfermer le nom de pays ; voir de sa capitale ; en effet il s'agit d'une formule très fréquemment utilisée dans des articles de presse où les actions du gouvernement seront évoquées par métonymie comme étant celles des pays.



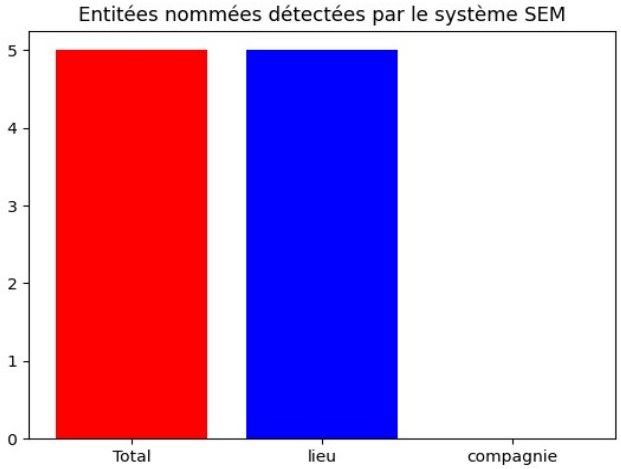
Aussi, la distinction d'un cas type

La France pousse au tourisme

ne sera pas nécessairement évident à saisir. Est-ce qu'il s'agit de :

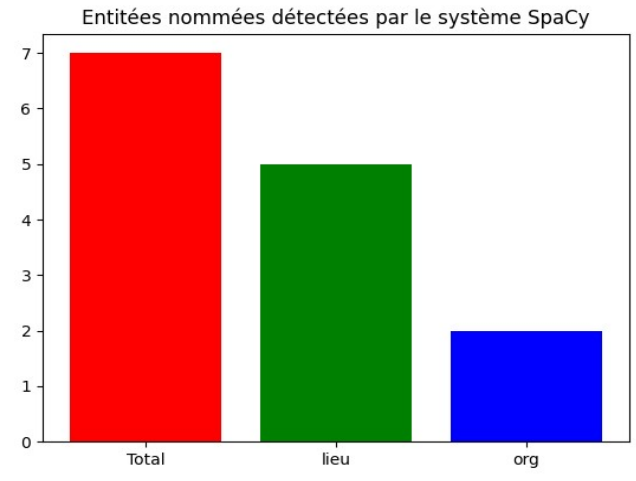
(i) *Les belles villes et paysages de ce pays qu'est la France pousse les touristes à s'y rendre.* ou plutôt de :

(ii) *Le gouvernement français met en place une politique visant à favoriser l'attraction de la France vis-à-vis des touristes.*



A priori, le contexte immédiat permet de dénouer cet ambiguë.

Par exemple, SpaCy a repéré l'*entité nommée* « URSS » comme étant une « ORG » quoique ce système n'a pas perçu « *Moscou* » et « *Tokyo* » comme telles alors que celles-ci sont sujets de la proposition infinitive « signer ». Du reste, nous ne pouvons observer cette distinction avec SEM dans la mesure où celui-ci ne fait pas la distinction entre la géographie et la gouvernance d'un pays dans son tag LOC.



Quoi qu'il en soit, cette distinction n'est pas nécessairement évidente à relever, elle dépend généralement du sens du verbe dont la formule serait sujet, or comme nous venons de le voir la polysémie peut s'y glisser.

3.8 Synthèse sur ces observations

Mesurer la qualité d'un système d'*entité nommée* dans ces conditions, suppose d'adopter une annotation qui correspond à celui-ci. Autant, s'il existe une quantité importante de systèmes particulièrement efficaces, ceux-ci utilisent une annotation qui leur est propre. Du reste, pour définir un objet aux frontières aussi poreuses, évolutives et individuelles que peuvent l'être les *entités référentielles* en *linguistique*, aucune méthode ne serait parfaite. À cela s'ajoute que, bien qu'elles soient peu ou prou semblables, chaque langue possède son propre système qui ne permet pas réellement de développer un système multilingue prêt à faire face à toutes leurs subtilités.

Aussi, l'amélioration de ces systèmes supposerait qu'une théorie linguistique à même de précisément définir leurs problèmes existe – or, lorsqu'on parle de *sens*, ces définitions explicites et unanimement acceptées sont malheureusement rares.

