Acquisition de relations entre entités nommées à partir de corpus

Mani Ezzat

Institut National des Langues et Civilisations Orientales

École doctorale N°265

Langues, littératures et sociétés du monde

[ER – TIM, Equipe de recherches Textes, Informatique, Multilinguisme]

THÈSE

présentée par :

Mani EZZAT

soutenue le : 6 Mai 2014

pour obtenir le grade de : **Docteur de l'INALCO**

Discipline : Traitement automatique des langues

Acquisition de relations entre entités nommées à partir de corpus

Thèse dirigée par :

M. Thierry POIBEAU	Directeur de recherche CNRS, Université Sorbonne Nouvelle Paris 3

RAPPORTEURS :

M^{me} Tita KYRIACOPOULOU	Professeur des universités, Université Paris-Est Marne-la-Vallée
M. Christophe ROCHE	Professeur des universités, Université de Savoie

---

MEMBRES DU JURY :

M. Tanneguy DULONG	Directeur technique, Invoxis
M. Mathieu VALETTE	Professeur des universités, Inalco
M^{me} Tita KYRIACOPOULOU	Professeur des universités, Université Paris-Est Marne-la-Vallée
M. Christophe ROCHE	Professeur des universités, Université de Savoie
M. Thierry POIBEAU	Directeur de recherche CNRS, Université Sorbonne Nouvelle Paris 3

ተቋም национален שפה 文化 شرقية
inalco
Institut national
des langues
et civilisations orientales

Remerciements

Je tiens à remercier en premier Thierry Poibeau, pour avoir accepté de diriger ce travail, pour sa compétence sur le sujet, son aide et sa patience à toute épreuve.

Je voudrais également remercier Arisem, qui a rendu tout cela possible, en particulier Nicolas Dessaigne, Sylvain Grand et Tanneguy Dulong sans qui, rien de cela ne serait arrivé.

Je porte également une attention particulière à Stéphanie Brizard, Aurélie Migeotte, Fabrice Blondeau, Michael Remars et Gaël Patin pour leur collaboration et leur enthousiasme qui m'ont donné un cadre de travail plaisant et sérieux.

Merci à Mathieu Valette, Jean-Michel Daube, Monique Slodzian et tout le laboratoire ER-TIM pour avoir élargi mes horizons, notamment sur le domaine du traitement automatique des langues.

Merci à Marguerite Leenhardt pour son soutien sans faille et qui m'a insufflé le courage et la motivation nécessaire.

Enfin, un grand merci à toutes les personnes, qui de près ou de loin, ont contribué à ce travail, que cela soit par des encouragements de parents ou par des tappes amicales dans le dos , qui vous poussent toujours à aller de l'avant. A tous ceux-là, merci infiniment.

Table des matières

Introduction générale

Le développement des données disponibles sur Internet a considérablement changé le domaine du traitement des langues. Les systèmes qui traitaient, il y a peu encore, quelques phrases isolées, doivent maintenant faire face à des déluges de documents variés. Lancés par les conférences MUC (Message Understanding Conference) au début des années 1990, de nombreux travaux ont porté sur un type d'unités appelées entités nommées. Elles correspondent généralement à l'ensemble des noms propres (noms de personne, noms de lieu, etc. ainsi qu'aux dates, etc.) et sont actuellement bien reconnues par les systèmes automatiques. Ces éléments sont importants pour indexer les textes et aider les analystes à en prendre connaissance. Cependant, ces séquences ne deviennent pleinement significatives que lorsqu'elles sont reliées entre elles. Il est par exemple intéressant de savoir qu'un texte contient des occurrences des mots Google et Youtube ; mais l'analyse devient beaucoup plus intéressante si le système est capable de détecter une relation entre ces deux éléments, voire de la typer comme étant une relation d'achat (Google ayant acheté Youtube en 2006).

Les entités nommées ont été l'objet de nombreuses études durant les années 1990. Leur reconnaissance dans les textes a atteint un niveau de maturité suffisant pour aller plus loin dans l'analyse, vers la reconnaissance de relations entre entités. Notre contribution s'appuie sur cet état de fait et s'articule autour de deux grands axes : tracer un contour plus précis autour de la définition de la relation entre entités nommées, notamment au regard de la linguistique, et explorer des techniques pour l'élaboration de systèmes d'extraction automatique qui sollicitent des linguistes.

Contexte et contribution

Si les études sur les relations entre entités nommées abondent aujourd'hui, nul ne peut considérer la tâche comme résolue, et ce, même dans des domaines restreints. La complexité de la langue naturelle reste un frein certain

à la modélisation formelle des relations et à son exploitabilité par un système informatique. Malgré deux décennies de recherche, les dernières conférences sur le sujet rencontrent encore des de nombreuses difficultés.

A cause de cela, nous avons orienté notre démarche vers un double objectif. Il s'agit dans un premier temps d'examiner les relations entre entités nommées au regard de la linguistique et des différentes conférences d'extraction d'information, pour décrire les difficultés rencontrées lorsqu'il est question d'identification de relations dans les textes. Dans un second temps, nous nous intéresserons aux techniques pour l'élaboration de systèmes d'extraction de relations, mêlant conjointement processus automatiques et intervention de linguistes. Notre contribution s'inscrit principalement sur ces deux points.

Précisons également que nos travaux se sont déroulés dans le contexte du dispositif CIFRE (Conventions Industrielles de Formation par la REcherche), qui permet à toute entreprise d'embaucher un doctorant pour le placer au cœur d'une collaboration de recherche avec un laboratoire. Ceci a orienté notre exploration vers des considérations pragmatiques, comme l'exploitation des techniques décrites dans ce mémoire dans un contexte industriel réel.

Plan de mémoire

Notre étude se divise en deux parties. La première s'intéresse aux relations entre entités nommées sous un regard théorique. La seconde décrit la mise en place d'un système d'extraction des relations dans les textes.

Notre première partie débute par un rappel sur les relations entre entités nommées, en faisant un parallèle avec la notion de prédicats (chapitre 1) et introduit d'emblée certaines lignes floues quant au périmètre de cet objet. Nous en analysons la cause dans le chapitre 2 à travers un historique des conférences traitant des relations, en pointant notamment le manque d'homogénéité dans la définition de la tâche d'extraction. Enfin, nous établissons dans le chapitre 3 les difficultés rencontrés lorsqu'il s'agit d'identifier manuellement les mentions de relations dans les textes, du point de vue de l'ingénierie de connaissance et de la linguistique. A l'issue de cette analyse, nous présentons dans le chapitre 4 une expérience et son évaluation, autour de la notion d'incertitude, à travers un schéma d'annotations exploitable et performant.

La seconde partie de notre mémoire aborde des points plus techniques, menant à l'élaboration d'un système d'acquisition de règles à partir de corpus, pour la détection automatique des relations entre entités nommées. Nous commençons par présenter dans le chapitre 5, un état de l'art décrivant quelques domaines d'application, suivi des types de données analysées ainsi que la représentation des résultats, et les techniques existantes menant à ce résultat. Nous mettons en exergue dans le chapitre 6, les différentes contraintes posées pour la suite des expériences, notamment à travers le prisme industriel. Puis nous décrivons dans le chapitre 7 des expériences portant sur la constitution de ressources ontologiques peuplées d'entités nommées et de relations, destinées à améliorer les systèmes d'extractions de relations. Nous tentons de monter comment des données publiques issues d'internet, comme Wikipédia, peuvent être utilisées, pour enrichir et filtrer ces ressources. Dans le chapitre 8, nous décrivons une chaîne de traitements semi-automatique, pour l'élaboration d'un système d'extraction de relations, dans lequel le linguiste a une place privilégiée. Nous en évaluons les performances en utilisant les mesures classiques de rappel et précision, mais aussi en tentant de rendre objectives certaines notions comme la lisibilité des résultats. Enfin, au terme de notre étude, nous évoquerons les différentes pistes d'amélioration et perspectives que nous pensons être adéquates.

En ce qui concerne les corpus utilisés, le travail en entreprise est marqué par les demandes clients, très variées et qui ne permettent pas le travail sur la durée sur un même corpus. La thèse n'est donc pas bâtie autour d'un corpus unique mais elle a impliquéede nombreux corpus, de taille et de thèmes différents. La plupart étaient cependant basés sur des articles de presse, issus de journal Le Monde principalement. L'annexe A présente une description des corpus utilisés dans nos travaux.

Première partie

Les relations entre entités nommées, un objet défini par l'extraction d'information

Chapitre 1

Définition : un préambule

Si la notion d'entités nommées a pu faire l'objet d'études aussi bien en linquistique qu'en traitement automatique des langues, les relations entre entités n'ont pas autant retenu l'attention. La question des relations entre entités a émergé, dès la fin des années 80, au centre du domaine de l'extraction d'information, et a été un sujet de prédilection de plusieurs conférences depuis lors. Au fil du temps, les principales caractéristiques de cet objet ont été mises en lumière. Nous nous inscrivons dans la foulée de ces travaux sur la question, qui font pour nous figure de références. Dans ce préambule, nous adossons notre proposition de la notion de relation entre entités nommées aux éléments de représentation que nous fournissent la logique des prédicats et le modèle actanciel de Tesnière (Arrivé, 1969; Tesnière, 1959).

1.1 la représentation des structures prédicatives : entre logique des prédicats et modèle de Tesnière

La logique des prédicats, ou logique du premier ordre, peut être mise en correspondance avec la relation entre verbes et actants chez Tesnière au moyen d'une variation terminologique : le verbe de Tesnière correspond au prédicat des logiciens, et les actants correspondent aux arguments. Les logiciens décomposent les énoncés en empruntant aux mathématiques la notion de fonction. L'énoncé *Pierre tombe* peut être décomposé de la façon suivante :

- *tombe* est un élément stable appelé « prédicat », qui est incomplet car il lui manque un argument
- *Pierre* est un élément variable qui sature le prédicat pour former avec

lui une proposition.
Mais le prédicat, comme toute fonction mathématique, peut avoir un nombre d'arguments variables. Ce qui amène les logiciens à définir les prédicats par le nombre d'arguments afin de constituer un énoncé complet. Ainsi, les prédicats « unaires », sont ceux qui n'appellent qu'un seul argument (*Pierre tombe, Pierre dort*). De même, les prédicats « binaires » doivent être saturés par deux arguments (*Pierre aime Marie*), et les prédicats « ternaires » par trois arguments (*Pierre donne un cadeau à Marie*). Les logiciens empruntent également aux mathématiques une représentation comparable, avec les arguments entre parenthèses qui suivent le prédicats :

- tomber(Pierre)
- aimer(Pierre, Marie)
- donner(Pierre, cadeau, Marie)

Dans le modèle linguistique de Tesnière, l'énoncé peut être organisé de la même manière. Les verbes sont assimilés aux prédicats et les arguments sont appelés actants. Les verbes sont également classés en fonction du nombre d'actants, c'est-à-dire leur valence. A l'instar des prédicats, ils peuvent être « monovalents », « bivalents » ou « trivalents ».
Il existe cependant des différences importantes entre les prédicats des logiciens et les verbes chez Tesnière. Pour les logiciens, le nombre d'arguments d'un prédicat est un élément définitoire : tout changement de nombre d'arguments correspond à un changement de prédicat. En linguistique, il est fréquent qu'un actant d'un verbe n'apparaisse pas dans un énoncé : l'actant et l'argument n'ont pas le même comportement. Cette différence est matérialisée par la notion sémantique de valence. Lorsqu'un actant n'est pas exprimé syntagmatiquement, il l'est implicitement sur le plan sémantique. Ainsi la construction de l'énoncé est complète car le verbe est saturé par les arguments attendus.
L'autre différence importante entre prédicats logiques et théories linguistiques se situe dans l'opposition que Tesnière effectue entre actant et circonstant. Les actants désignent les acteurs du procès tandis que les circonstants expriment le support dans lequel se déroule le procès. Ce sont généralement les circonstances de temps, de lieu ou de manière. Les circonstants ne trouvent aucun équivalent dans la logique des prédicats (Touratier, 2004). Ils sont simplement considérés comme des arguments supplémentaires. La théorie linguistique de la valence a donc une portée syntaxique ignorée par la logique des prédicats.
Ce rapide panorama des approches, logique d'un côté, linguistique de l'autre, fournit des éléments en faveur d'une approche d'inspiration linguistique pour

aborder les relations entre entités nommées.

1.2 Modélisation des relations entre entités nommées

A partir des considérations préalables que nous venons d'évoquer, nous allons tenter de circonscrire le périmètre de l'objet d'étude auquel nous nous intéressons dans le cadre de ce travail. Les entités nommées correspondent généralement à l'ensemble des noms propres augmenté d'expressions numériques ou spatio-temporelles. Elles peuvent également faire référence à des termes dans un domaine spécifique. A la façon de la terminologie, les entités nommées nommées sont traditionnellement considérées comme directement référentielles, d'une manière univoque, aux objets du monde. Mais cette conception est aujourd'hui souvent contestée, car *elle rend compte d'une façon très simplifiée de la complexité de la langue* (Poibeau, 2005). Cependant, il est possible de définir une relation entre entités nommées dans un texte en nous appuyant sur la théorie de la logique des prédicats : une relation entre entités nommées est un prédicat dont les actants sont réalisés par des entités nommées. De ce point de vue, ils'agit donc de la spécialisation d'une relation de prédication, en posant une contrainte sur le type des actants. La relation peut alors être conçue comme un lien significatif, matérialisé par un prédicat entre entités nommées dans un énoncé. Les notions d'instance et de référentialité, directement héritées des modèles de représentation des entités nommées, tiennent également ici une place importante : plusieurs mentions d'une relation peuvent faire référence à la même instance, identifiée idéalement de manière unique. Parmis les instances, il est possible d'en distinguer deux types :

1. Les *relations statiques*, qui sont essentiellement des états. Ce qu'on appelle état se caractérise par l'absence de changement. Un état qui est vrai pour un intervalle temporel donné, est vrai pour tout point de cet intervalle. C'est donc un lien stable à un instant t entre deux entités nommées.
2. Les *évènements*, qui peuvent être assimilés à une phrase d'action, mettant en cause plusieurs entités, apportant une nouvelle information sur les participants qui peuvent éventuellement avoir une localisation spatio-temporelle.

Dans le cas des relations statiques, les circonstants peuvent parfois être

considérés comme optionnels. Leur présence n'influe pas sur la résolution de la référence de l'instance en question. On dit alors qu'ils ne sont pas obligatoires à l'instanciation de la relation. Exemple :

- Arisem est une filiale de Thalès.
- Arisem est une filiale de Thalès depuis 2004.

L'information apportée par « depuis 2004 » ne change en rien le fait auquel font référence ces deux énoncés. Ce n'est pas le cas pour certains évènements :

- Nicolas Sarkozy a rencontré Jacques Chirac le lundi 24 Janvier lors d'un entretien confidentiel.
- Nicolas Sarkozy a rencontré Jacques Chirac jeudi dernier.

La date joue ici un rôle important car elle différencie les deux évènements qui n'ont pas ici, la même référence. Selon les cas, les circonstants peuvent alors définir la référentialité d'un évènement. Mais ils peuvent être implicites ou explicites. En ce sens, nous pouvons faire un rapprochement entre les relations entre entités nommées et les théories linguistiques de la valence, car celles-ci admettent la variation du nombre d'arguments pour un prédicat donné.

Ce préambule permet d'esquisser notre objet d'étude. Celui-ci demeure approximatif : ceci est dû au fait que nous avons affaire en réalité à un objet émergent, abordé dans le cadre de différentes conférences dont l'existence, souvent guidée par les applications, n'a pourtant jamais fait l'objet d'une définition théorique. Et les limites de ce préambule apparaissent ainsi rapidement. Tout d'abord, une relation sémantique peut exister entre deux entités nommées sans que celle-ci soit matérialisée par un schéma prédicatif dans les textes comme le montre l'exemple suivant :
« Les comportements de François et Pierre sont très différents. Le cadet a été un nourrisson câlin, sensuel. Tandis que l'aîné est désireux de comprendre les règles des adultes. »
Le lecteur comprend aisément que François et Pierre sont frères, même si le texte ne le précise jamais dans un schéma prédicatif tel que « Pierre est le frère de François ». Dans d'autres cas, il est difficile de réduire l'information à une structure prédicative. Prenons l'exemple d'un attentat terroriste. Cet évènement à une localisation ou une date. Potentiellement, il a été revendiqué par un groupe et a fait un certains nombres de victimes. Dans une tâche d'extraction d'information typique, ces différents éléments seront réunis dans un formulaire. Mais ils ne peuvent pas être réduits par une représentation

prédicative. Il s'agit plutôt d'une chaîne de prédication réunie sous le faisceau d'un même évenement. L'exemple de l'attentat terroriste n'est pas anodin et provient d'une conférence d'extraction d'information. C'est pourquoi, dans le chapitre suivant, nous effectuerons tout d'abord un historique de ces conférences. Puis nous verrons les limites et les difficultés qui en découlent, malgré un nombre toujours croissant de travaux.

Chapitre 2

Historique des conférences d'extraction d'information

2.1 Extraction d'information, quelques notions

La notion de relation émerge dans le domaine de l'extraction d'informations, que l'on peut définir comme l'ensemble des méthodes permettant d'identifier l'information contenue dans un grand volume de documents et de les organiser de manière structurée ; autrement dit, il s'agit de retrouver certaines entités textuelles pertinentes par rapport à un besoin, et éventuellement de les annoter. Grishman (1997) propose une définition plus intéressante de la tâche, en faisant référence explicitement à l'extraction de relations : « C'est l'identification d'instances particulières de classes d'évènements ou de relations dans un texte en langage naturel, et l'extraction des arguments de ces évènements ou relations ». A ce titre, l'extraction d'information implique la création de structures de représentation de l'information recherchée et selon la formule de Poibeau (2003), « c'est l'activité qui consiste à remplir automatiquement une banque de données à partir textes écrits en langue naturelle ». Ces deux définitions font émerger clairement la notion de structuration des données : des entités sont groupées dans une structure tabulaire. Les modèles de la logique du premier ordre, notamment le modèle relationnel des bases de données, s'imposent naturellement comme les modèles de représentation de l'information extraite.

Cette idée de réduire l'information à travers une telle structure de représentation a déjà été émise précédemment par Harris dans les années 1960. Nous pouvons à cet égard citer les travaux de Sager et al. (1987) de l'université de New York, qui ont mené à la mise en place d'un système permettant de structurer l'infor-

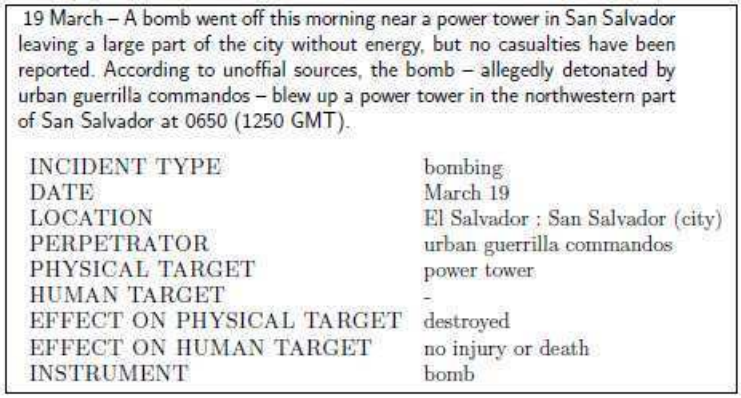
mation extraite dans des textes médicaux. Mais ce n'est qu'à la fin des années 80, à travers les conférences MUC (*Message Understanding Conference*), dont les campagnes s'étalant sur une décennie, que la notion d'extraction d'information a réellement été formalisée. Viennent ensuite les conférences ACE dans les années 2000 (*Automatic Content Extraction*), qui contribuent à l'avancée des techniques d'extraction d'informations, notamment pour les relations entre entités nommées. Enfin plus récemment, les conférence TAC (*Text Analysis Conference*) prennent le relais, tout en montrant les limites des approches développées précédemment. Nous décrirons l'historique de ces conférences en mettant en exergue les différentes évolutions qui ont permis l'émergence d'un nouvel objet, les relations entre entités nommées.

2.2 Message Understanding Conference (MUC)

Les conférences MUC ont été lancées sous l'impulsion du département R&D de la NOSC (Naval Ocean System Center) avec le support de l'agence de la défense américaine pour les projets de recherche (DARPA).
Les conférences MUC prennent la forme de campagnes dans lesquelles il s'agit de soumettre un système d'extraction d'information à une évaluation pour ensuite comparer les approches lors d'une manifestation rassemblant différents participants. Ces derniers reçoivent à l'avance le type de textes et les indications sur le type d'information à extraire à partir desquels chaque participant développe un système d'extraction. Celui-ci est par la suite évalué par rapport à un corpus manuellement annoté.

La plupart des systèmes d'extraction d'information issus des conférences MUC sont élaborés selon le même principe général. Dans un premier temps, le système extrait les entités candidates estimées pertinentes pour une application donnée , à partir des textes en utilisant une analyse locale. Puis, à l'aide de cette première analyse, d'autres éléments d'information, dans une fenêtre plus large, sont repérés notamment grâce à un processus d'inférence. Enfin, ces évènements sont filtrés selon l'information recherchée et intégrés à des formulaires structurés. Selon la terminologie de MUC, ces formulaires sont appelées *scenario template* dont un exemple est présenté dans la figure 2.1. Ce formulaire est une sorte de résumé de l'information recherchée dans un texte. Il contient une dizaine de champs à remplir et porte sur un type d'évènement particulier (ici, des actes terroristes).
La première conférence de 1987 porte sur des messages militaires de la Navy et s'inscrit dans une démarche plus exploratoire, dans laquelle la tâche d'extraction n'est pas définie précisément. Aucune indication n'a réellement été

FIGURE 2.1 – exemple de formulaire de MUC



19 March – A bomb went off this morning near a power tower in San Salvador leaving a large part of the city without energy, but no casualties have been reported. According to unoffial sources, the bomb – allegedly detonated by urban guerrilla commandos – blew up a power tower in the northwestern part of San Salvador at 0650 (1250 GMT).

INCIDENT TYPE	bombing
DATE	March 19
LOCATION	El Salvador : San Salvador (city)
PERPETRATOR	urban guerrilla commandos
PHYSICAL TARGET	power tower
HUMAN TARGET	-
EFFECT ON PHYSICAL TARGET	destroyed
EFFECT ON HUMAN TARGET	no injury or death
INSTRUMENT	bomb

fournie et chaque participant définit le format de sortie ainsi que les informations à extraire.

La seconde conférence de 1989 marque un premier tournant. La tâche s'articule autour du remplissage de formulaires dont le format est fourni à l'avance et une évaluation formelle est introduite avec des métriques précises (rappel, précision, F-score). La troisième conférence de 1991 marque un changement de domaine ainsi que la complexification graduelle de la tâche d'extraction. Il s'agit dorénavant de repérer des évènements relatifs au terrorisme en Amérique du Sud dans des textes fournis par une agence de presse. Les formulaires comportent près du double de champs à remplir par rapport aux conférences précédentes (jusqu'à 18). La conférence suivante, en 1992, suit le même sillon avec une nouvelle augmentation du nombre de champs à remplir (jusqu'à 24).

La cinquième conférence MUC, qui a lieu en 1993, marque le début d'une évolution importante des objets textuels à repérer et à extraire. Le domaine des textes à analyser traite des opérations commerciales conjointement menées par des entreprises, tout en augmentant la complexité de la tâche. Il y a maintenant 11 « scenario template » et un total de 47 champs à remplir, soit le double de l'année précédente. Avant MUC-5, un « scenario template » correspondait au résultat d'une analyse sur un seul texte. De ce fait, le nombre de champs à remplir était assez conséquent. A partir de MUC-5, un formulaire est dédié à un évènement plus circonscrit : il y a moins de participants pour chaque évènement et la granularité sémantique de chaque formulaire est plus fine. En d'autres termes, il y a une volonté d'associer des faits de plus en plus locaux dans les textes à un seul et unique formulaire.

MUC-5 introduit également le multilinguisme.

La sixième conférence MUC, en 1995, continue dans cette voie avec la volonté de définir plusieurs composants indépendants et réutilisables dans les systèmes d'extraction d'information. Le plus important est la détection des entités nommées, qui s'attache à extraire les noms propres (personne, lieu, organisation). En réalité, les systèmes s'intéressaient déjà à ce qui s'appellera par la suite entités nommées, en extrayant déjà des entités bien particulières comme les organisations, les dates ou des expressions numériques. Durant cette édition de MUC, cette tâche liée aux entités nommées a fait l'objet d'un cadre définitoire et d'une évaluation plus précise. En parallèle, un autre formulaire a été crée en conjonction de la tâche de détection des entités nommées : le *template element*, qui agrège en quelque sorte les informations relatives à une entité nommée. MUC-6 marque également un tournant au ni-

FIGURE 2.2 – exemple de template element

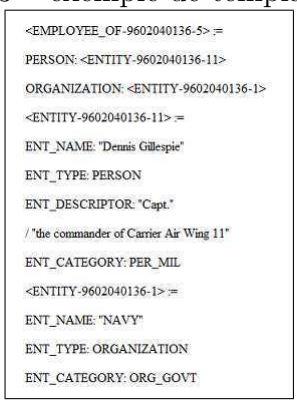
```
<ORGANIZATION-9402240133-5> :=
    ORG_NAME: "Coca-Cola"
    ORG_ALIAS: "Coke"
    ORG_TYPE: COMPANY
    ORG_LOCALE: Atlanta CITY
    ORG_COUNTRY: United States
```

veau de la performance des systèmes. Pour la détection des entités nommées les F-score vont jusqu'à 97 (pour l'anglais et sur des catégories restreintes), tandis que les classiques « scénario template » peinent à dépasser les 50%.

Enfin la septième et dernière conférence MUC, en 1998, introduit une nouvelle tâche : l'extraction de relations entre entités nommées, définie en tant que relations significatives et locales entre deux entités nommées. Trois classes de relations sont arbitrairement choisies (*employee of*, *location of*, *product of*) sans préciser plus avant le périmètre de ce qu'elles recouvrent (Chinchor, 1998b). Les textes proviennent d'articles de presse, montrant la volonté de MUC de ne plus se restreindre à un domaine spécifique. Le formulaire correspondant à cette tâche se nomme *template relation*.

Conclusion S'étalant sur une dizaine d'année, les conférences MUC ont fait évoluer la tâche de l'extraction d'information. Partant au départ de formulaires générique de plus en plus complexes (MUC 1 à 4), la tâche d'extraction d'information a évolué d'année en année pour finalement faire émerger de nouveaux objets aux périmètres de plus en plus précis.

FIGURE 2.3 – exemple de template relation



Les entités nommées ont bénéficié d'une attention particulière. Les performances des systèmes, au terme des deux dernières conférences, se rapprochent des performances humaines. L'intérêt pour les entités s'explique par le fait qu'on peut les retrouver dans tout type de textes. Elles constituent une ancre de base permettant aux systèmes d'extraction de rendre compte de l'information contenue dans un texte et d'identifier les actants des évènements relatés et recherchés.

Les relations entre entités nommées n'ont finalement été abordées de façon spécifique que dans le cadre de la dernière conférence MUC, en se concentrant sur trois classes de relations. Cependant, cet objet était, de façon implicite, présent dans toutes les conférences par le biais des *scenario template*. En effet, la détection de faits locaux dans des textes pour le remplissage de formulaires se rapproche grandement de la tâche d'extraction de relation. Mais au terme des conférences MUC, la définition des relations entre entités reste encore obscure. Par exemple, la différence entre un *element template* et un *relation template* n'est pas très explicite, car il s'agit souvent dans les deux cas, d'entités nommées qu'on a regroupées dans une structure tabulaire. A partir des années 2000, le programme ACE (Automatic Content Extraction) reprend le flambeau et fournira de plus amples précisions quant à la tâche d'extraction de relation entre entités nommées, notamment par le biais de guides d'annotation manuelle, élaborés en prenant en compte les spécificités qui y sont liées.

2.3 Automatic Content Extraction (ACE)

Le programme ACE débute durant l'année 2000. La dernière campagne d'évaluation date de 2008. Contrairement à MUC, le programme ACE ne se focalise pas sur des domaines particuliers. En effet, il s'agit d'un programme orienté sur les technologies d'extraction plutôt que sur les applications. C'est le *Linguisitic Data Consortium* (LDC) qui se charge de l'évaluation. Le LDC est un consortium de recherche et de développement qui s'intéresse aux technologies liées à la langue et vise à partager des ressources et définir des standards. Pour ACE, le LDC fournit les corpus et ressources linguistiques, ainsi que des recommandations d'annotation. Les corpus rassemblent des articles de presse. Ces corpus peuvent également contenir des enregistrements audio retranscrits ou de textes générés par des technologies de reconnaissance optique de caractères (ROC).

Historiquement, il est possible de diviser ACE en trois phases distinctes. ACE Pilot (2000) et ACE Phase 1 (2001) marquent le début du programme. Les données ainsi mises à disposition sont en anglais. Seule la reconnaissance d'entités nommées est traitée et une première version des recommandations pour cette tâche est fournie. Les entités nommées sont catégorisées selon 7 classes distinctes (« Personne », « Organisation », « Lieu », « Véhicule », « Arme », « Bâtiment » et « Entité Géo-Politique »), elles-mêmes subdivisées en plusieurs classes (par exemple, les sous-types de la classe « Organisation » peuvent être « commercial », à « but non lucratif », « gouvernemental », etc.). Selon ACE, cette tâche est le cœur du programme car elle est le point de départ obligatoire pour les autres tâches.

ACE 2002, ACE 2003 et ACE 2004 forment un autre cycle. La tâche d'extraction de relation est introduite. Il s'agit de relations entre entités nommées portant sur 6 catégories (également subdivisées en plusieurs sous-catégories, cf. figure 2.4).

Deux nouvelles langues sont introduites dans le programme : l'arabe et le chinois. Des corpus sont également créés à cet effet.

ACE 2005, 2007 et 2008 forment un troisième cycle. Une nouvelle tâche vient se joindre aux précédentes : l'extraction d'évènements. Selon ACE, cet objet est similaire aux relations entre entités nommées mais reste différent : un évènement est caractérisé par un changement d'état et comporte un cadre spatio-temporel. Il y a 8 catégories (« Life », « Movement », « Business », « Conflict », « Contact », « Justice », « Personnel ») également subdivisées

FIGURE 2.4 – Typologie des relations dans ACE

Type	Subtype
ART (artifact)	User-Owner-Inventor-Manufacturer
GEN-AFF (General affiliation)	Citizen-Resident-Religion-Ethnicity, Org-Location
METONYMY*	*None*
ORG-AFF (Org-affiliation)	Employment, Founder, Ownership, Student-Alum, Sports-Affiliation, Investor-Shareholder, Membership
PART-WHOLE (part-to-whole)	Artifact, Geographical, Subsidiary
PER-SOC* (person-social)	Business, Family, Lasting-Personal
PHYS* (physical)	Located, Near

en sous-catégories, dans une typologie rendant compte plus précisément des différentes classes et des relations entre elles. Une nouvelle langue fait également son apparition : l'espagnol. Des recommandations pour l'annotation des relations entre entités nommées sont fournies, permettant ainsi de tracer un contour plus net de cet objet. Nous allons synthétiser ceci à l'aide de la dernière version des recommandations datant de 2008.

2.3.1 Fenêtre d'annotation

Si la modélisation et la définition sémantique des concepts de classe de relation s'avèrent être une étape importante et complexe pour la stabilisation d'étiquettes et des notions qu'elles recouvrent, la réalité des mentions des relations dans les textes soulève également des complications. Les diverses recommandations de ACE sont le reflet d'un questionnement important sur les segments de texte qu'il faut annoter. Il n'existe pas réellement de définition précise des relations entre entités nommées qui permet d'annoter un segment en tant que mention de telle ou telle catégorie de relation. Les recommandations de ACE tentent néanmoins de donner quelques indications sur les parties de texte à annoter.

L'impression générale qui se dégage des recommandations des conférences ACE reflète une grande modestie dans les objectifs. Le premier principe porte sur le contexte d'apparition d'une relation. Les guides de ACE précisent explicitement que la tâche de détection de relation se borne à la phrase. Plus

précisément, une des conditions de l'annotation d'une mention d'une relation entre entités nommées et que ces mêmes entités nommées apparaissent dans la même phrase.

L'autre principe qui a attiré notre attention est celui de la « preuve explicite et interne » du lien sémantique de la relation. Les guide de ACE illustrent ce principe par l'explication suivante (adapté de l'anglais) :
Dans l'exemple « Christophe et Jeanne vont à l'école ensemble », on ne peut guère utiliser l'étiquette de relation frère/sœur, même si on découvre plus tard dans le document que Jeanne et Christophe sont effectivement liés par cette relation. En d'autres termes, l'utilisation de connaissances pragmatiques ou de connaissances déduites au sein du même texte ne permet pas de catégoriser une relation s'il n'y a pas dans la phrase une preuve explicite de ce lien. Généralement, cette preuve explicite et interne à l'occurrence se définie au niveau syntaxique. Nous présentons sa typologie.

2.3.2 Typologie des constructions syntaxiques pour la sémantique du lien entre deux entités

Le lien sémantique entre deux entités, en tant que « preuve explicite interne » est classé selon une plage de constructions syntaxiques possibles. ACE fournit à ce propos des recommandations précises sur les constructions syntaxiques à identifier permettant d'annoter ou non une mention de relation. Cependant, ces recommendations illustrent les constructions syntaxiques de l'anglais uniquement. Nous avons donc laissé cette typologie dans sa langue d'origine.

Possessif

– *Example : America's Department of Defense*

Type	Argument 1	Argument 2
Part-Whole.Subsidiary	America	Department of Defense

Préposition

– *Example : Officials in California*

Type	Argument 1	Argument 2
Phisical.Located	Officials	California

Epithète

- *Example : Palestinian leaders*

Type	Argument 1	Argument 2
Org-Aff.Employment	leaders	Palestinian

Coordination

- *Example : he and a hunting partner*

Type	Argument 1	Argument 2
Per-Social.Lasting	He	Partner

Convention

- *Example : Jane Clayson, ABC News, SouthLake Tahoe*

Type	Argument 1	Argument 2
Phisical.Located	Jane Clayson	South Lake Tahoe
Org-Aff.Employment	Jane Clayson	ABC News

Participial

- *Example : the crowd trapped inside the compartment*

Type	Argument 1	Argument 2
Physical.Located	crowd	compartment

Verbal

- *Example : Death Valley is in the Mojave Desert*

Type	Argument 1	Argument 2
Part-Whole.Geo	Death Valley	The Mojave Desert

2.3.3 Modalité

La modalité dans ACE tient une place relativement réduite. Il s'agit de dire si la relation est réelle (« assertée ») ou potentielle (« autre »). Ainsi, il n'y a aucune contrainte au niveau de la fenêtre d'annotation d'une relation car il s'agit juste de donner une valeur d'ensemble sur toute l'occurrence de la relation sans préciser les indices qui ont permis cette interprétation. Laisser trop de liberté d'interprétation aux annotateurs peuvent conduire à de nombreux désaccords.

Récit d'évènement et temps d'énonciation

ACE propose de préciser le temps de l'énonciation selon 4 valeurs possibles :

- Passé, lorsque la relation se déroule avant l'énonciation du locuteur.
- Futur, lorsque la relation se déroule après l'énonciation du locuteur.
- Présent, lorsque la relation se déroule pendant l'énonciation du locuteur.
- Non spécifié.

Pour les trois première valeurs, il s'agit principalement du temps des verbes qui portent la relation. Certains marqueurs comme « ex » dans « l'ex-président » sont également pris en compte.
Une note précise l'utilisation de la valeur « non spécifié » (traduit de l'anglais) :

Le temps ne sera annoté pour les relations que lorsqu'il est explicitement présent dans la fenêtre de la mention de la relation. Pour la plupart des relations exprimées par un syntagme nominal, cela signifiera que la valeur du temps est « non spécifiée ».

La volonté générale de la part d'ACE de ne pas étendre la fenêtre de l'occurrence de la relation (exprimée par un prédicat verbal) et de simplement mettre de coté le problème si le temps de l'énonciation n'est pas marqué explicitement à l'intérieur de la relation.

Les évènements

Les guide d'ACE introduisent également la notion d'évènements, qui suivent le même schéma que les relations, avec néanmoins trois grandes différences.

La première est que les participants à l'évènement ne sont plus forcément au nombre de deux. Un évènement peut avoir une localisation ou une date qui lui est associée. De telles entités nommées seront dits « attributs » de l'évènement, car ils ne sont pas obligatoires pour l'annotation d'une mention dans un texte.

La seconde différence est l'annotation du déclencheur de l'évènement. Il s'agit selon ACE du verbe ou du nom qui porte la sémantique de l'évènement. Par exemple :

- *In 1927, she married William Gresser, a New York lawyer and musicologist.* (Married)
- *Shenson, who was born in San Francisco, was working in London at the time.* (Born)
- *Quaker Oats rejected PepsiCo's takeover offer as too low.* (Takeover)

Enfin, la troisième différence se situe au niveau d'un champ appelé polarité. Celui-ci doit indiquer si l'évènement est caractérisé par une phrase négative ou affirmative. Dans le cadre des relations, si une mention est caractérisée par un verbe à la forme négative, celle-ci ne devait pas être annotée.

2.3.4 Conclusion

Si ACE a permis de définir plus précisément le périmètre de la tâche d'extraction de relations entre entités nommées, certaines recommendations fournies par les guides d'annotations demeurent encore floues. On y occulte certaines problématiques, comme la modalité ou la sémantique des prédicats. Néanmoins, ACE permet une première modélisation des relations (évènements ou relations statiques, arguments optionnels ou obligatoires, structures syntaxiques, contexte des mentions, etc.) qui sera reprise et étendue à la fin des années 2000 par TAC.