Qui plus est, ces systèmes tendent à être *linguistiquement viable*, mais leur but premier reste de pouvoir être utilisés : ils doivent fonctionner et, de fait, fonctionnent tout à fait correctement. Par ailleurs, chacun fournit un moyen de modifier ses paramètres de façon à l'adapter à nos besoins. En ce sens, ils sont à la fois accessibles pour du traitement de masse, à gros grains, et du traitement précis : dans le cadre d'une étude *linguistique* par exemple.

4 Conclusion

4.1 Des problèmes inévitables

L'*entité nommée* est la parfaite illustration de la nature bivalente du *traitement automatique des langages naturels*, à la fois d'obédience *linguistique* par son objet, et à applicabilité *informatique.* Ainsi, le *TAL* viserait en quelques sortes à la symbiose de ces disciplines *a priori* bien différentes.

En ce sens, certains échecs du *TAL* seraient imputables à la linguistique, du moins, à l'imprécision des notions qui la composent – une résultante directe des multiples courants d'une science jeune et foisonnante qui cherchait à se définir et à s'affirmer comme belle et bien indépendante… des courants multiples et variés, des écoles auxquelles les uns souscriront, honnies par d'autres. À cela s'ajoute selon nous que la *linguistique* n'avait pas eu nécessité d'*accélérer le rythme*, le processus de développement théorique étant d'ailleurs une activité relativement lente, procédant par tâtonnements calmes, mesurés et méthodiques.

Le *TAL* quant à lui est à caractère applicatif, une applicabilité qui génère également d'importants profits, ce qui a conduit les informaticiens à produire et à combler les trous là où la *linguistique* était incapable d'apporter une base précise. Ainsi, comme le note F. Rastier « *Dans leur coopération, linguistes et informaticiens apportent des préoccupations et des points de vue différents. (…) Aussi, en matière de théories linguistiques, les informaticiens font preuve d'une réjouissante absence de pruderie, là où les linguistiques craignent frileusement de s'aventurer hors d'une orthodoxie*. » [RAS91].

Pour le *TAL*, l'applicabilité reste l'objectif fondamental, il s'agit de développer des systèmes qui peuvent être utilisés dans la simple mesure où ils répondent à un besoin. Ainsi, d'autres problèmes pourraient être imputables à la dimension *informatique* du *TAL* où, nécessiteux de produire du résultat, certains ingénieurs ne s'intéresseraient pas à la théorie d'une discipline touchant pourtant à l'objet qu'ils manipulent « *par leur formation, et par leur*

familiarité avec les langages de programmation, les informaticiens sont naturellement attirés par la linguistique formelle dans la mesure où elle utilise les instruments de la logique mathématique pour rassembler sous une même théorie les langues naturelles et les langages artificiels. Toutefois, ils n'ont pas de révérence a priori pour les formalismes inutilement complexes, qui gênent l'implantation informatique, plutôt qu'ils ne la favorisent. »[RAS91]

Du reste, si les théories *linguistiques* s'avèrent demain résolues et unilatéralement déclarées vraies – cela ne les rendrait pas nécessairement traitables en *informatique*, que ce soit en raison de la formalisation binaire de ces données ou plus prosaïquement encore, par limite technique. Un système particulièrement robuste incorporant l'intégralité des règles de grammaire, incluant chaque petit détail de chaque petite occurrence, pourrait tout simplement être trop coûteux pour la machine. Aussi, dans une perspective applicative, nombreux sont ceux qui se rangeraient vers les systèmes déjà existants qui sont, nous l'avons-vu, tout à fait fonctionnels.

4.2 Ouverture et perspectives

Nuançons cette perspective, si cette symbiose apparemment singulière peut occasionner de nombreux problèmes, elle peut tout autant occasionner de très grand progrès, non seulement pour le *traitement automatique des langages naturels*, mais également pour la *linguistique*. De fait, Mańczak Witold remarque les faiblesses statistiques de la *linguistique* :

« *En ce qui concerne les critères de vérité, un abîme sépare la linguistique des sciences naturelles : les naturalistes ne font aucun secret que, pour eux, la pratique, la statistique et l'expérience sont des critères de vérité. Dans cet état de choses, il nous est venu à l'esprit de réfléchir sur les critères de vérité susceptibles d'être utilisés par la science du langage et nous sommes arrivé à la conclusion que les linguistes, en suivant l'exemple des naturalistes, devraient recourir à la linguistique et, exceptionnellement à l'expérience.* » et d'ajouter *: « Si les linguistes décidaient d'agir à l'instar des naturalistes et de confronter toutes sortes d'opinions avec des données statistiques et expérimentales, la linguistique deviendrait une science*

comparable aux sciences naturelles et un grand progrès serait atteint. »[MAN91].