2.4 Text Analysis Conference (TAC)

Avec une première édition en 2008, la conférence TAC se situe dans le sillon des conférences ACE, TREC (*Text Retrieval Conference*) et DUC (*Document Understanding Conference*), la première pour sa tâche de question-réponse et la seconde pour sa tâche de résumé automatique. A l'instar de DUC et TREC, la conférence TAC est née sous l'impulsion de l'institution gouvernementale américaine *National Institute of Standards and Technology* (NIST). L'objectif est de fournir un cadre et des infrastructures pour des évaluations à grande échelle. La compétition entre les systèmes n'est pas le but premier de TAC.Il s'agit plutôt d'avoir un panorama de l'avancement de la recherche et des systèmes de l'état de l'art par une évaluation normalisée.

Le cycle de la conférence est défini précisément et s'étale sur une durée d'un an. Au début de l'année, la NIST distribue des données textuelles brutes pour chaque tâche. S'ensuit une période d'incubation d'environ 3 mois durant laquelle les différents participants et responsables affinent la définition des tâches afin de pouvoir fournir des guides précis concernant divers aspects comme la méthodologie d'évaluation, la collection de textes de test ou l'annotation. Les premiers résultats des systèmes sont communiqués aux participants environ 4 mois plus tard. La fin de cycle est consacrée à l'organisation d'un évènement pour la publications d'articles de recherches ainsi que différents ateliers.
TAC est divisé en 3 grandes tâches principales (tac, 2011) : le résumé automatique, les systèmes de question-réponse et la « population » automatique de bases de connaissances. Cette dernière tâche (*Knowledge Base Population*) rappelle naturellement les différents programmes mentionnés précédemment, comme ACE ou MUC mais présente néanmoins des nuances. Nous allons les aborder, en commençant par la définition de la tâche, puis en décrivant les deux-sous tâches associées en s'appuyant sur les guides d'annotation fournies par les organisateurs.

2.4.1 Définition de la tâche « Knowledge Base Population » (KBP)

Le principal objectif de la tâche telle qu'elle est définie dans TAC 2011 est de promouvoir la recherche et d'évaluer la capacité des systèmes automatiques à découvrir des informations sur des entités nommées afin de nourrir un base de connaissances. Une base de connaissance de référence est fournie, ainsi qu'une collection de documents destinée à l'apprentissage des systèmes qui participent à l'évaluation. La constitution de la base de connaissances

de référence s'appuie sur des mini-formulaires décrivant un sujet donné dans Wikipédia appelés *infobox* (fig. 2.5). Une infobox est une table préformatée de données dynamiques qui présente sommairement des informations importantes sur un sujet dans un cartouche ou un encadré placé en général à droite de l'article. Les infobox et leurs paramètres sont régis indépendamment par chaque projet.

FIGURE 2.5 – exemple d'infobox de Wikipédia

La tâche KBP va plus loin que les précédentes campagnes d'évaluation telles que ACE et MUC en aiguillant les travaux vers de nouveaux axes :

- Extraction à grande échelle (sur plus d'un million de documents)
- Utilisation d'une collection de textes représentatifs
- Résolution des coréférences d'entités à travers des documents différents (dans la continuité des prémices de ACE)
- Alignement entre les évènements extraits et une base de connaissance
- Adaptation rapide à de nouvelles relations

La tâche KBP est divisée en deux sous-tâches distinctes : L'« alignement » des entités nommées, où les entités découvertes dans les différents documents doivent être alignées avec les entités de la base de connaissances de référence ; et le remplissage de formulaires où sont ajoutées les informations extraites à propos d'une entité nommée en particulier (cette dernière tâche peut être vue comme une tâche d'extraction d'information traditionnelle).

2.4.2 Alignement des entités nommées avec la base de connaissance

l'alignement des entités se définit comme suit. Etant donnée une requête et un document y répondant, les systèmes participants doivent fournir l'identifiant d'une entité dans la base de connaissance qui correspond à la mention de cette requête dans le document. Plusieurs requêtes peuvent désigner une même entité dans la base en utilsant des variantes de surface (exemple : « Nicolas Sarkozy » et « N. Sarkozy »). Les classes d'entités nommées concernées sont au nombre de deux :

- Les personnes, restreintes aux personnes individuelles humaines (les personnages de fiction ou les groupe des personnes ne rentrent pas dans cette catégorie)
- Les organisations, qui sont des entreprises, des agences ou tout autre groupe de personnes définies par une structure organisationnelle. Les groupes de musiciens sont considérés comme des organisation, tandis que les projets comme les congrès ou séminaires ne le sont pas.

Un exemple typique d'une requête pour l'évaluation de cette tâche (tirée de KBP 2010) se présente de la manière suivante (au format XML) :

```
 <query id="EL000304">
  <name>Barnhill</name>
  <docid>eng-NG-31-100578-11879229</docid>
</query>
```

La requête est marquée par un identifiant (« EL00304 »), et a pour valeur « Barnhill ». Le document « eng-NG-31-100578-11879229 » apporte des informations contextuelles sur « Barnhill ». Les systèmes doivent à partir de ce couple identifier sa correspondance dans la base de connaissance.
Le multilinguisme a également été introduit pour KBP2011. Les requêtes ne sont plus uniquement en anglais mais doivent toujours être alignées avec un concept d'une base de connaissance qui elle, reste toujours en anglais.

2.4.3 Remplissage de formulaires dans KBP

La sous-tâche de remplissage de formulaires s'apparente à une tâche d'extraction d'informations traditionnelle. L'objectif est de collecter des informations relatives à une entité nommée à partir d'un corpus et de les ajouter dans un formulaire. Les emplacement dédiés dans le formulaire sont pré-définis. Par exemple, une personne aura un champ pour sa date de naissance, des liens de parenté ou sa religion. Il s'agit de relations sémantiques dont les classes ont été fixées *a priori*. Un guide d'annotation détaillé est fourni pour chaque relation. La table 2.6 les recense telles que définies dans KBP2011.

FIGURE 2.6 – Relations définies dans KBP2011

Person	**Organization**
per :alternate_names	org :alternate_names
per :date_of_birth	org :political/religious_affiliation
per :age	org :top_members/employees
per :country_of_birth	org :number_of_employees/members
per :stateorprovince_of_birth	org :members
per :city_of_birth	org :member_of
per :origin	org :subsidiaries
per :date_of_death	org :parents
per :country_of_death	org :founded_by
per :stateorprovince_of_death	org :founded
per :city_of_death	org :dissolved
per :cause_of_death	org :country_of_headquarters
per :countries_of_residence	org :stateorprovince_of_headquarters
per :stateorprovinces_of_residence	org :city_of_headquarters
per :cities_of_residence	org :shareholders
per :schools_attended	org :website
per :title	
per :member_of	
per :employee_of	
per :religion	
per :spouse	
per :children	
per :parents	
per :siblings	
per :other_family	
per :charges	

Le type et le format des 42 relations dans le formulaire sont définis précisément. Un emplacement peut recevoir une valeur unique (la date de naissance par exemple) ou multiple (liste des enfants d'une personne). Il n'y a pas de normalisation de valeur. Une expression numérique pourra être écrite en toute lettre. Pour une entité nommée de type personne ou organisation, seule une chaîne de caractères d'une mention est requise. Les références dans la base de connaissance ne sont donc pas utilisées pour cette tâche. Les valeurs dans les formulaires ne sont que des chaînes de caractères qui ne pointent pas vers un concept d'entité nommée dans la base. La tâche considère que l'ajout des relations dans la base de connaissance n'est pas centrale. L'accent est mis sur les entités auxquelles sont ajoutées des informations supplémentaires. L'organisation de la table 2.6 en atteste et propose une vue par rapport à la classe d'entité nommée (personne et organisation) et non par relations. Ainsi, nous pouvons considérer que les résultats de l'extraction sont davantage destinés à une interprétation humaine de l'information : un humain consultera une entité nommées et aura directement accès à la lecture d'attributs supplémentaires découverts dans un corpus. Ce genre de format n'est pas directement exploitable par une machine, dans le cadre d'applications pour lesquelles les références des entités nommées dans la base de connaissances sont nécessaires.

2.4.4 Conclusion

La tâche d'alignement des entités montre une volonté prononcée pour la capitalisation des connaissances dans une base de référence alors que la tâche de remplissage de formulaires sort de cette optique. Pour l'annotation, KBP fournit un guide décrivant les classes de relations et ce qu'elles recouvrent. Cependant, aucune consigne clairement formulée aborde les difficultés dues aux phénomènes linguistiques dans les textes. La consigne générale dictée est de seulement extraire les informations qui seraient suceptibles d'apparaître dans une infobox de Wikipédia.

Au niveau des résultats des sytèmes évalués pour KBP, les performances sont relativement faibles, avec des F-mesure tournant autour de 28%. La robustesse et l'adaptation des différents systèmes sont également remises en cause par une tâche « suprise », qui consiste en une évaluation improvisée des systèmes dans un délai court (4 jours) afin de tester la rapidité d'adaptation des participants. Aucun système n'est parvenu a obtenir un score dépassant les 10% de F-mesure. Nous verrons les raisons de cet échec dans le chapître suivant qui aborde les difficultés liées à l'extraction de de relations.

Chapitre 3

Quelques difficultés liées à l'extraction de relations

Le domaine du Traitement Automatique des Langues (TAL) a accordé, ces deux dernières décennies, une importance croissante aux travaux de recherche portant, non seulement sur les entités nommées, mais aussi sur les relations entre entités nommées (muc, 1995; Appelt et Martin, 1999; Mikheev et al., 1999). A l'origine, objet du TAL défini à partir de besoins applicatifs, pour mener à bien des objectifs de Recherche et d'Extraction d'Information (REI), la relation entre entités nommées s'est avérée être un observable complexe à saisir, à définir et donc à modéliser, parce que tributaire d'une matérialité linguistique soumise aux remous des relations contextuelles. Cela ayant un impact concret sur les systèmes de REI : malgré la variété des stratégies applicatives mises en œuvre – apprentissage statistique supervisé, recours aux méthodes symboliques, pour n'en citer que quelques unes – la littérature [1] souligne le potentiel d'amélioration des performances des systèmes.

3.1 La modalité, peu traitée par la représentation des connaissances en TAL

On a commencé à le dire : les problématiques du TAL quant à l'identification et à l'extraction d'entités nommées sont directement liées aux objectifs qui président à la mise en œuvre des systèmes de REI. Rappelons que tout système de REI est conçu pour la réalisation de deux tâches principales,

1. Nous renvoyons le lecteur aux résultats des campagnes d'évaluation KBP2011, qui proposent des résultats détaillés sur cette question http ://nlp.cs.qc.cuny.edu/kbp/2011/

comme le précise d'ailleurs Jurafsky et Martin (2009) :

- la détection d'éléments d'information pertinents à extraire – des entités nommées notamment – dans des ensembles de textes disponibles en entrée ;
- une fois ces éléments d'information identifiés, il s'agit, pour le système en question, de parvenir à établir des relations entre eux ; ainsi, les résultats produits peuvent être utiles à différentes applications d'aide à la prise de décision ou encore de découverte de connaissances nouvelles.

Lorsque l'on aborde les recherches du domaine, on remarque une grande variation dans l'état de maturité de ces deux domaines et nous allons d'abord en décrire l'écart.

Pour la reconnaissance d'entités nommées, de nombreux outils, données et standards existent, dans différentes langues. Certains travaux (Collins et Singer, 1999) montrent que la reconnaissance d'entités nommées atteint des performances élevées, avec des évaluations présentant des F-scores allant jusqu'à 0.97 pour les catégories d'entités les plus fréquentes – personne, lieu, organisation – dans les corpus de presse. Pour ce qui relève des schémas de catégorisation des entités, les recherches menées par Sekine et al. (2002) ont, *de facto*, constitué un standard, par l'étendue des catégories décrites – plus de 200.

Cela est moins vrai pour les relations entre entités nommées ; cependant, certains travaux proposent des pistes pour pallier cette différence, dont les plus notoires sont très probablement FrameNet (Fillmore et al., 2003) et WordNet (Miller, 1995; Fellbaum, 1998), dont nous allons rappeler les grands principes ci-après, en regard de notre problématique d'extraction de relations entre entités nommées. FrameNet, projet adossé à la théorie dite de la Sémantique des cadres – une théorie lexico-sémantique rattachée aux travaux du linguiste américain Fillmore – est un projet qui mobilise les efforts de chercheurs depuis 1997. Il s'inscrit dans un programme de recherches dont l'objectif est de produire une ressource lexicale, exploitable par une machine aussi bien que lisible par un humain, décrivant le cadre sémantique de quelque 13 000 entrées, dont chacune est accompagnée d'exemples annotés manuellement. L'approche proposée par FrameNet – dont la notion centrale de « frame » n'est pas sans rappeler la notion de valence du modèle de Tesnière – est la suivante. Le cadre[2] de chaque entrée indique la structure d'arguments –

2. « frame » en anglais

autrement dit les rôles sémantiques définis à partir d'une connaissance encyclopédique du lexique – qui lui sont afférents, ainsi que les autres cadres auxquels il peut être relié. Différents types de relations entre les cadres sont distingués, qui se fondent sur les niveaux de description linguistique lexical et propositionnel, reposant essentiellement sur le comportement co-textuel des unités lexicales, sur le plan de leur combinatoire syntaxique et sémantique. FrameNet peut ainsi être conçu comme une ressource intéressante à mobiliser pour l'analyse des relations entre entités nommées.

De la même façon, WordNet, programme de recherches débuté en 1985 et rattaché aux travaux de la linguiste américaine Fellbaum, a vocation à classifier et relier le contenu sémantique des unités lexicales de la langue. Cette ressource organise les unités lexicales par synonymie, ce qui permet de dessiner les contours d'un concept : autrement dit, les unités sont regroupées selon le sens unique auquel elles réfèrent dans la langue, donc selon le ou les concepts sémantiques associées à chaque unité. Plus précisément, la notion centrale de ce modèle est le « synonym set », que l'on peut décrire comme un groupe d'unités lexicales sémantiquement équivalentes ou associées au même référent du monde. Du point de vue de l'extraction de relations entre entités, WordNet pourrait par exemple fournir un groupe de verbes liés à une relation sémantique donnée. FrameNet et WordNet, en particulier parce que ces ressources sont librement accessibles, structurées, de large couverture, d'une qualité lexicographique réputée bonne et adossées à des procédures d'exploitation informatisées, peuvent donc présenter un intérêt dans le cadre de la mise en œuvre de systèmes dédiés à l'extraction de relations. Cependant, il faut indiquer que, confrontés à la matérialité de textes non techniques, ces ressources butent sur des phénomènes fréquents en discours (Valette, 2008) que sont l'ambiguïté sémantique, les chaînes de coréférence et la modalité. En effet, comment traiter une unité lexicale lorsqu'elle prend un sens non prévu par la ressource ? Comment répercuter le sens associé à un référent sémantique lorsque celui-ci est inclus dans une relation anaphorique avec d'autres éléments d'un texte donné ? Enfin, comment gérer l'impact de la modalité sur la sémantique d'une unité ? Il semble légitime de penser que, si ces questions restent entières du point de vue de l'analyse des entités nommées, elles le sont tout autant concernant les relations entre ces entités.

Cette différence de maturité entre les entités et les relations peut s'expliquer par les difficultés qu'engendre l'élaboration d'un système d'extraction de relations. Nous nous proposons de rappeler dans ce chapitre les principales difficultés auxquelles ceux qui, s'intéressant à la même problématique que nous, se trouveront couramment confrontés. Nous avons identifié deux

paradigmes :

1. En premier lieu, la résolution de coréférences, dans le cadre des systèmes dédiés à l'extraction de relations entre entités nommées ;
2. En second lieu, l'annotation manuelle de ces relations dans des documents pris en charge par de tels systèmes ; nous verrons que la tâche d'annotation nous confronte à deux autres problématiques spécifiques, dont la première est liée à l'ingénierie des connaissances et la seconde, à la modalité linguistique.

3.2 La coréférence : quels problèmes pour l'extraction de relations ?

La tâche de résolution de coréférence occupe les esprits depuis plusieurs décennies car les mentions d'entités peuvent être des coréférences. Les premiers travaux proposant des algorithmes permettant de les traiter automatiquement remontent à la fin des années soixante-dix (Hobbs, 1986; Sidner, 1981). Les systèmes automatisés de l'époque ont recours à l'utilisation de ressources linguistiques spécifiques (Brennan et al., 1987; Tetreault, 1999; Strube, 1998). A cet égard, ils diffèrent des approches récentes, qui ont plus fréquemment recours aux méthodes d'apprentissage statistique (Ge et al., 1998; Soon et al., 2001; McCallum et Wellner, 2003; Cardie et Wagstaff, 1999). Il faut noter que le traitement automatique de la coréférence a, bien souvent, été confondu avec celui de l'anaphore grammaticale (Perdicoyanni-Paléologou, 2001; Van Deemter et Kibble, 2000). Les différentes recherches sur le sujet ont d'ailleurs mis en exergue la variété des difficultés rencontrées, lorsqu'il s'agit d'élaborer un système automatique dédié à la résolution des anaphores (Miktov, 1998; Cornish, 1995; Corblin, 2007). Cependant, si ces deux phénomènes sont proches, ils ne sont pour autant pas identiques ; on a pu constater dans la littérature que certaines approches placent sous le paradigme de la coréférence, des phénomènes linguistiques procédant en réalité de relations sémantiques plus complexes (Elango, 2006). D'un point de vue linguistique, il s'agit, pour la coréférence, d'une relation sémantiquement symétrique entre deux segments textuels, indépendants sur le plan de l'interprétation et dont la nature catégorielle est, ou non, homogène. Ce n'est pas le cas de la relation anaphorique, dans le cadre de laquelle un premier segment communique une valeur référentielle à un autre, ultérieur dans la

chaîne textuelle. Ainsi, une anaphore reprend le genre et le nombre de son antécédent :

- Coréférence : Apple a été fondé par Steve Jobs. La firme à la pomme connaît aujourd'hui un succès retentissant. (Apple - firme à la pomme)
- Anaphore et coréférence : Pierre a acheté des stylos. Il les a donnés à Jean. (Pierre - il, stylos - les)

Sous l'angle du Traitement Automatique des Langues, la résolution de coréférences est une tâche à part entière, non triviale, dont les enjeux ont directement maille à partir avec les systèmes d'extraction de relations entre entités nommées. L'impact de la résolution de coréférences sur la performance de ces systèmes, est loin d'être négligeable : cela tient au fait que les mentions d'entités nommées qui composent une relation, sont fréquemment incluses dans une chaîne de coréférence.La reconnaissance est donc partielle si la chaîne de coréférence n'a pas été analysée correctement par le système. Force est de constater que les approches précédant la mise en œuvre d'applications de traitement automatique ne se sont peut être que trop rapidement posé la question de la nature de ces phénomènes, sur lesquels la réflexion linguistique s'est sans doute plus longuement attardée. Toutes choses égales par ailleurs, nous nous en tiendrons à l'idée que la complexité inhérente au traitement automatisé de ces phénomènes renvoie à la problématique de leur modélisation. Différents types de problèmes se posent lorsque de tels systèmes se trouvent confrontés à la matérialité linguistique des textes. En effet, la prise en compte des coréférences pour l'extraction de relations entre entités nommées, a un impact direct sur les performances des systèmes, en termes de rappel. Des travaux récents ont montré que la non prise en compte des coréférences induit une baisse de performance de 30% en français (Ezzat, 2010) et de 25% en anglais (Grishman, 2011). Des cas de figure variés peuvent avoir une incidence négative sur le rappel des systèmes d'extraction de relations, que nous indiquons rapidement ci-après :

1. L'anaphore grammaticale - **Le ministre des Affaires étrangères Alain Juppé** a effectué vendredi 18 novembre une visite à Ankara et Istanbul, où **il** a rencontré le président Abdullah Gül.
2. **Le chef de l'Etat** a annoncé, le 27 octobre à Poligny (Jura), le déblocage de 650 millions d'euros d'aides et 1 milliard d'euros de prêts à taux bonifiés pour soutenir le secteur de l'agriculture qui subit « une crise absolument exceptionnelle ».
3. Les coréférences implicites - Grand changement chez **EA Sports**. Peter Moore, ancien dirigeant de la branche américaine de Sega et de

Microsoft, va prendre la position de CEO [de EA Sports] à la fin de l'année fiscale.

4. Les coréférences marquées - **Twix**, anciennement **Raiders**, racheté par le groupe Mars incorporated.

Lorsque l'argument d'une relation entre entités nommées est instancié dans le cadre d'une anaphore grammaticale, le système doit être en mesure d'en identifier l'antécédent – comme « Alain Juppé » dans l'exemple 1. Chose non triviale, les segments textuels anaphoriques peuvent se présenter sous diverses formes grammaticales et la place de l'antécédent n'est pas toujours simple à identifier – quoique généralement à gauche de l'anaphore. Phénomène du même acabit, l'anaphore pronominale de l'énonciateur – matérialisée par le pronom personnel de première personne, au singulier ou au pluriel – peut poser le problème particulier de la longueur de la chaîne de coréférence à résoudre, notamment lorsque la fenêtre textuelle où elle se trouve est de plusieurs phrases ou paragraphes. Enfin, l'absence d'antécédent grammatical dans le co-texte est un autre verrou fréquent, qu'il est parfois possible de lever en recourant aux informations relevant du hors-texte, telles les métadonnées – on pense par exemple à la métadonnée que représente le pseudonyme d'un auteur de message contribuant à un forum de discussion.

Parmi les mécanismes qui peuvent être mobilisés dans l'activation d'un référent permettant de résoudre une coréférence, on trouve le cas de la coréférence nominale, qu'il est possible de rapprocher de l'anaphore associative, en tant qu'elle repose sur une connaissance du monde supposément partagée, hors du texte. C'est le cas dans l'exemple 2 où le recours à la connaissance empirique – savoir qui est le chef de l'Etat – permet une interprétation correcte de la phrase. Parce qu'elles mobilisent des systèmes de connaissances extra-linguistiques, les coréférences nominales comptent parmi les plus difficiles à résoudre pour un système automatique, que son antécédent ait été mentionné dans le co-texte ou non (Kunz, 2006).

Le mécanisme du sous-entendu est une stratégie courante, notamment pour éviter les répétitions – comme dans l'exemple 3, où nous avons restitué le référent implicite entre crochets. Dans ce cas de figure, auxquels on réfère généralement par la notion de coréférence implicite, l'interprétation correcte de la phrase induit la nécessité de rétablir l'inférence contextuelle. A notre connaissance, l'inférence et la restitution du référent implicite, qui permettent en principe la détection complète de la relation entre entités et de l'ensemble des arguments de cette dernière, ne sont implémentées dans aucun système

d'extraction, à ce jour.

L'exemple 4 donne à voir un cas de coréférence marquée explicitement ; dans ce cas de figure, la chaîne de coréférence est exprimée dans un contexte où, soit les marques typologiques ou de ponctuation, soit la syntaxe de la phrase, permettent d'identifier et de résoudre la chaîne de coréférence. Ce qui, toutes proportions gardées, les rend proches de structures telles que les énoncés définitoires ou les relations sémantiques hiérarchiques – l'hyponymie, l'hypéronymie, la méronymie par exemple.

Notons que la plupart des systèmes participant aux campagnes d'évaluations ne comportent pas de composant dédié à la résolution de coréférences. Dans certaines de ces campagnes, notamment ACE, la résolution de chaînes de coréférences est d'ailleurs considérée comme une tâche indépendante. Afin de donner une idée d'importance du poids de cette question pour la performance des systèmes d'extraction de relations entre entités nommées, il paraît pertinent de rappeler que les occurrences de relations qui comportent des entités nommées liées à une chaîne de coréférence, représentent en moyenne 25 à 30% du nombre total des occurrences d'entités nommées rencontrées dans un corpus de textes à traiter (Ezzat, 2010; Grishman, 2011).

3.3 L'annotation manuelle : deux problématiques spécifiques

Le second paradigme de difficultés que nous souhaitons mentionner ici est lié à l'annotation manuelle des relations entre entités nommées dans les corpus de textes. On l'a évoqué plus haut, la tâche d'annotation nous confronte à deux problématiques spécifiques, l'une renvoyant à l'ingénierie des connaissances, l'autre à la modalité linguistique ; points que nous nous proposons d'étayer ici.

3.3.1 Une question d'ingénierie des connaissances

Quoi catégoriser ? Un problème de représentation et de modélisation des connaissances. Affecter une catégorie à une occurrence de relation entre entités nommées, parmi l'ensemble des catégories disponibles, est loin d'être une action triviale à réaliser. Pour catégoriser une relation entre plusieurs entités nommées, il faut déjà parvenir à catégoriser les entités elles-mêmes, étape qui influe directement sur la nature de la relation qu'il sera possible de déterminer entre elles. Typiquement, cette démarche de catégorisation

des entités s'appuie sur une représentation des connaissances particulière, illustrée dans un modèle. Or, les représentations du sens sont difficiles à stabiliser : la littérature nous en fournit d'ailleurs plusieurs exemples. En effet, on a pu constater un véritable foisonnement typologique sur les catégories, d'une part du point de vue de leur nombre – certains travaux du début des années 2000 ont par exemple développé des systèmes typologiques très précis, comportant plus de 200 catégories distinctes (Sekine et al., 2002) – d'autre part, du point de vue du domaine de connaissance décrites : d'abord génériques, telles les « personnes », « lieux », « organisations » par exemple, des catégories spécifiques à des domaines de connaissances particuliers sont apparues, tels les « véhicules » ou encore les « quantités numériques ». Il faut ajouter que les systèmes, qui ont pris une envergure industrielle importante pour les entreprises, ont été adaptés à des domaines métiers de plus en plus spécifiques, guidés par la demande du marché – on pense par exemple à l'extraction d'entités nommées dans des domaines circonscrits tels que la veille concurrentielle dans le secteur bancaire, ou le système mobilisera des catégories différentes ce celles de la veille concurrentielle dans le secteur de la santé-beauté ou de l'automobile. La littérature indique qu'il aura fallu une dizaine d'années pour que les typologies de catégories d'entités nommées parviennent à se stabiliser.

En ce qui concerne les relations entre entités nommées, le constat est plus mitigé. Cette tâche, émergente, est encore tributaire d'une forte fluctuation typologique. En effet, un examen des relations considérées par les principales conférences du domaine – MUC, ACE et KBP – mène au constat d'une grande variabilité dans les connaissances représentées et leur organisation :

- Conférence MUC - « employee of » (employé de), « location of » (emplacement de), « product of » (produit de)
- Conférence ACE - Typologie générale, qui se veut adaptable à des domaines hétérogènes et à des textes produits sur des supports et dans des contextes variés
- Conférence KBP - Typologie centrée sur les relations dont les arguments principaux sont des personnes ou des organisations

La situation peut être résumée ainsi : à partir de chaque corpus relevant d'un domaine spécifique, un ensemble de catégories est développé, si bien qu'il semble difficile d'en établir un panorama plus ou moins stabilisé, par exemple à partir des relations entre entités les plus fréquemment utilisées.

Couverture sémantique des classes de relations. La représentation et la structuration des connaissances sous forme de typologies de classes de relations entre entités nommées n'est que l'entrée en matière de cet état de fait. En effet, une fois cette étape réalisée, il faut parvenir à préciser ce que chaque catégorie recouvre sur le plan sémantique, car il peut y avoir des ambiguïtés. Examinons les exemples ci-dessous :

1. Dès le lendemain de son arrivée, le mercredi 18 septembre, le président Laurent Gbagbo aura un tête-à-tête suivi d'un déjeuner au Palais du Quirinal avec le président italien, Carlo Azeglia Ciampi.
2. Jacques Chirac a eu un entretien téléphonique avec Laurent Gbagbo au cours duquel les deux hommes ont « évoqué la possibilité que le président ivoirien vienne à Paris après la formation du gouvernement d'union nationale », a annoncé mercredi soir l'Elysée.
3. Gildas le Lidec, a rencontré jeudi le ministre de la Défense ivoirien, Kadet Bertin.
4. Rachat d'un bloc de 20.8% de capital
5. Partygaming a racheté en 2007 le groupe spécialisé dans le poket Empire Online.

Peut-on considérer que les relations identifiées dans les exemples 1 à 3 relèvent du type « Contact »au sens de contact entre deux personnes ? Autrement dit, les prédicats « rencontrer », « avoir un entretien téléphonique avec » et « avoir un tête-à-tête » sont-ils rattachables au même concept de relation sémantique ? Pour les exemples 4 et 5, le « rachat d'un bloc de 20.8% de capital » et « racheter » sont-ils des prédicats à rattacher au même type de relation ?
Ces questions renvoient à la définition d'une relation et à ce qu'elle recouvre sémantiquement. Cela nécessite de devoir trancher selon les cas, lorsque l'on est confronté à une tâche d'annotation. Pour tenter de pallier ce problème, les conférences MUC fournissent un guide d'annotation (muc, 1995) qui peut s'avérer non trivial à appliquer à des cas d'usages spécifiques, rencontrés en corpus. Sur les corpus annotés les plus récents fournis par le Linguistic Data Consortium (LDC, responsables des ressources pour les conférences ACE et KBP), les évaluations montrent que les annotations atteignent 54% de rappel ; il faut ajouter à cela que les annotations ne convergent en moyenne qu'après 6 essais sur corpus, ce qui représente une contrainte non négligeable, mise en regard du coût induit par une démarche d'annotation manuelle. De plus, le LDC met en exergue la difficulté à obtenir des jugements consensuels lors des campagnes d'annotation, à l'issue desquelles la précision est

généralement de 70%. Cela a des conséquences directes le travail préparatoire en amont des campagnes, les guides d'annotation devant être systématiquement révisés et étendus. Certains travaux (Shirky, 2005) proposent par exemple sur ce point une liste de principes généraux à suivre, qui pourraient selon nous former un socle pertinent pour la réflexion sur l'élaboration de typologies de classes de relations ; en particulier :

- Disposer d'un corpus de taille modérée pour élaborer la typologie.
- Formaliser systématiquement les classes, ou catégories, appartenant à la typologie.
- Elaborer un ensemble restreint de types d'entités ou de relations entre entités.
- Déterminer des frontières précises pour définir les différents types.

De surcroît, l'absence d'une définition consensuelle des relations entre entités nommées constitue un obstacle à l'élaboration d'un standard d'annotation. Comme on l'a évoqué plus haut, il s'agit d'un problème d'ingénierie des connaissances : les classes de relations représentées dans une typologie devraient, en principe, être définies en regard d'un domaine donné et, si possible, avec l'aval d'un expert du domaine en question. Etant entendu que, le temps passant, la typologie est potentiellement sujette à modifications, qu'il convient de répercuter sur celle-ci.

Type des entités nommées et métonymie. Le choix de la classe à laquelle il est possible d'associer une entité nommée peut s'avérer problématique en raison des glissements de sens observés dans les textes ; en particulier, les emplois métonymiques sont fréquents et représentent une difficulté majeure pour mener à bien une tâche d'extraction d'entités nommées (Poibeau, 2005). Cette question a, nécessairement, une incidence directe sur le typage des relations entre entités nommées. Selon la définition que nous avons proposée, les relations comprennent plusieurs entités – les arguments de la relation – liées par un prédicat. Il est alors légitime de s'interroger sur l'impact du type de la relation sur la classe des entités nommées qu'elle implique : comment appréhender la définition des relations entre entités, lorsque celles-ci sont tributaires des glissements sémantiques de leurs constituants ? Prenons quelques exemples pour illustrer concrètement notre point :

1. Bill Gates avait racheté le système d'exploitation QDOS pour en faire MS-DOS, puis a conçu Windows, tous deux en situation de quasi-

monopole mondial.

2. Total a racheté Elf-Aquitaine le 22 mars 2000 pour former TotalFinaElf, rebaptisée Total S.A. en 2003.
3. En 1992, la compagnie Motorola a investi 120 millions de dollars américains dans la ville de Tianjin.
4. Il est surprenant de constater que la France est le principal actionnaire de Quick.
5. La France a décidé d'intervenir dans le conflit libyen.
6. Le numéro un mondial de la bière AB Inbev affiche un bénéfice net dopé par le rachat en novembre dernier de l'américain Anheuser-Busch.
7. Après différents propriétaires, le Flamingo est racheté par Hilton en 1992.

Si l'on considère les deux premières entités nommées mises en exergue dans l'exemple 1, l'on peut être tenté de dire qu'il s'agit d'une relation de type « Rachat » impliquant une entité de type « Personne » – Bill Gates – et une entité de type « Artefact » – QDOS. De la même façon, dans l'exemple 2, il est possible de considérer qu'on est confronté au même type de relation, cette fois-ci entre deux entités nommées de type « Organisation » – Total et Elf-Aquitaine. Cependant, s'agit-il vraiment du même type de relation ? Ce pourrait être le cas, si l'on considère que, dans l'exemple 1, l'entité de type « Personne » procède d'un emploi métonymique, dont le référent sémantique réel est finalement une entité nommée de type « Organisation » – en l'occurrence, Microsoft.

Ces considérations nous confrontent à toute l'épaisseur du problème de la représentation des connaissances et de sa mise en œuvre, lorsque l'on passe les concepts circonscrits dans une typologie à l'épreuve des textes. Mais annoter une relation entre entités nommées attestée dans un texte, ne relève pas uniquement du modèle de connaissances disponible pour les annotateurs. En effet, ceux-ci se trouvent confrontés à la richesse de la matérialité linguistique, en particulier aux expressions de la modalité, aspect dont nous proposons d'aborder les principaux problèmes.

3.3.2 Les écueils de la modalité

C'est en travaillant sur des corpus en français, dans le cadre de nos travaux de recherche, que nous avons pu constater la richesse de l'expression de la modalité. Et c'est bien cette richesse qui participe de la difficulté à établir

les modèles de représentation des connaissances exploités par le TAL, qui soient stables et non ambigus. Mécaniquement, la tâche d'extraction de relations entre entités nommées pâtit de cet état de fait. Avant d'aller plus avant dans l'exposé des problématiques que nous y voyons, il semble pertinent de rappeler ce que nous entendons en parlant de modalité.

Quelques généralités sur la modalité, du point de vue du TAL. La notion de modalité fait l'objet de débats quant à sa définition linguistique, que nous n'abroderons pas en détail ici. Ainsi, nous invitons le lecteur à consulter le récent ouvrage de Laurent Gosselin (2010) pour un panorama détaillé sur ce sujet (Gosselin, 2010).
La modalité est généralement conçue comme concernant la totalité de ce qui est dit par un énoncé, relevant du cadre général de celui-ci. Ainsi, on y reviendra plus bas, la modalité est particulièrement difficile à traiter automatiquement : les systèmes informatiques se heurtent à ce que la modalité n'est possiblement interprétable qu'en contexte, i.e. dans le cadre d'une situation de communication située, complexe à gérer par une procédure automatique informatisée. Traditionnellement, trois paradigmes modaux sont distingués :

- les modalités logiques ou aléthiques – catégorique (objet assertée comme un fait « La Terre est ronde. »), hypothétique (objet présentée comme une possibilité « La Terre est possiblement ronde. »), apodictique (objet présentée comme une nécessité « La Terre est nécessairement ronde. »).
- les modalités épistémiques – qui procèdent des croyances du locuteur(« Je crois que la Terre est ronde. »).
- les modalités déontiques – qui procèdent d'une appréciation morale ou sociale (« Je suis obligé de croire que la Terre est ronde. »).

La modalité est, en outre, liée au temps, à l'aspect et aux modes verbaux, prenant des formes variées quant à sa manifestation concrète dans les textes dont voici quelques façons de l'exprimer :

- Mode indicatif : mode du réel, de la certitude
- Mode subjonctif : mode de l'éventualité, de l'irréel
- Mode conditionnel : mode de l'hypothèse, de la supposition
- Verbes et auxiliaires modaux : « devoir », « pouvoir »
- Adverbes modalisateurs : de renforcement, d'atténuation, d'intensification (resp. « certainement », « peut-être », « toujours »)

– Connecteurs : « malgré », « même si »
– Quantificateurs : « peu », « le / la moindre »

Notons que nous n'abordons la négation que ci-après, étant au fait que cette notion est l'objet, en linguistique, de définitions variées. Nous retiendrons, pour notre part, que la négation a un statut particulier, en tant qu'elle peut, d'une part, être considérée comme une catégorie assertive à part entière, faisant donc partie de ce qui est asserté ; d'autre part, envisagée comme un marqueur de la modalité, reflétant l'attitude du locuteur quant à son énoncé. Nous faisons le choix de la traiter comme relevant du paradigme de la modalité, car les difficultés qu'elle pose pour le traitement automatique sont analogues. En tous cas, la négation s'exprime par différents moyens grammaticaux ; en discours, son instanciation renvoie à des emplois particuliers, renvoyant aussi bien à des usages linguistiques – tels que le niveau de langue – qu'à des effets stylistiques :

– Adverbes, prépositions, conjonctions : « non », « ne ... pas », « sans », « ni »
– Adjectifs, substantifs, pronoms : « aucun », « nul », « carence », « rien », « personne »
– Préfixes : « in- », « dé- »,« a- »
– Propositions à valeur modale : « il est faux de dire que », « ce n'est pas que »
– Registre soutenu : « je ne puis vous dire », « si je ne m'abuse »
– Registre familier (omissions ou constructions elliptiques) : « je sais pas », « pas encore »

Du point de vue du TAL, la question de la négation reste, pour ainsi dire, entière (Horn, 2001; Benveniste, 1966; lan, 2006). D'après notre examen des conférences du domaine, liées à l'évaluation des systèmes d'extraction d'informations et notamment d'entités nommées et des relations qui y sont afférentes, seule la campagne d'évaluation menée dans le cadre de la conférence ACE aborde la question de la modalité ; ce au travers de recommandations d'annotation quant à la temporalité des verbes, où l'introduction de la notion de « doute » est indiquée. Il faut souligner que ces recommandations procèdent d'un choix, plutôt guidé par un choix qui consiste à les élaborer à partir d'exemples tirés des données rencontrées en corpus. Ainsi, plutôt que de se fonder sur une typologie linguistique à proprement parler, l'élaboration des recommandations fournies par la conférence ACE se fonde sur la position des segments textuels porteurs de la modalité dans les énoncés attestés analysés. Notons que cette question de position, peu fréquemment abordée,

soulève des problèmes pratiques pour le repérage de la relation entre entité nommées, considérée comme un objet identifiable automatiquement dans un texte. Il n'y a en effet pas d'ancre ni de repère stable, pour identifier une occurrence de relation. A ce sujet, les recommandations d'annotation de la conférence ACE donnent pour objectif « d'annoter la plus petite ou la plus proche relation », donc de repérer une relation entre entités dans la plus petite fenêtre possible. Autrement dit, seule la portion de texte contenant les entités liées et les segments participant de la sémantique de la relation doivent être pris en compte. Concrètement, comme on le voit dans les exemples présentés plus haut, la fenêtre est susceptible de varier, selon la position des entités associées par une relation donnée.