En l'occurrence, une formalisation des connaissances pour permettre une confrontation empirique de ces théories à travers des corpus – qui est du reste, l'objet d'étude de la *linguistique* – pourrait conduire au développement de celle-ci. Par un exemple plus concret, Mańczak soulève par exemple que « Nous n'arrivons pas à comprendre pourquoi on a émis tant d'hypothèses au sujet de la différence entre les noms propres et les noms communs et pourtant personne n'a essayé de les vérifier et de fournir une preuve statistique à l'appui de telle ou telle conception. » [MAN91].

Aussi, si les critères de la théorie *linguistique* s'affine ainsi, ceux-ci seront autant de données traitables pour le *TAL* comme l'évoque F. Rastier « *La conception de l'interdisciplinarité que nous mettons alors en œuvre diffère selon les disciplines. Quand il s'agit d'une technologie comme l'IA, nous cherchons d'une part comment ses formalisations et procédures peuvent être utilisées en sémantique, et d'autre part comment la sémantique peut contribuer aux traitements automatique du langage*. » [RAS91].

Néanmoins, celui-ci met en garde dans la réédition de son ouvrage *Sémantiques et recherches cognitives* de 2010 [RAS91] que « *Plus encore que le poids relatif des différentes disciplines, le mode de l'interdisciplinarité a changé. Alors que les échanges et les débats publics étaient constants, un courant de disciplinarisation, lié à l'élargissement des communautés scientifiques et techniques, a conduit à séparer des domaines académiques : par exemple, les informaticiens connexionnistes ont créé leur propre revues, congrès, associations, etc. ; au lieu d'une alternative théorique globale, leurs travaux se présentent à bon droit comme un complément technique, et si l'efficacité y gagne peut-être, le débat s'est à peu près éteint.* ». L'échange continu entre ces disciplines peut contribuer aux bénéfices de chacune, le risque étant que ces échanges cessent, que priorité soit donné à l'une d'elles au déficit de l'autre.

Bibliographie

PAN-6: Pāṇini, Aṣṭādhyāyī of Pāṇini, VI° siècle avant J.C

MEI21: Antoine Meillet, Linguistique historique et linguistique générale, 1921

SAU10: Ferdinand de Saussure, Linguistique générale, 1910

KLE99: Georges KLEIBER, Problemes de référence : descriptions définies et noms propres, 1999

GAR91: Marie-Noëlle Gary-Prieur, Langue française : syntaxe et sémantique des noms propres. Le nom propre constitue-t-il une catégorie linguistique ? , 1991

W&P92: Robert-Léon Wagner, Jacqueline Pichon, Grammaire du français classique et moderne, édition 1992

AGD89: Michel Arrivé, Françoise Gadet et Michel Galmiche, La grammaire d'aujourd'hui, 1989

LER06: Pierre Lerat, Quelques réflexions sur le traitement terminographique unilingue de la métalangue grammaticale en français, 2006

VAX16: Jean-Louis Vaxelaire, De la définition linguistique du nom propre, 2016

MAN91: Witold Manczak, La nature du nom propre. Prolégomènes., 1991

M&V63: A. Meillet et J. vendryes, Traité de grammaire comparée des langues classiques,

JON93: Kerstin Jonasson, Le nom propre, constructions et interprétations, 1993

SAU16: Ferdinand de Saussure, Cours de linguistique général, 1916

JAK63: Roman Jakobson, Essais de linguistique générale, 1978

ULL52: Ullmann, Précis de sémantique française, 1952

BRE24: Michel Bréal, Essai de sémantique, 1924

JES05: Otto Jespersen, Growth and Structure of the English language, 1905

D&P39: Damourette et Pichon, , 1911 - 1939

RAS95: François Rastier, , 1995

COS52: Eugenio Coseriu, Système, norme et discours, 1952

GAR91deux: Marie-Noëlle Gary-Prieur, Le nom propre constitue-t-il une catégorie linguistique ? , 1991