Dans cette optique, il nous paraît plus à propos de fonder notre réflexion sur la question de la modalité dans l'extraction de relations entre entités nommées, sur un découpage pragmatique, réalisé à partir des exemples attestés en corpus dans nos données de recherche. Ce parti-pris nous semble présenter deux avantages pour notre démarche. Tout d'abord, il est ainsi possible de rendre compte de deux problématiques liées au TAL : le repérage et la catégorisation de la modalité. Ensuite, cela nous permet de mettre en œuvre une procédure implémentable dans une optique d'industrialisation, de l'apprentissage automatisé de grammaires pour l'extraction de relations entre entités nommées, cette grammaire se fondant sur la position des différents composants d'une instance de relation. Cela nous mène donc à envisager la position du segment textuel porteur de la modalité, en fonction de la relation entre entités. Celui-ci peut être interne ou externe à l'occurrence de la relation, comme le montre les exemples ci-dessous :

Modalité interne - temps des verbes
1. George W. Bush et sa femme Laura **accueilleront** la chancelière allemande Angela Merkel.
2. Google **aurait racheté** la société Plannr, qui propose un système d'agenda social pour petits groupes de mobinautes.
3. Après une visite surprise en Afghanistan, le Premier ministre britannique, Gordon Brown, **s'est rendu** au Pakistan pour évoquer avec le président Asif Ali Zardari, une nouvelle stratégie destinée à combattre le terrorisme.

Modalité interne - predicat modal
4. VMware **pourrait** donc ainsi acquérir la partie Suse Linux, et Attachmate Corps, un éditeur de logiciel détenu par Golden Gate Capital.
5. Francisco Partners, **pourrait**, lui, récupérer les autres morceaux, tels Netware et Zenworks.

6. Rice **devrait** rencontrer son homologue iranien Manouechehr Mottaki en marge de la conférence ministérielle internationale sur l'IRA.
7. Le député PS Gérard Bapt **veut** se rendre au Tibet.

Modalité interne - adverbe
8. Selon les spécialistes, Google, Apax Partners et Warburg sont les entreprises qui pourraient **potentiellement** racheter le portail Seznam.
9. Wavre pourrait ensuite recevoir le vainqueur du duel entre Couvin et Fernelmont puis **éventuellement** se rendre à Hoggstraeten une semaine plus tard.
10. Plaxo probablement racheté par Comcast.

Modalité interne - adjectif qualificatif
11. Le géant coréen Samsung (42,3% de parts du marché de la mémoire flash) réfléchit au rachat **éventuel** de Sandisk.
12. Dans une brève déclaration à la presse faite en marge de la clôture de la session d'automne du Conseil de la nation, Belkhadem prédit une rencontre **probable** entre Bouteflika et Mohamed IV.

Modalité interne - négation
13. En préférant NeXT, Apple n'a finalement pas racheté le système Be OS créé sous l'impulsion de Jean-Louis Gassé.
14. Le président Sarkozy n'a donc pas rencontré d'opposants lors de l'entretien avec les représentants de la majorité au pouvoir au Tchad.
15. En raison de son état de santé, le président de l'UFC (opposition) Gilchrist Olympio ne s'est finalement pas rendu à Lomé mercredi.
16. Microsoft n'est pas parvenu à acquérir Yahoo ! malgré plusieurs tentatives depuis janvier 2008.

Modalité externe - antéposition
17. Or **il semblerait** qu'Apple ait décidé d'embaucher Michael R. Sweet, son créateur, ainsi que d'acheter la propriété du code source de CUPS.
18. Luca Di Montezemolo dément toute **éventualité** de rachat d'Alfa Romeo par Volkswagen.
19. L'expertise a été demandée dans le cadre de l'éventuel rachat de 20,4% de parts supplémentaires du capital d'Expand par StudioCanal (Canal +).
Modalité externe - postposition
20. La rencontre Wade-Ouattara **s'est bien déroulée** selon le rapport de nos enquêteurs.
21. La rencontre entre les deux Premiers ministres Poutine et Tusk le 1er septembre **pourrait** ne pas avoir lieu.

Il semble intéressant de noter différentes caractéristiques des cas de figure dans lesquels la modalité est exprimée par un segment textuel inclus dans la fenêtre de la relation entre entités. Nous proposons d'en donner un aperçu ci-après.

- Sur le temps des verbes : l'expression de la modalité est directement indiquée par le temps verbal, qui porte la sémantique de la relation identifiée ;
- Sur le prédicat modal : les verbes et auxiliaires modaux, tels qu'ils sont définis par la grammaire traditionnelle, sont généralement positionnés à proximité du prédicat de la relation. Nous avons observé dans nos données d'analyse, que ces prédicats sont fréquemment conjugués au mode conditionnel, ce que nous attribuons à la nature des données, journalistiques, qui constituent notre corpus de travail : la diffusion d'informations, qui relaient des actions non encore réalisées ou des faits dont la fiabilité demande à être validée pour les asserter, contribue à l'expliquer. Nous en retenons que le domaine dont sont extraites les données influe probablement sur l'expression de la modalité, donc sur la valeur sémantique qu'elle porte.
- En ce qui concerne l'adverbe, dans la plupart des cas rencontrés en corpus, il jouxte le prédicat verbal, qu'il vient modifier. Notons cependant que dans les textes issus de la presse, certaines parties telles que le titre, ne correspondent pas toujours à des phrases complètes : le prédicat modal peut être élidé à des fins stylistiques, comme c'est le cas dans l'exemple 10.
- En ce qui concerne les adjectifs qualificatifs, ils modifient la valeur du prédicat lorsque celui-ci est réalisé par un nom. D'après nos observations, on les retrouve systématiquement en fonction d'épithète du prédicat nominal, lorsqu'ils sont inclus dans la fenêtre de la relation. Ils peuvent aussi renforcer la modalité portée par le verbe, comme le montrent les exemples 11 et 12 – respectivement, « réfléchir » et « prédire ».
- Concernant la négation, qui habituellement est déclarative – donnée à voir dans les exemples 13, 14 et 15 – elle répond à une assertion pour la réfuter ou désapprouver ; en contexte, on observe qu'elle s'accompagne souvent d'autres indices modaux dans la fenêtre de la relation, même si ce n'est pas systématiquement le cas – comme on le voit dans l'exemple 16.

Nous nous proposons de poursuivre le panorama des caractéristiques que

nous avons identifiées en corpus, en nous attardant maintenant sur celles des cas où la modalité est portée par un segment textuel positionné hors de la fenêtre dans laquelle une relation entre entités est repérée. Le segment en question peut être soit antéposé à la relation – donc à gauche de celle-ci, soit postposé – donc à droite de celle-ci.
Concernant l'antéposition, le segment textuel porteur de la modalité peut être matérialisé autour d'une construction verbale et l'instance de la relation prend la forme d'une subordonnée complétive – exemples 17. Dans certains cas, la construction est nominale et la relation prend la forme d'une proposition complément du nom – exemple 18. Enfin, le segment textuel porteur de la modalité peut prendre la forme d'un adjectif épithète du prédicat nominal de la relation – exemple 19.
Concernant la postposition : nous n'en avons observé, en corpus, qu'un seul type, où la relation est exprimée par un prédicat nominal et a la fonction de sujet d'un verbe, lequel est modifié par des segments textuels porteurs de la modalité – exemples 20 et 21.

Ainsi, s'appuyer sur la notion de fenêtre de la relation entre entités nommées, pour observer les éléments porteurs de la modalité du point de vue de leur position par rapport à cette dernière, permet de fournir de premiers éléments de description, à partir des occurrences de relations attestées en corpus. Si ce travail doit, selon nous, être poursuivi par des travaux ultérieurs, il donne déjà un aperçu des cas de figure que l'on peut rencontrer dans le cas d'un corpus journalistique, exploité en contexte industriel.

Problématiques spécifiques posées par la modalité dans les textes.
Les précédentes observations ont mené à identifier des problématiques spécifiques, liées à la présence de la modalité dans les relations entre entités nommées :

- tout d'abord, il semble important de souligner que la richesse des modèles linguistiques de la modalité semble difficile à appliquer de façon systématique, si l'on souhaite par exemple s'y adosser pour mener à bien une tâche d'annotation de la modalité dans les textes, en particulier quant à la question de la catégorisation des relations identifiées ;
- ensuite, s'il est possible d'observer que la présence des segments textuels porteurs de la modalité, liés à une occurrence de relation, peuvent être instanciés dans différentes constructions, nous constatons que cette question, qui semble pourtant revêtir un intérêt de premier plan pour

le traitement automatique, est quasi absente des travaux de recherche relatifs à l'extraction de relations entre entités nommées, que ce soit dans les guides d'annotation ou les systèmes d'extraction existants ;
– étant donné que, pour une occurrence de relation donnée, plusieurs indices de modalité peuvent être présents, comment alors déterminer, de façon systématique, la valeur générale de la modalité de la relation ?

Ce sont différentes problématiques auxquelles nous tenterons d'apporter des éléments de réponse, au travers d'une expérience décrite et menée dans le chapitre suivant, dont l'objet sera d'élaborer un modèle d'annotation de la modalité axé sur la gestion des relations entre entités nommées.

Chapitre 4

La question de la modalité : vers une annotation manuelle robuste

Nous l'avons évoqué précédemment : l'extraction de relations entre entités nommées ne peut ignorer la question de la modalité. Il nous paraît important d'aborder cet aspect, avant de parvenir au cœur de notre démarche sur l'extraction de relations entre entités nommées. L'expérience en contexte industriel montre que selon les besoins applicatifs, la nécessité pour les systèmes de parvenir à repérer ces phénomènes et nuancer les extractions obtenues peut être primordiale. Par exemple dans le domaine économique, les enjeux financiers sous-jacents rendent nécessaire le fait de parvenir à une telle finesse d'analyse : on pense d'emblée aux applications de veille économique, où il est vital de savoir si le rachat d'une entreprise par une autre relève de la probabilité ou de la certitude, cela conditionnant de façon très différente les flux boursiers. La notion d'incertitude d'un évènement est souvent liée à la modalité mais ce phénomène nous semble poser des difficultés d'annotation : le nombre d'étiquettes et de consignes seraient très importants et difficiles pour des annotateurs si ils suivaient un schéma inspiré des études linguistiques de ce phénomène. Nous avons voulu avoir un schéma d'annotation simple permettant de représenter l'incertitude et dont la mise en œuvre reste abordable

4.1 Corpus retenu pour les expériences

Pour notre expérience, notre choix s'est porté sur des corpus de presse. Plusieurs raisons nous ont incité à nous appuyer sur des corpus de cette na-

ture. Tout d'abord, ce sont des données qui connaissent une grande variété quant aux thématiques abordées, ce qui permet d'éprouver la démarche que nous présentons sur différents domaines de la culture générale. Ensuite, la pluralité des auteurs d'articles de presse donne accès à une diversité de productions linguistiques, intéressantes pour tester la démarche proposée dans un contexte d'expression écrite relativement peu contraint – en tous cas par rapport à une stylistique de domaine spécialisée fortement normée, comme peut l'être le discours juridique, par exemple. A cela, il faut ajouter que, dans le milieu industriel, les systèmes automatisés et les linguistes qui interagissent avec eux sont tous deux confrontés à des données extrêmement variables, qui peuvent aussi bien provenir de textes issus de domaines de spécialité que de domaines généralistes. Enfin, de nombreuses conférences dont l'objet est celui de l'analyse des entités nommées, comme les conférences ACE [1], proposent l'utilisation de corpus de presse.

Ainsi, nous avons choisi de réaliser nos expériences sur deux corpus de taille différente. Le premier corpus, d'un volume de plus de 17 millions de mots – que nous nommerons C17 – est constitué par l'ensemble des articles parus dans Le Monde durant l'année 2007 [2]. Le second corpus, de taille plus modeste, comporte plus de 2 millions de mots – nommons-le C2 – et comprend des données issues de flux d'informations variés, auxquelles nous avons eu accès dans le cadre de l'entreprise où nous avons effectué nos travaux de recherche [3].

Code du corpus	Nom du corpus	Nombre de mots
C17	Le Monde 2007	17 621 487
C2	Corpus « news »	2 912 232

Dans l'industrie, lorsque l'on s'intéresse à des données de presse, la source d'information la plus couramment exploitée est Internet. L'on y trouve des sites d'information et les dépêches des agences de presse y sont diffusées en continu. La plupart de ces contenus peuvent être facilement récupérés via

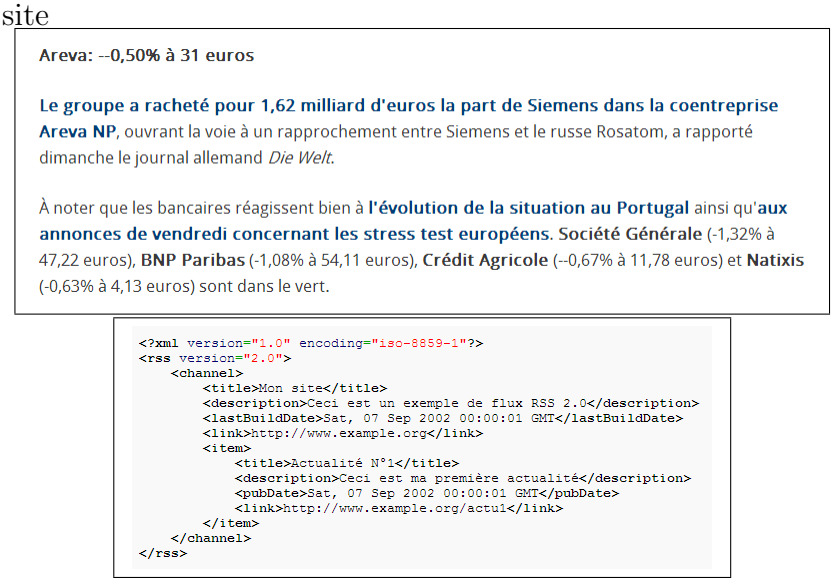
1. cf. chapitre 2 et 3

2. Afin de pouvoir avoir accès à l'intégralité des contenus des articles, nous avons choisi d'acquérir le corpus Le Monde 2007 auprès de la société ELDA, qui dispose des droits négociés avec le journal pour une exploitation intégrale des données publiées à des fins de recherche.

3. Ces données nous ont gracieusement été fournies par l'entreprise ARISEM ; initialement destinées à une exploitation spécifique à l'entreprise, la présentation que nous pouvons en faire est limitée, afin de satisfaire à la discrétion à laquelle nous sommes tenus.

différents standards, donc le plus répandu est le RSS – littéralement Really Simple Syndication – formé de textes et d'information péritextuelles, appelées métadonnées, disponible grâce à un système de flux automatiques. Une autre façon de récupérer ces contenus et de les extraire automatiquement des sites où ils sont publiés.

FIGURE 4.1 – Exemple de flux RSS et exemple d'une information disponible sur un site



L'intérêt de construire notre archive de données d'expérimentation avec ces deux corpus est double : nous pouvons décrire, d'une part, un cas concret de problématique métier rencontré en entreprise, en présentant le corpus C2 ; d'autre part, nous pouvons évaluer le potentiel d'industrialisation des méthodes que nous proposerons, dans la suite de ce travail, avec un corpus de taille importante : le corpus C17.

4.2 Expérience sur corpus autour d'une annotation simple et robuste de l'incertitude

4.2.1 Principe de l'expérience

Un schéma de représentation des connaissances relativement simple peut avoir une robustesse et de finesse suffisantes pour répondre à des besoins

applicatifs, conçus dans une perspective de mise en œuvre industrielle. Nous avons mené une expérimentation préliminaire à notre travail de recherche, avec l'objectif :

- d'apporter des éléments de réflexion aux questions précédemment posées sur la relation entre la modalité et les relations entre entités nommées ;
- de proposer une évaluation des résultats obtenus.

Cette expérience, qui a fait l'objet d'un article (Ezzat et Poibeau, 2011) publié lors de la conférence RANLP 2011 – Recent Advances in Natural Language Processing – ébauche un schéma d'annotation qui prend en compte les phénomènes de la modalité dans les textes. Elle contribue à appuyer l'idée selon laquelle un schéma d'annotation simple peut être appliqué aux textes ; à partir de quoi nous formulons l'hypothèse qu'un tel schéma d'annotation permet d'obtenir un accord inter-annotateur très élevé, dont on présume qu'il est suffisant pour la majorité des applications industrielles.

Plus précisément, nous sommes partis du point de vue suivant : les relations entre entités nommées qui nous intéressent correspondent généralement à un évènement – une rencontre entre deux personnes, un rachat d'une entreprise par une autre – enrichi d'informations liées à cet évènement – sa nature hypothétique, son caractère attendu, ses conséquences probables, ou encore le fait que l'action liée à l'évènement soit ou non réalisée, par exemple. Ce sont ces informations, cet enrichissement, dont procède la modalité, qui nous intéresse particulièrement ici. Par ailleurs, les différentes campagnes d'évaluation menées sur les systèmes d'extraction d'informations ne prennent que rarement en compte la gestion de ces phénomènes. Nous souhaitons prendre les choses du point de vue de la robustesse, défini comme étant la capacité d'un modèle de représentation donné à pouvoir catégoriser des entrées textuelles, de la façon la plus univoque possible. En effet, il apparaît légitime de penser qu'une modélisation linguistique fine de la modalité n'est pas systématiquement appropriée, étant données les applications industrielles. Dans ce sens, le but de notre expérience est de voir s'il est possible de proposer un schéma de représentation des connaissances, orienté par la prise en compte de phénomènes dont on peut dire qu'ils relèvent de la détection d'incertitude, et qui soit à la fois simple et opérationnel pour une application en contexte industriel. Partant, nous choisissons de nous intéresser spécifiquement au degré de réalisation d'un évènement, pour savoir s'il est achevé, en cours ou s'il relève encore du potentiel, du possible. Nous souhaitons également pouvoir indiquer si l'évènement est l'objet d'une négation, ainsi que la variété de discours qui rapporte l'évènement ; en outre, lorsque

ce dernier est indirect, la source qui en fait mention est spécifiée. Enfin, nous décidons de nous concentrer sur l'analyse des relations entre entités nommées, de type « Achat ». Le tableau 4.1 présente le jeu d'étiquettes définies pour mener l'expérience.

TABLE 4.1 – Etiquettes définies pour l'expérimentation

Etiquettes	**Cible**	**Description**
COMPLETED	Evènement	Indique si l'action liée à l'évènement est réalisée et s'est achevée
ONGOING	Evènement	Indique que l'action liée à l'évènement est en cours et n'est pas encore achevée
POSSIBLE	Evènement	Indique que l'action liée à l'évènement est du domaine du potentiel
NEGATED	Evènement	Indique que l'action liée à l'évènement est l'objet d'une négation
DIRECT	Source	Indique si l'action de l'évènement est énoncée en discours direct
INDIRECT	Source	Indique si l'action de l'évènement est énoncée en discours indirect
SOURCE	Source	Spécifie la source si le discours est indirect

Pour illustrer la façon dont ce jeu d'étiquettes peut s'appliquer à des données concrètes, extraites de corpus qui ont été constitués à des fins d'applications sur le terrain, il nous a paru pertinent de le mettre en œuvre sur l'un des corpus disponibles pour notre recherche. En l'occurrence, nous avons choisi le corpus C2, dont nous avons prélevé un échantillon. Arrêtons-nous un instant sur la procédure d'échantillonnage appliquée ici, avant de poursuivre.

4.2.2 Une méthode d'extraction de phrases pertinentes qui exploite les cooccurrences

Dans le travail quotidien auquel nous avons été confronté, la tâche d'identification de phrases pertinentes a été l'un des premiers problèmes que nous avons voulu résoudre. Dans notre cas, une phrase pertinente est une phrase contenant une relation entre des entités nommées. Le repérage de ces phrases en corpus peut très rapidement s'avérer chronophage : il est en effet généralement

nécessaire de lire une grande quantité de textes, pour n'en identifier au final qu'un modeste ensemble exploitable. Dans le but de réduire le temps nécessaire à la constitution de collections de telles phrases, nous avons développé un outil permettant d'accélérer le processus de repérage. Il était pour nous essentiel que ce processus repose sur une méthode simple et facilement reproductible, afin qu'elle s'intègre de façon fluide dans un flux de travail industrialisable, rapide à mettre en œuvre et peu coûteux à maintenir.

Pour y parvenir, nous nous sommes inspiré d'une idée présentée dans les travaux de Riloff (1993) : si des patrons d'extraction contenant des prédicats peuvent être mobilisés pour découvrir de nouvelles entités incluses dans une relation, alors inversement, des patrons établis à partir d'entités nommées peuvent être exploités pour la découverte de nouveaux prédicats de relations. Nous avons adapté cette idée et nous nous sommes appuyé sur le principe du calcul de cooccurrences[4] pour réaliser notre algorithme de filtrage des phrases pertinentes dans un corpus. Concrètement, l'algorithme pour le calcul de cooccurrences est configuré comme suit :

- d'une part, il prend en compte une contrainte imposée sur le type des entités nommées à détecter ; à cette étape, le programme fait appel à un outil de l'état de l'art dédié à la reconnaissance d'entités nommées ;
- d'autre part, les entités doivent se trouver dans une fenêtre précise, c'est-à-dire que le nombre maximal de mots possible inséré entre les deux entités est contrôlé d'après un seuil fixé à l'avance ; idée par ailleurs appliquée dans les travaux de Freitag (1998) ;
- enfin, il prend en compte le nombre d'entités devant être contenues dans la fenêtre définie.

Il est ainsi possible de pouvoir lancer l'outil de filtrage avec ces paramètres minimaux, qui vont avoir une incidence sur la taille de l'échantillon de phrases produit en sortie. On peut par exemple choisir de se concentrer sur un échantillon dont les phrases ne contiennent que deux entités nommées, pour explorer de façon plus fine les prédicats qui instancient une relation entre ces entités.

Avant d'entrer dans le détail de l'expérience elle-même, un dernier mot sur le corpus constitué à partir de la méthode d'extraction présentée plus haut. Celle-ci, appliquée au corpus C2, nous a permis d'extraire plus d'un millier de phrases potentiellement pertinentes, en quelques minutes : la vérification

4. Nous fournissons plus de précision sur la cooccurrence dans le chapitre 8

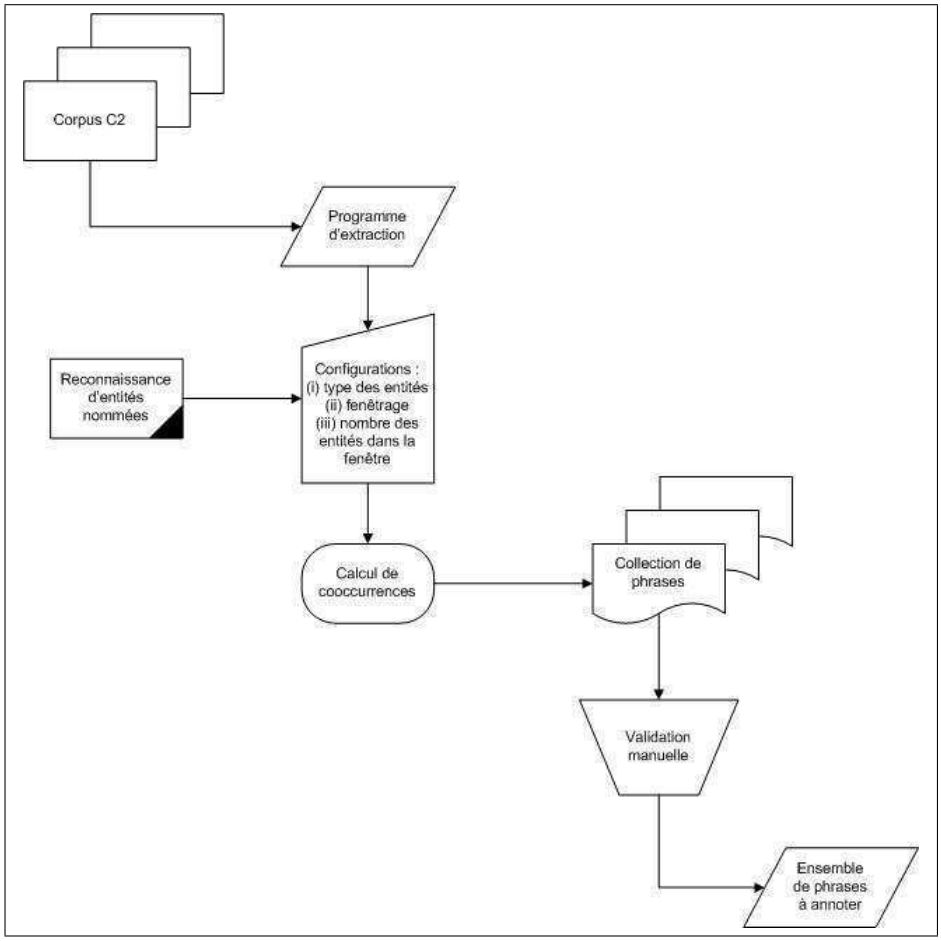


FIGURE 4.2 – Schéma de la procédure d'extraction proposée

manuelle de l'intégralité de l'échantillon a également été rapide – moins d'une heure. On l'a dit, le rapport entre le temps mis à appliquer une procédure et la qualité de son résultat, est un facteur crucial en contexte industriel. Nous avons constaté, après validation manuelle, que près de 40% des phrases extraites étaient effectivement pertinentes, c'est-à-dire porteuses de la relation « Achat » sur laquelle nous avons choisi de nous concentrer initialement. Finalement, cette procédure d'extraction présente plusieurs avantages : rapide à mettre en œuvre, reproductible, facilement intégrable dans une chaîne de traitement, ses résultats sont estimés suffisamment de bonnes qualité en regard d'une exploitation industrielle.

4.2.3 Tâche d'annotation et évaluation de l'accord inter-annotateur

A partir de l'ensemble de phrases pertinentes obtenu, nous avons établi un échantillon destiné à la tâche d'annotation en sélectionnant arbitrairement les 100 premières phrases de l'ensemble. L'échantillon de données pertinentes pour l'annotation constitué, la liste des étiquettes – présentée plus haut – a été fournie à deux annotateurs, appelés A et B, sans guide ou explicitation particulière ; ceux-ci ont eu pour tâche d'annoter chacun l'intégralité de l'échantillon disponible. Il s'est donc agi pour eux d'attribuer les étiquettes adéquates à chacune des phrases de l'échantillon, en se concentrant sur la présence explicite d'un prédicat renvoyant à la relation recherchée – ici « Achat ». Le but de l'opération était, *in fine*, d'établir des conditions d'expérimentation propres à une évaluation de l'accord inter-annotateurs, à l'issue de la tâche en question. Nous présentons dans la figure 4.2 quelques exemples annotés, avant de préciser la procédure d'évaluation appliquée.

L'accord inter-annotateur a, classiquement, été évalué avec le coefficient kappa de Cohen (Cohen, 1960). Il s'est agi de comparer les annotations opérées par l'annotateur A à celles de l'annotateur B ; nous avons paramétré l'évaluation comme suit : si les étiquettes affectées par A et B à une phrase ne présentent pas de correspondance exacte, alors nous considérons que A et B sont en désaccord. Notons que ce choix peut être considéré comme strict, en ce sens qu'il ne permet pas de pondérer le désaccord entre annotateurs. Nous avons néanmoins obtenu un score kappa de 0,94 : selon l'interprétation courante de ce score [65], l'accord inter-annotateur est quasi-parfait ; cela permet de considérer, légitimement, que la méthode proposée est efficace et précise, étant donnés les objectifs initiaux que nous nous sommes donnés. A l'issue de la tâche d'annotation, les annotateurs ont échangé sur les cas problématiques. Notamment, l'affectation des étiquettes DIRECT ou INDIRECT aux phrases présentant une négation été difficile à trancher et compte pour l'essentiel des cas de désaccord inter-annotateurs. Par exemple, dans la phrase : « Twitter dément la rumeur de rachat par Apple. », le désaccord a porté sur le fait de savoir si l'étiquette DIRECT ou INDIRECT devait être utilisée pour indiquer la source de l'information, étant donné que l'énoncé est porteur d'une négation – donc que l'étiquette NEGATED lui est associé. Il leur a cependant été aisé de parvenir à s'accorder, pour définir des règles d'application des étiquettes : dans le type de cas problématique précité, il a été choisi de privilégier l'étiquette INDIRECT.

Il ressort de cette expérimentation préliminaire que tous les cas de désaccord

ont pu être résolus, à l'issue de la tâche, par une discussion entre les annotateurs : aucun désaccord n'est donc resté insoluble. Si ces résultats préliminaires sont encourageants, ils demandent à être confirmés par une expérience à plus grande échelle, adossée à un corpus plus important et mobilisant un plus grand nombre d'annotateurs. De plus, il nous semble indispensable de mettre à l'épreuve la robustesse du schéma d'annotation proposé, notamment en cherchant à l'étendre selon les besoins du terrain, en explorant les spécialisations possibles et en évaluant sa maintenabilité dans le temps, de nouveaux besoins d'applications étant sans aucun doute susceptibles d'émerger en contexte industriel. C'est d'ailleurs l'un des principaux axes de travail que nous souhaiterions pouvoir poursuivre à l'avenir. Nous retenons en tout cas plusieurs enseignements, au terme de ce chapitre :

- un schéma d'annotation simple peut être mobilisé pour caractériser la modalité dans les phrases présentant des relations entre entités nommées ;
- la simplicité du schéma d'annotation participe de la rapidité de sa mise en œuvre, qui, étant donnés les résultats encourageants obtenus sur l'accord inter-annotateur, peut présenter un intérêt réel pour les applications opérationnelles sur corpus en contexte industriel.

TABLE 4.2 – Exemples de phrases du corpus annotées.

Twitter dément la rumeur de rachat par Apple	NEGATED, INDIRECT, SOURCE='rumeur'
Areva a racheté pour 1,62 milliard d'euros la part de Siemens dans la co-entreprise Areva NP, ouvrant la voie à un rapprochement entre Siemens et le russe Rosatom, selon le journal allemand Die Welt, qui cite les porte-parole des deux groupes, s'exprimant dans un document qui sera publié lundi.	COMPLETED, INDIRECT, SOURCE='les porte-parole des deux groupes'
Selon Apple4us, un des plus gros blogs chinois au sujet d'Apple, la firme de Cupertino aurait racheté EditGrid, un service de tableurs en ligne basé à Hong Kong, pour une somme comprise entre 10 et 30 millions de dollars.	COMPLETED, INDIRECT, SOURCE='Apple4us'
Amazon aurait racheté la jeune pousse américaine Touchco basée à New York pour développer son offre de lecteurs de livres numériques Kindle.	POSSIBLE, DIRECT
Le possible rachat du Parisien-Aujourd'hui en France par le groupe Dassault inquiète.	POSSIBLE, DIRECT
La société Acom27 dirigée par Monsieur et Madame Garnot n'a absolument pas été rachetée par les éts Cochet.	NEGATED, DIRECT

Deuxième partie

Elaboration d'un système d'extraction de relations entre entités nommées dans un contexte industriel

Chapitre 5

Panorama des applications et des méthodes pour l'extraction de relations entre entités nommées

Les dernières décennies ont été le théâtre d'une profonde mutation dans la façon de produire, échanger, stocker et archiver les informations de toutes sortes, contenues dans des documents de nature variée. Du fait de leur quantité pléthorique, gérer ces informations est devenu un enjeu majeur, dont découle la nécessité de mettre en place des procédures pour les exploiter rapidement et efficacement. Ces procédures recourent essentiellement à des outils informatiques, dont l'objectif général est d'identifier, dans la masse, un sous-ensemble d'éléments pertinents, étant donné un besoin spécifique d'accès à l'information. Par exemple, pour une entreprise qui commercialise des services destinés au grand public, si traiter des informations fournies par ses clients est indispensable, le faire dans un délai minimal peut s'avérer vital, quand bien même ces informations ne se présentent pas toujours dans un format homogène et structuré. C'est dans ce cas de figure qu'interviennent les applications informatiques relevant du paradigme de l'Extraction d'Information (EI). Nées à la fin des années soixante-dix, ces applications ont pour objectif général l'extraction d'informations à partir de textes, pour produire des sorties structurées. Ce champ étant vaste et complexe, des domaines d'applications plus spécifiques ont fait l'objectif de recherche, en particulier l'extraction de relations entre Entités Nommées. Notre plan s'articule autour de trois parties : les applications de l'EI mobilisant l'extraction de relations entre entités Nommées, le type et le format des données analysées ainsi que les structures de données extraites. Enfin, nous tenterons d'esquisser

les problèmes spécifiques que peut poser l'intégration des techniques actuelles dans une chaîne de traitement industrielle, pour aborder le chapitre suivant où nous examinerons plus spécifiquement ce sujet.

5.1 Les grands types d'applications ayant recours à l'extraction d'information

Selon Poibeau (2003), l'extraction d'information a pour objectif général d'analyser automatiquement des textes et en extraire un ensemble d'informations pour remplir une base de donnée. Dans ce champ relativement vaste, nous nous intéressons aux relations entre entités nommées. Ce type de traitement est aujourd'hui utilisé par un vaste ensemble d'applications courantes, qu'elles soient implémentées dans les systèmes d'information des entreprises ou relèvent des applications destinées au grand public. Nous en donnons, dans cette section, un panorama non exhaustif.

5.1.1 Le traitement des interactions avec des clients ou des usagers

Les grandes entreprises – les banques ou les compagnies de téléphonie, par exemple – comme les grandes institutions de service public – les universités, les services administratifs, notamment – reçoivent des quantités importantes d'informations émises, pour les premières, par leurs clients, pour les secondes, par leurs usagers. L'un des défis que rencontrent ces organisations est de pouvoir convertir ces données dans un format qui soit facilement exploitable pour qu'il soit possible de les traiter informatiquement. Pensons par exemple à la gestion de données issues de formulaires de renseignement, présentant l'identité et les informations de contact de chaque individu ; gérer ces informations peut présenter plusieurs difficultés comme le temps mis pour les traiter, les mettre à jour, en étant sûr de n'avoir que ce qu'il faut comme information utile. L'un des enjeux est de pouvoir passer d'éléments d'information, initialement disponibles sous la forme de chaînes de caractères, à des données structurées qu'il est possible d'enregistrer dans des bases informatiques, afin justement de les traiter plus rapidement et d'éviter les doublons : cela passe par la structuration homogène des informations (Borkar et al., 2001; Sarawagi et Bharnidipaty, 2002). Les entreprises, soumises à des problématiques de performance sur leurs produits et services, ont vu l'utilité de recourir aux solutions d'extraction d'information automatisée pour gérer des cas de figure tels que l'identification des produits visés par les requêtes soumises par des

clients (Chakaravarthy et al., 2006), ou encore l'identification de la satisfaction de ces derniers à partir des transcriptions de conversations téléphoniques (Jansche et Abney, 2002).

5.1.2 Le traitement des informations biologiques et médicales

Dans le domaine de la santé, en particulier de la bio-informatique, l'extraction d'information concerne des entités telles que les noms de protéines ou de gènes, par exemple. Ces données permettent notamment de peupler des bases de connaissances [1], utilisées en génomique. Il faut souligner que la nature spécifique de ces données, d'un type différent de celles sur lesquelles l'intérêt des conférences traitant de l'extraction d'information – MUC en particulier – portait habituellement, a motivé l'amélioration des systèmes dédiés à cette tâche (Bunescu et al., 2005; Plake et al., 2006).

5.1.3 Les applications liées à l'explosion des données disponibles sur Internet

Les applications sur le web sont probablement les plus nombreuses et c'est sur ce domaine sur lequel la plus grande partie des travaux se concentre. Cependant, il est possible de les diviser en plusieurs catégories car elles demeurent assez hétérogènes et diffèrent grandement d'une étude à l'autre.

Comparateur de prix

Depuis maintenant quelques années, les comparateurs de prix sur internet prennent une place de plus en plus importante. Pour s'en apercevoir, il suffit de taper une requête portant sur un produit de consommation dans un moteur de recherche web, et de constater les résultats en première page. Sponsorisés ou non, il s'agit souvent de sites recensant différentes enseignes et comparant leurs prix. Beaucoup d'intérêt a donc été porté à la création de ce genre de *métamoteur* qui parcourt automatiquement les sites de plusieurs enseignes pour extraire les produits et leurs prix associés à des fin de comparaisons depuis la fin des années 90 (Etzioni et al., 1997). Aujourd'hui, la tâche de ces systèmes est devenue plus difficile car les sites utilisent des langages de scripts qui interrogent dynamiquement des base de données, ce

1. dont les plus connues sont Pubmed et Mesh

qui oblige les systèmes à extraire les informations à partir des formulaires d'affichage (Huang et al., 2008).

Base de citations Plusieurs bases de données de citations sont créees à partir de mécanismes d'extraction d'informations sur des sources allant des sites internet des conférences scientifiques aux pages personnels des auteurs de publications. Les plus célèbres sont Google Scholar [2], Cora (McCallum et al., 2000b) et Citeseer (Lawrence et al., 1999). L'élaboration de telles bases de données nécessite d'extraire des informations à différents niveaux, en commençant par repérer les pages contenant les listes de publications, puis d'extraire les titres, auteurs et références bibliographiques pour chaque publication. Enfin, pour chaque référence, il faut être en mesure d'extraire les auteurs, le titre, l'année de publication et le moyen de publication. La structure ainsi extraite apparaît comme un graphe de publications, dont les relations sont représentées par les citations. Il est alors possible de calculer l'importance et l'influence d'un auteur en se basant sur le nombre de fois où il est cité par ses pairs.