RAS89: François Rastier, Sens et textualité, 1989

WIKIPARIS: , , , https://fr.wikipedia.org/wiki/Ville_de_Paris

WIL91: Marc Wilmet, Syntaxe et sémantique des noms propres, sous la direction de Marie-Noëlle Gary-Prieur, 1991

DAU65: Maurice Daumas, Les machines à traduire de Georges Artsrouni, 1965

TUR50: Alan M. Turing, Computing Machinery and Intelligence, 1950

HUT07: W. John Hutchins, Machine translation : a concise history, 2007

BHI53: Yeshoshua Bar Hillel, Some linguistic problems connected with machine translation, 1953

COL02: Marcel Cori & Jacqueline Léon, La consitution du TAL, 2002

HUT97: John Hutchins, ALPAC : the (in)famous report, 1997

ENJ05: Patrice Enjalbert, Machine translation: a concise history, 2005

PIE00: Jean-Marie Pierrel, Ingénierie des langues, 2000

RAS91: François Rastier, Sémantique et recherches cognitives, 1991

BNN01: P. Bessières, P. Nazarenko, A. Nédellec, Apport de l'apprentissage à l'extraction d'information : le problème de l'identification géniques, 2001

MUC19: , Site du MUC, , https://dl.acm.org/conference/muc

GSU96: Ralph Grishman & Beth Sundheim, Message Understand Conference - 6 : A Brief History, 1996

CLH93: , Evaluating Message Understanding Systems: An Analysis of the Third Message Understanding COnference (MUC-3), 1993

EHR08: Maud Ehrmann, Les Entités nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation., 2008

CHI98: Nancy Chinchor, Overview of MUC-7/MET-2, 1998

CHI97: Nancy Chinchor, MUC-7 Named Entity Task Definition, 1997

TKM03: Erik F. Tjong Kim Sang & Fien De Meulder, Introduction to the CoNLL-2003 Shared task : language independent named entity recognition., 2003

POI06: T. Poibeau, Extraction automatique d'information. Du texte brut au web sémantique., 2006

SSN02: Satoshi Sekine & Kiyoshi Sudo & Chikashi Nobata, Extend Named Entity Hierarchy, 2002

FRI02: Nathalie Friburger, Reconnaissance automatique des noms propres. Application à la classification automatique de textes journalistique, 2002

VIC05: Rangel Vicente, La glose comme outil de désambiguïsation référentielle des noms propres pures, 2005

WIKI: , , , https://en.wikipedia.org/wiki/Named-entity_recognition

WIKIFR: , , , https://fr.wikipedia.org/wiki/Reconnaissance_d%27entit%C3%A9s_nomm%C3%A9es

DUB94: Jean Dubois, Mathée Giacomo, Louis Guespin, Chirstiane Marcellesi, Jean-Baptiste Marcellesi, Jean-Pierre Mével, Dictionnaire de linguistique, 1994

KLE81: Georges Kleiber, Problemes de référence : Descriptions définies et noms propres, 1981

FEH12: Hela Fehri, Reconnaissance automatique des entités nommées arabes et leur traduction vers le français, 2012

KLE99deux: Georges Kleiber , Problèmes de sémantique la polysémie en question, 1999

RAS94: François Rastier, Marc Cavazza, Anne Abeillé, Sémantique pour l'analyse de la linguistique à l'informatique, 1994

FRI13: Denis Maurel et Nathalie Friburger, CasSys Un système libre de cascades de transducteurs, 2013

PAU16: Sébastien Paumier, Claude Martineau, Unitex 3.1 Manuel d'utilisation, 2016

CasEN11: Denis Maurel, Nathalie Friburger, Jean-Yves Antoine, Iris Eshkol, Damien Nouvel, Cascades de transducteurs autour de la reconnaissance des entités nommées, 2011

CasENsite: , , , https://tln.lifat.univ-tours.fr/version-francaise/anciens-projets/ortolang

SEMsite: Ilaine W., , , https://www.lattice.cnrs.fr/sites/itellier/guide.html

SEMsite2: Ilaine W., , , https://www.lattice.cnrs.fr/sites/itellier/SEM.html

SpaCyDOC: , , , https://tln.lifat.univ-tours.fr/version-francaise/anciens-projets/ortolang

NRRMC10: Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, James R. Curran, Learning multilingual named entity recognition from Wikipedia, 2010