Analyse d'opinions

Internet regorge de sites web rapportant des opinions non modérées sur des sujets variés tels que des livres, des films, de la musique ou des personnes. La plupart de ces opinions sont émises à travers des textes non structurés comme les blogs ou les forum de discussion. Structurer ces données de sorte à savoir ce que les gens pensent d'un sujet donné représente une forte valeur ajoutée pour la plupart des entreprises. Par exemple, pour un produit donnée comme une voiture, il est intéressant de connaître l'opinion émise par les internautes sur différents aspects (les équipements, la conduite, l'esthétique, etc.).

Dépasser la recherche par mots-clés

L'un des plus gros enjeu aujourd'hui pour les systèmes de recherche d'informations est de permettre l'utilisation de requêtes structurées. La recherche

2. www.scholar.google.com

par mots-clés est utilisée pour récupérer des documents pertinents. Typiquement, les requêtes sont composées de noms, de groupe nominaux ou d'entités nommées. Elles échouent lorsqu'il faut rechercher des relations particulières entre plusieurs entités (Chakrabarti, 2002). Par exemple, si on recherche des documents contenant des textes à propos du « rachat de la société Youtube par Google », l'emploi de requêtes par mots-clés semble assez limité. L'utilisateur devra étendre sa requète en ajoutant des mots comme « rachat », utilisant ses connaissances basiques du moteur qu'il utilise afin d'affiner les résultats renvoyés. Les premiers prototypes permettant l'utilisation de requête structurées sont aujourd'hui assez rares et commencent seulement à apparaître (Suchanek et al., 2007; Cafarella et al., 2007; Suchanek et al., 2006).

5.2 Type de données analysées

Les données analysées en entrée d'un système d'extraction peuvent être décrites selon deux dimensions, hétérogénéité des formats et des genres textuels. Ces deux axes ont une grande influence sur les systèmes, qui doivent s'articuler autour de ces dimensions et prendre des décisions en amont. La visée applicative, les approches, les prétraitements des données, tous ces composants varient en fonction de la composition des données en entrée des systèmes.

5.2.1 Formats hétérogènes

Si internet a permis la prolifération et l'accessibilité d'une grande masse de documents, le processus de numérisation du contenu textuel a entraîné parallèlement un éclatement des standards, avec une différence qui s'effectue principalement au niveau du degré de structuration du document.

Les documents les plus structurés en entrée sont issus de systèmes de génération automatique. Typiquement, il s'agit de documents HTML dynamiquement générés à partir d'une base de données (Kushmerick et al., 1997; Liu et al., 2000; Muslea et al., 1999). La structure résultante de la génération des documents varie d'un site à l'autre. La principale difficulté pour les systèmes réside en la compréhension automatique ou semi-automatique de la structure des pages générées, en se basant majoritairement sur les régularités détectées de l'utilisation des balises HTML.

La majorité des travaux en extraction d'informations se focalisent sur des

documents partiellement structurés et spécifiques à un domaine. Il s'agit d'articles de journaux (Consortium, 1998-2008, 2004; Grishman et Sundheim, 1996) ou des citations (Peng et McCallum, 2004; Borkar et al., 2001). Certains documents sont déjà structurés avec par exemple le corps d'un article différencié de son titre. D'autres métadonnées peuvent être également ajoutées, comme l'auteur ou la date de parution de l'article. Avec suffisamment de données annotées, il est possible de développer des modèles d'extraction performants, malgré des variations plus importantes au niveau des textes.

Plus récemment, un intérêt important s'est développé pour des systèmes capables de traiter des textes provenant de domaines et de sources variées. Les systèmes doivent alors extraire des entités et leurs relations sur internet, où il reste difficile de prédire une quelconque homogénéité des données analysées (Banko et al., 2007; Cafarella et al.; Shinyama et Sekine, 2006). L'approche utilisée consiste à exploiter la redondance des informations extraites à travers différentes sources. Ceci est particulièrement vrai pour les relations entre entités.

5.2.2 Genres textuels hétérogènes

Le genre textuel des documents analysés influence l'approche et la vision du problème de l'extraction de l'information. Nous donnons ici l'exemple de deux types de documents non-structurés dont les genres respectifs changent de manière importante les techniques employées par les systèmes afin de les analyser.
Les adresses et les citations sont ainsi des séquences devant être extraites des textes par un grand nombre d'applications différentes (Agichtein et Ganti, 2004; Borkar et al., 2001; Peng et McCallum, 2004; Soderland, 1999). Ce type de données peut être analysé comme une succession de champs juxtaposés qui sont concaténés selon un ordre pertinent. De ce fait, il devient possible d'aborder l'extraction comme un problème de segmentation, où il s'agit de fixer les frontières entre chaque entités correspondant à un champ. Par exemple, une adresse comme « 175, Boulevard Saint-Marcel 75013 Paris » devra être segmentée et remplir les champs définis de la manière suivante :

Numéro	Type	Nom	Code Postal	Ville
175	Boulevard	Saint-Marcel	75013	Paris

Mais de nombreuses tâches ont pour but d'avoir accès à des contextes plus importants comme des phrases dans un paragraphe ou dans un document entier. C'est le cas par exemple de l'extraction d'évènements dans les journaux (Chinchor, 1998a; Grishman et Sundheim, 1996; Grishman, 1997). La difficulté réside alors dans la capacité des systèmes à établir des techniques pour le filtrage des éléments pertinents contenus dans des documents plus longs.

5.3 Structures de données extraites

Nous distingons trois types généraux de structure en sortie des systèmes. Le premier type correspond à la tâche d'extraction des entités nommées où des classes prédéfinies peuvent être projetées sur un ou plusieurs mots dans les textes. Le second type de structure traite les relations. Il s'agit de grouper des entités nommées entre elles par une représentation tabulaire comme des formulaires. Enfin, il existe un troisième type de structure plus complexe destiné à la représentation des données par une structure arborescente. C'est notamment le cas pour des résultats du type ontologie.

5.3.1 Entités nommées

Les entités nommées les plus étudiées sont généralement les noms de personne, de lieu ou d'organisation. Elles ont été popularisées par des campagnes d'évaluation telles que MUC (Chinchor, 1998a; Grishman et Sundheim, 1996) ou ACE (Consortium, 1998-2008). Aujourd'hui ces classes ont été largement étendues à de nombreux domaines et certains systèmes d'extraction d'entités nommées s'intéressent notamment à des classes diverses commes des noms de protéines ou de gènes. Les résultats d'un système d'extraction d'entités nommées sont généralement représentés par une structure *à plat* : il est ainsi possible de projeter les annotations directement sur les parties de textes concernées. Le langage XML est largement privilégié pour cette description. Il permet entre autre de garder une certaine lisibilité pour le lecteur humain grâce à l'injection de balises explicites jouant à la fois le rôle de bornes et de typage de classe. Exemple :

– *Le candidat* ***<Person>****François Hollande****</Person>*** *entend « faire pression » sur le groupe* ***<Organisation>****Unilever****</Organisation>***

> pour qu'il ne ferme pas son usine près de ***<Location>****Marseille****</Location>****.*

5.3.2 Relations entre entités nommées

Les relations sont des objets où deux entités nommées ou plus sont liées selon une catégorie sémantique précise, comme la relation d'acquisition entre deux entités du type « organisation » ou la relation de déplacement entre une entité de type « personne » et une entité de type « lieu ». Exemple :

- Aujourd'hui, **Microsoft** a confirmé l'acquisition du logiciel **Skype** pour **8,5 milliards** de dollars en cash.
- **Google** s'apprête à effectuer une acquisition significative en rachetant le fabricant de téléphones mobiles **Motorola Mobility**.

Le résultat de l'extraction de relations diffère de celui des entités nommées sur un point bien précis. Les entités nommées représentent des suites de mots dans le texte sur lesquels il est possible de projeter une annotation (généralement le type de sa classe). Les relations sont, quant à elles, représentées par des entités regroupées dans une structure tabulaire. Par exemple, les résultats des deux exemples ci-dessus pourront prendre la forme d'un tableau représenté par la figure 6.1.

Relation de Fusion-Acquisition		
Microsoft	Skype	8.5 milliards
Google	Motorala Mobility	

FIGURE 5.1 – Exemple de résultats d'extraction de relations

Le tableau regroupe les entités selon le concept sémantique de la relation. Il s'agit de « Relation de Fusion-Acquisition ». A ce titre, le tableau joue également le rôle de réificateur où chaque ligne représente un ensemble d'entités nommées mises en relation par le lien sémantique porté par le tableau.

De ce point de vue, la compatiblité de ce genre de structure avec le modèle relationnel est évident. Que cela soit des formulaires ou des tableaux, les résulats d'un composant d'extraction de relations entre entités nommées correspond parfaitement au formalisme des bases de données rélationelles.

A noter que dans l'exemple de la figure 6.1, les couples d'entités nommées ainsi réifiées ne sont pas ordonnés. Une permutation des éléments ne changeraient en aucun cas la valeur du résultat. Cependant, la tâche et les structures de données évoluant, le typage de cellules des structures tabulaires devient de plus en plus important. Par exemple dans les formulaires des conférences MUC sur les actes terroristes, on distinguera des cellules typés par rapport à l'information recherchée, comme le nom du groupe revendiquant un attentat, le nombre de victimes ou l'arme utilisée. Certaines cellules ne sont pas obligatoirement renseignées. Ainsi, certaines informations sont typées comme étant optionnelles à l'existence de l'évènement relaté. L'exemple de la fusion-acquisition ci-dessus l'illustre bien : les deux entreprises participantes à la transactions sont des éléments nécessaires à l'instanciation de cette relation, tandis que le montant ou la date de la transaction apparaissent comme des informations supplémentaires et optionnelles.
Ainsi, il en découle une autre tâche recevant un intérêt grandissant dans la communauté du TAL et agissant sur la structure des resultats : l'annotation des rôles sémantiques (*Semantic Role Labeling*) (Wen-tau Yih et Toutanova, 2006; Fillmore et al., 2003). Etant donné un prédicat et ses arguments comme dans la relation de fusion-acquisition entre deux entreprises, l'objectif est d'identifier les rôles sémantiques des arguments en précisant quelle societé est « acheteur » et quelle societé est « achetée ». Cette tâche est importante dans des applications telles que les systèmes de questions-réponses ou tout système d'extraction d'informations complexe.

5.3.3 Ontologies

Le résultat des systèmes d'extraction d'informations peut aujourd'hui être représenté par des structures complexes qui ne sont pas nécessairement à plat. Il s'agit en général d'arbre ontologique, comportant une hiérarchie au niveau des classes sémantiques. Les technologies développées dans le cadre de l'extraction d'informations peuvent alors être considérées comme un pont entre les textes non-structurés et la représentation formelle des connaissances exprimée dans une ontologie. On parle alors d'extraction d'informations basée sur les ontologies (OBIE, *Ontology-Based Information Extraction*, Li et Bontcheva (2007)). Une partie des systèmes utilise directement les connaissances apportées par l'ontologie cible dans leurs algorithmes d'extraction (Kogut

et Holmes, 2001; Li et Bontcheva, 2007). D'autres n'incorporent pas ces données mais doivent établir une correspondance entre leurs résultats et la modélisation des classes et instances de l'ontologie cible (Handschuh et al., 2002).

Les systèmes OBIE font l'hypothèse la plupart du temps que les relations entre les différents concepts des ontologies sont de nature hiérarchiques (hyperonymie/hyponyme). Cela va jusqu'aux processus d'évaluation, qui prennent en compte ce type de relations dans leurs calculs. Si X et Y sont deux concepts directement liés hiérarchiquement dans l'ontologie, l'erreur consitant en la classification d'une instance de X sous le concept Y devra avoir un coût minimal. Par exemple, on ne considèrera pas totalement faux le fait que « Paris » soit classé sous le concept « Lieu » plutôt que « Ville », car une ville est un type de lieu.

L'autre problèmatique importante relève de la compatibilité entre les « différents mondes » décrits pas des ontologies. Les systèmes incorporant directement l'ontologie utilisateur dans leur algorithme d'analyse ne souffrent pas de cet inconvénient, car ils utilisent directement la hiérachie de classes pour produire leurs résultats. Mais les systèmes qui établissent une correspondance entre leur résultat et l'ontologie cible sont directement confrontés à cette difficulté. Il y a souvent un réel écart entre les classes définies dans leurs résultats et la modélisation de l'ontologie cible. Une des solutions proposés est de créer des règles spécifiques de correspondance entre classe pour chaque ontologie. Mais la difficulté pour un utilisateur d'établir de telles règles augmente le coût global de la mise en place de ces systèmes. Nous aborderons dans le chapître 7 cette problématique à travers une expérience menée dans le cadre de nos travaux.

5.4 Aperçu des méthodes utilisées en extraction d'information

5.4.1 Techniques d'extraction d'information par analyse distributionnelle

L'extraction de relations par étude statistique s'intéresse à la distribution des mots dans un corpus, et propose des couples d'unités linguistiques en relation. Cette approche émet l'hypothèse qu'*il existe un système qui sous-tend le fonctionnement de la langue et que ce système est ramenable à un*

ensemble de règles relativement à un domaine et son sous-langage associé (Seguela, 2001). Nous pouvons observer ce phénomène dans l'exemple qui suit, tiré d'un corpus sur les droits à polluer :

- *Les* ***quotas de gaz à effet de serre*** *fixés par le protocole de Kyoto semblent insuffisants au regard du mouvement écologiste.*
- *Les* ***quotas marchandises*** *sont transformés en véritable droits à polluer par les industries gouvernementales.*

Les approches statistiques par distribution du contexte vont alors repérer que *gaz à effet de serre* et *marchandises* sont tous deux en position de détermination par rapport au mot *quotas*, et supposent alors qu'en contexte, ils sont reliés par un lien sémantique que l'on peut extraire à partir d'une méthode statistique si le phénomène se répète suffisamment par rapport à un seuil fixé. Ceci dit, l'extraction ne précise pas le type de relation. Ici, on peut en déduire qu'il s'agit d'une relation de synonymie *textuelle*, entre *marchandises* et *gaz à effet de serre*.

Une des règles caractérisant cette approche a été formulée par Van Rijsbergen (1979) et avance que *l'emploi de deux termes en cooccurrence est l'expression d'une relation sémantique entre eux*. Des unités linguistiques distribuées d'une même manière auraient des éléments de sens communs.
Cette hypothèse a donné naissance à des travaux et des outils de lexicométrie. Alceste, par exemple, est un logiciel de lexicométrie dont une des fonctions est de faire émerger des *classes* de mots qui auraient un lien sémantique entre eux. Ces classes sont construites en se basant sur les cooccurrences et la distribution des mots qui les composent. Alceste ne propose pas de nommer ces classes et de leur donner une sémantique. Il les fait émerger à partir d'hypothèses statistiques. De la même manière, Smajda (1993)étudie les fréquences de ces cooccurrences afin de proposer des relations entre des unités linguistiques.

D'autres travaux utilisent de véritables patrons morpho-syntaxiques (*Nom-Adj*, *Adj-Nom*, *Nom-Prep-Nom* etc...) et exploitent ces connaissances morpho-syntaxiques par leur distribution statistique, ce qui nécessite donc un prétraitement des textes en amont. Le système SEXTANT (Grefenstette, 1994) en est un bon exemple pour l'anglais et utilise des méthodes statistiques sur des connaissances linguistiques (un arbre syntaxique). Pour le français, le

système ZELLIG (Habert et Nazarenko, 1996) traite les données issues de LEXTER (Bourigault, 1994). Ce dernier décompose les syntagmes en tête et expansion. Puis ZELLIG est chargé de réduire les informations proposées par LEXTER en trouvant des formes dans ces arbres élémentaires, qui mettront en relief les relations entre des mots pleins.
Plus récemment, Fabre et Bourigault (2006) utilisent une méthode couplant indice de proximité distributionnelle et indice de cooccurrence sur des couples nom-verbe.

« *Un couple est extrait si le nom et le verbe apparaissent avec les mêmes arguments sur l'ensemble du corpus, d'une part, et s'ils apparaissent au moins une fois dans un même paragraphe munis du même argument* ».

La méthode est appliquée à un corpus très volumineux (200 millions de mots) et démontre ainsi que l'information n'est pas forcément contenue dans des arguments de relations morphologiquement analogues, et que l'approche statistique relève d'une certaine robustesse sur des corpus volumineux, qui tendent à perdre leur homogénéité au fur et à mesure qu'ils grossissent.

Ces méthodes par approche statistique ne proposent pas d'extraire des relations à proprement parler. Elles n'extraient que des couples de mots, qui ne sont pas organisés en terme de classe de relation. Ils sont essentiellement représentés en un nuage de points pour lesquels on ne caractérise pas les liens qui les relient mais qui relèvent assez souvent de la synonymie. C'est donc une approche émergentiste qui nécessite généralement un processus d'interprétation et de validation. Les couples ainsi extraits peuvent former des classes de mots qui partagent une somme statistique d'environnements significative par rapport à un seuil. L'approche statistique s'avère alors robuste et générique. Mais les processus d'interprétation et de validation en aval comportent néanmoins de nombreuses difficultés. Si les critères statistiques sont connus, la variété de la langue et des types de relations font qu'il est difficile d'adopter une position systématique pour l'expert ou le linguiste, et d'objectiver les résultats validés.

5.4.2 Approche à base de règles

Historiquement, les premiers systèmes d'extraction d'informations ont été construits à partir de règles créées manuellement ou issues d'un apprentissage (Appelt et al., 1993; Riloff, 1993; Cunningham et al., 2002). Après l'échec historique de la traduction automatique, ces systèmes tentent de résoudre des tâches plus modestes, dont les données présentent une diffi-

culté de représentation moindre. C'est le cas par exemple de l'extraction de codes postaux ou de numéro de téléphone. Mais les systèmes à base de règles perdurent encore aujourd'hui et sont appuyés par la linguistique afin de capturer et modéliser des phénomènes plus complexes (Shen et al., 2007; Jayram et al., 2006).

L'approche à base de règles s'attache à rechercher et décrire des formules linguistiques qui attestent d'une relation donnée, et qui se répètent systématiquement dans un corpus. Nous distinguons alors deux grandes méthodologies. La première se base sur l'hypothèse que la langue naturelle relève de principes généraux, récurrents dans tous les textes, fussent-ils techniques ou non. On parle alors de langue transversale à tous types de production langagière. La seconde approche touche aux problématiques de l'apprentissage au sens large et considère que les textes et les phénomènes linguistiques qui y sont attestés ne le sont que dans le cadre du domaine auquel est circonscrit ce texte. Dans les sections 5.4.2, nous décrirons la manière dont ces règles sont représentées et formalisées. Puis nous examinerons dans la section 5.4.2 les degrés de généralisation associés aux symboles utilisés dans les différentes règles. Nous verrons ensuite les techniques de constitution de ces règles, en commençant par leurs élaboration manuelle, pour finir par les techniques d'apprentissage automatique ou semi-automatique.

Représentation

Les systèmes d'extraction d'informations basés sur des règles emploient différents formats. Cela va des expressions régulières classiques (Soderland, 1999) aux requètes formulées dans un langage comme le SQL (Shen et al., 2007). Si la forme des règles peut varier d'un système à un autre, elles gardent néanmoins une structure commune et un pouvoir expressif identique que nous allons décrire par la suite.

Une règle d'extraction est typiquement constituée en deux phases, avec une partie se chargeant de capturer le phénomène et son contexte à extraire, qui vient ensuite activer une seconde phase définissant la projection de l'annotation résultant de cette capture. La règle de capture est en général équivalente au pouvoir expressif d'une expression régulière. Il s'agit de repérer l'entité (au sens général) à extraire en se basant sur son contexte d'apparition. Généralement séquentiel, ces contextes d'apparition peuvent être définis selon différents niveaux d'analyse. Lorsqu'une séquence de texte est repérée par ce biais, la règle active alors sa phase d'annotation en projetant une étiquette sur la séquence trouvée. L'ensemble des règles est appliqué à un

texte. Ces règles relèvent généralement d'une organisation dite « en cascade », où l'ordre des règles est ordonné et les règles appliquées en dernier utilisent des symboles représentant une étiquette qui aura été préalablement projetée par les règles appliquées en premier.

Par exemple, pour les relations entre entités nommées, un premier ensemble de règle sera chargé de détecter et annoter les entités nommées elles-mêmes. Ces symboles seront utilisées dans un second temps par d'autres règles qui vont déterminer si oui ou non les entités nommées ainsi détectées sont en relation. D'une certaine manière, cela renvoit directement au formalisme des graphes conceptuels de Sowa (Sowa, 1984; Gerbé). Des concepts (comme des entités nommées) sont reliés par un arc qui porte le type de la relation dans une disposition qui rappelle la logique des prédicats. Pour les systèmes à base de règles, les concepts sont détectés par les règles appliquées en premier tandis que l'enjeu réel réside dans la modélisation de règles postérieures représentant l'arc de jointure entre deux concepts dans le formalisme de Sowa.

Différents niveaux de description

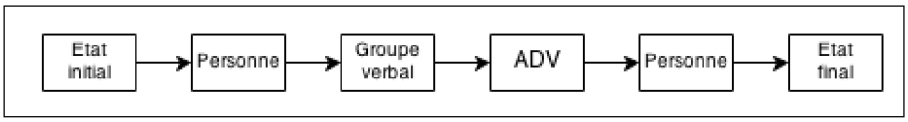
Les systèmes d'extraction d'informations basés sur des règles ne s'appliquent généralement pas sur les textes bruts. La plupart du temps, les règles sont appliquées dans le contexte d'une phrase, et la description des contexte du phénomène à capturer s'effectue au niveau d'un mot ou d'une séquence de mots. Il est donc nécessaire d'effectuer une première phase d'analyse qui segmente le texte en phrases, puis en mots. Vient ensuite une seconde phase d'analyse augmentant le niveau de description des mots sur divers axes (sémantique, syntaxique, typologique, etc.) et ayant des portées d'ordre différent (un seul et unique mot ou toute une séquence).
Un système d'extraction d'informations à base de règles applique donc différentes analyses avant l'application des règles. Les mots sont alors associés *a priori* à plusieurs étiquettes symbolisant des niveaux de description divers tels que :

- la forme de surface du mot telle qu'elle apparaît dans le texte
- la nature grammaticale
- le lemme ou la forme du dictionnaire
- le genre et le nombre (selon la langue)
- la forme typographique (mot commençant par une majuscule)
- des traits sémantiques issus généralement de dictionnaires électroniques
- toute autre annotation issue d'une analyse précédente (groupe nominaux, entités nommées)

Les symboles associés à ces différents niveaux d'analyse peuvent alors être utilisés au sein des règles d'extraction. Par exemple, soit la phrase « *Sarkozy a rencontré hier Chirac* ». Un système d'analyse linguistique pourrait associer aux mots de la phrase les étiquettes suivantes :

- Sarkozy : `Personne`
- Chirac : `Personne`
- a rencontré : `Groupe Verbal`
- hier : `ADV`

La règle qui reconnaîtra cette séquence peut être alore représentrée par un automate de la forme suivante :



Cela offre un degré de généralisation de la règle supérieure et elle pourra reconnaître également des phrases comme « *François Hollande a rencontré aujourd'hui Martine Aubry* ».

Règles définies à la main

Nous distinguons deux courants selon que le corpus sert à la validation ou comme un dépôt à partir duquel les connaissances sont acquises.

Dans le cas du corpus servant à la validation, il s'agit de se baser sur les connaissances générales de la langue afin de déterminer les formules caractéristiques d'une relation lexicale. La qualité de ce travail est donc largement dépendante de la compétence du linguiste, qui scrute ces connaissances par introspection afin d'en produire des formules en amont. Le corpus agit ici en guise de validateur, par attestation. Il n'est en aucun cas un outil d'acquisition propre. Si le linguiste cherche à être précis au niveau de sa description et de la production de formules linguistiques, il ne cherche pas la performance à outrance en terme de rappel. Son objectif, en tant que linguiste, est une modélisation du système de la langue générale par des formules. Partant de ces choix, la sémantique des relations lexicales se restreint à quelques relations dites linguistiques. Il s'agit généralement des relations d'hyponymie/hyperonymie, d'ingrédience, ou de synonymie. Nous allons illustrer ce

courant par les travaux de Jackiewicz (1996).
Jackiewicz a établi une liste de marqueurs pour la relation partie-tout, basée sur les copules *être* et *avoir* ainsi que sur les constructions génitives. Du fait de l'ambiguïté des marqueurs, Jackiewicz passe par le concept de classe de mots. Il s'agit de regrouper les mots qui ont des traits sémantiques communs. Par exemple, en regroupant des verbes comme *réunir, rassembler, collecter* sous la dénomination de *verbe de compositionnalité*, Jackiewicz arrive ainsi à lever certaines ambiguïtés lors de la description des patrons. Ces classes de mots élaborées par introspection peuvent difficilement être exhaustives et cette approche, si elle désambiguïse partiellement, a l'inconvénient de diminuer la capacité opératoire des patrons et leur rappel. Ce sont en effet des contraintes importantes qui privilégient la précision au rappel. Nous pouvons signaler d'autres travaux de manière non-exhaustive proposant des listes de marqueurs analogues sur la relation d'hyponymie (Borillo, 1996) ou sur les énoncés définitoires en terminologie (Pearson, 1999).

Une seconde approche vise à utiliser le corpus non plus comme simple validateur de formules linguistiques, mais comme un véritable « dépôt de connaissances ». Le linguiste, au lieu de scruter ses connaissances par introspection, va systématiquement se servir du corpus afin de produire les patrons d'une relation donnée. Le corpus devient le plus large et hétérogène possible, constituant alors un échantillon représentatif de la langue dite générale. Il s'agit donc de méthodes complètement descriptives, et cela implique que les connaissances de la langue générale peuvent être extraites à partir d'un corpus large et représentatif.
Le principe méthodologique de cette approche a donné naissance à des travaux et des outils durant les années 1990. Après s'être donné une relation à identifier, une liste de phrases du corpus dans lesquelles apparaît cette relation est établie. A partir de cette liste, les concepts sont hiérarchisés du plus au moins générique. Puis, les formules linguistiques ainsi que les patrons sont définis.
Le système SEEK (Jouis, 1993) par exemple, est dédié à la recherche de relations dites statiques (identification, appartenance, inclusion, possession, localisation...). Il en localise l'expression à l'aide d'une liste importante de patrons (environ 3300) et de règles morphologiques préétablies, se basant alors sur les grammaires applicatives et cognitives. De son coté, le système COATIS (Garcia, 1998), basé sur la même approche que SEEK, vise à repérer l'expression des relations de causalité. Cette relation est divisée hiérarchiquement en relations spécifiques de la causalité. Des classes de verbes faisant référence à la causalité comme *provoquer* sont ainsi créées. Mais des verbes plus ciblés comme *modifier* sont également relevés. Leur valeur de causalité est alors

confirmée par la présence d'indicateurs dans le contexte. Puis, ces formules sont hiérarchisés selon leur précision, suivant une approche heuristique. COATIS repère également le rôle des acteurs de la relation. Ainsi, il arrive à identifier les termes de la cause, et ceux de l'effet.

Les types de relations choisis se veulent transversaux à de nombreux domaines. Le corpus prend alors une place centrale. Il est le reflet des manifestations linguistiques des principes généraux de la langue naturelle. Un nouveau corpus implique souvent de nouvelles manifestations. Il se doit donc d'être le plus large et le plus hétérogène possible. Les résultats obtenus sont donc directement tributaires du corpus d'étude. Les limites de cette approche sont apparentes si les patrons qui en découlent sont appliqués sur des corpus très techniques. Après de nombreux affinages successifs des patrons et des formules linguistiques, les règles ainsi définies sont certes très précises, mais aussi très contraintes. Il en résulte une diminution de la capacité opératoire de ces patrons sur des textes provenant d'autre domaines. Sur les corpus techniques, de telles règles sont difficilement réutilisables. Certaines sont ambiguës et les corpus techniques peuvent être régis par des relations très spécifiques. De plus, contrairement aux marqueurs définis en amont par un linguiste, les contre-exemples des patrons trouvés dans des corpus viennent directement contredire la règle qu'ils supposent. Nous avons affaire en réalité à des systèmes hermétiques et fermés, difficilement réutilisables sur d'autres corpus techniques.

Algorithmes d'apprentissage de règles

La notion de corpus technique peut difficilement être écartée pour deux raisons :

- Comme nous l'avons vu précédemment, les marqueurs définis à partir de l'hypothèse d'une langue générale ne sont pas opératoires sur des corpus techniques.

- Un domaine spécifique peut considérer des relations spécifiques, qui ne sont pas nécessairement transversales à d'autres domaines. Et les relations dites « linguistiques » telles que l'hyponymie, hyperonymie, méronymie etc... ne sont pas forcément pertinentes sur de tels corpus.

Il est alors nécessaire d'avoir des systèmes avec une architecture souple pouvant s'adapter à plusieurs domaines spécifiques. L'acquisition de nouveaux

marqueurs à des fins d'enrichissement apparaît alors vital et l'apprentissage de règles à partir de corpus peut y répondre, du moins partiellement.
Si les règles peuvent être créées manuellement, dans beaucoup de cas, elles peuvent également provenir d'un apprentissage automatique à partir d'un corpus préalablement annoté. L'objectif est de trouver un nombre minimum de règles qui couvre le maximum des exemples annotés dans le corpus d'entraînement. Pour ce faire, les différents algorithmes apprennent les règles une à une en suivant un schéma commun :

```
Soit D l'ensemble des documents d'un corpus.
Soit R l'ensemble des règles (initialement vide).
Tant qu'il existe des exemples  E annotés dans D non-couverts par R
    Créer une règle qui reconnait E
    Ajouter la règle à R
```

La principale difficulté réside dans la création de règles avec une précision et un rappel élevés, sans être redondantes avec des règles déjà existantes. La plupart des stratégies correspond à l'une deux catégories suivantes : généralisation ascendante (*bottom-up*) (Califf et Mooney, 1999, 2003; Soderland et al., 1995) et spécialisation descendante (*top-down*) (Soderland, 1999). La généralisation ascendante produit des règles et généralise les contextes d'apparition par la fusion de règles initialement proches, tandis qu'à l'inverse, la spécialisation descendante affine de manière itérative des règles très générales.
Les systèmes utilisant ces algorithmes comportent également deux stratégies bien distinctes après la création de nouvelles règles. La première est de réduire les nombre de règles (compression) en éliminant celles qui sont à la base de la création des nouvelles. Cela évite la redondance de règles qui recouvrent les mêmes instances. La seconde politique est de supprimer les exemples positifs lorsqu'une nouvelle règle les reconnaît. La suppression des instances positives est plus efficiente car il n'est plus nécessaire de trouver des règles pour des instances déjà couvertes. Cependant, elle a l'inconvénient par rapport à la compression de ne pas choisir de règles qui pourraient couvrir les exemples restant et qui subsument les règles déjà existantes.
Dans les paragraphes suivant, nous allons décrire les principales caractéristiques des approches ascendantes et descendantes à travers des systèmes existants.

Généralisation ascendante, Bottom up Dans l'approche de généralisation ascendante, les règles de départ sont typiquement issues d'instances positives d'un corpus. Ces règles ont un rappel peu élevé mais une précision de

100%. L'algorithme va ensuite graduellement généraliser ces règles en augmentant le rappel au détriment de la précision. Nous prenons l'exemple du sytème RAPIER (Califf et Mooney, 2003). Il y a trois étapes essentielles dans ce système : la création de règles initiales qui alimentant l'algorithme, la généralisation des règles à proprement parler et la suppression des instances couvertes par les nouvelles règles afin d'éviter la redondance.
La création de règles de départ consiste en une représentation simple d'une instance positive, constituée par les mots qui composent cette instance. Par exemple, la séquence « Selon monsieur Hollande » génèrera une règle du type :

– « selon » - « monsieur » - type-Personne :Hollande

L'étiquette « Personne » sera adjointe à l'entité « Hollande ». Dans RAPIER, chaque règle de départ est constituée à partir de deux instances afin d'augmenter leur couverture initiale. La généralisation de cette règle intervient par la suite en remplaçant un des symboles de la règle précédente par une caractéristique plus générale recouvrant la même séquence. Par exemple :

– « selon » - « monsieur » - type-Mot-débutant-par-une-majuscule = type-Personne
– « selon » - « monsieur » - type-Mot-appartenant-à-un-dictionnaire-de-nom-propre = type-Personne

La première règle remplace le symbole de la forme de surface « Hollande » par un symbole généralisant qui regroupe tous les mots commençant par une majuscule. La seconde la remplace par un mot appartenant à un dictionnaire de nom propre. De manière évidente, plus le nombre de mots est important, plus le nombre de généralisations possibles augmente. Un paramètre fixé par l'utilisateur limite en général le nombre de généralisations maximum pour une règle initiale donnée.

Spécialisation descendante, Top-down La création de règles par la spécialisation descendante débute à partir d'une règle avec un rappel de 100% mais une très faible précision. Par exemple, dans le système Whisk (Soderland, 1999), la règle initiale donnée à l'algorithme se compose de la forme :

```
*(*)*(*)*(*)*
```

Cette règle représentée par une expression régulière capture à la base n'importe quelle séquence de mots avec une subdivision en trois groupes. Elle extraira de nombreuses instances incorrectes et devra être spécialisée de manière itérative.

- Soit R_n la règle courante à une itération donnée telle que R_0 soit la règle initiale
- Pour chaque règle R_n
  - Pour chaque position de symbole w de R_n
  - Générer des règles dont le symbole à la position w est spécialisé
  - Ecarter les règles dont la couverture est en dessous d'un seuil s

Pour ce genre d'approche, les systèmes interagissent souvent avec un être humain pour de meilleurs résultats. Celui-ci agit sur certains paramètres de seuil ou sur l'apport d'exemples d'instances positives. Le nombre d'itérations élevé freine grandement ce genre d'interaction. Les systèmes se doivent de refléter à l'utilisateur en temps réel les modifications que celui-ci apporte entre deux itérations de l'algorithme.

Méthodes à base de règles : conclusion

Nous avons présenté dans cette section les méthodes à base de règles. Le principal avantage de ces méthodes réside dans leur représentation qui demeure accessible à l'être humain. Il est capable d'interpréter et d'augmenter ces règles, rendant les systèmes basés sur ce type d'approche accessible à un expert du domaine. En pratique, ce n'est pas toujours le cas. Ces systèmes requièrent souvent de fixer des seuils agissant sur les algorithmes de génération. Or, cela nécessiste une certaine connaissance de l'algorithme en question. De plus, les validations successives, la représentation des règles et leur nombre viennent accroître la difficulté de l'intervention humaine avant, pendant et après l'apprentissage.

5.4.3 Méthodes à base d'apprentissage statistique

Les algorithmes basés sur une approche statistique ont été éprouvés au départ dans des domaines divers autre que celui de l'extraction d'information ou du traitement automatique des langues. Du domaine de la finance au traitement du signal, ils interviennent essentiellement lorsque les phénomènes étudiés sont complexes et avec une échelle de donnée importante. En réalité, ces algorithmes permettent, par des hypothèses statistiques, de modéliser automatiquement un ensemble de phénomènes représentés par des données

dans un domaine particulier. Il devient alors évident que leur apport peut être important pour la modélisation de la langue : avec un nombre de données grandissantes grâce à Internet, ces algorithmes peuvent être alimentés de manière suffisante de sorte que l'hypothèse statistique vient améliorer une modélisation apprise automatiquement.

Les systèmes basés sur de tels algorithmes ont connu un essor important durant les années 80, sous l'impulsion des conférences MUC. En extraction d'information, il s'agit de répondre premièrement à un problème de décomposition d'un texte non-structuré en une unité significative, puis d'étiqueter des séquences composées par ces unités avec une étiquette significative pour l'application. Nous allons donc voir en premier lieu comment les textes sont structurés en s'attachant particulièrement aux modélisations basées sur le mot, qui reste aujourd'hui la plus employée. Puis nous verrons deux algorithmes de catégorisation dont les résultats sont aujourd'hui les plus probants : les machines à vecteurs de support (*Support Vector Machine*, SVM) et les champs conditionnels aléatoires (*Conditionnal Random Field*, CRF).

Le mot, l'unité de décomposition des textes

La plus commune manière de décomposer un texte en unité atomique est de segmenter celui-ci au niveau du mot. Dans les langues latines, cela se traduit par une notion de mots graphiques délimités par des séparateurs tels que l'espace, la virgule, l'apostrophe, le point, etc. Chaque mot peut être augmenté par des méta-informations variées. Ces informations sont généralement d'ordre linguistique, comme la nature grammaticale du mot ou un trait sémantique issu d'un dictionnaire.
Pour accoler une étiquette à une séquence de mots, il est coutume de la décomposer en plusieurs étiquettes différentes marquant le début, la fin et le milieu de l'entité. Les systèmes de labélisation encodés de cette manières sont populaires et portent l'acronyme BCEO (Begin, Continue, End, Other) comme le montre la figure 5.4.3

La représentation la plus fréquente de la décomposition des textes en mots restent aujourd'hui le modèle vectoriel sur lequel se base les algorithmes d'apprentissage. Il s'agit de décrire un espace dans lequel chaque dimension correspond à un mot du vocabulaire. Il devient alors possible de représenter chaque instances en un point de cet espace. Prenons l'exemple d'un espace composé du vocabulaire suivant : Google, rachète, Groupon, Youtube. La dimension de cet espace est de 4. Soit les deux instances *Google rachète*

Et	**notre**	**correspondant**	**spécial**	**Nicolas**	**Banon**
ConjC	ADJ :POSS	N	ADJ	UKN	UKN
MAJ	MIN	MIN	MIN	MAJ	MAJ
-	-	-	-	PERS :BEGIN	PERS :END

FIGURE 5.2 – Décomposition en mot, avec augmentation d'étiquettes

Groupon et *Google rachète Youtube*. Après une première phase d'analyse venant segmenter ces instances en mot, nous pouvons modéliser leurs vecteurs correspondant. Par exemple, nous mettons la valeur 1 lorsque le mot est présent dans la phrase et 0 sinon.

Google	rachète	Groupon	Youtube
1	1	1	0
1	1	0	1

Notons tout d'abord, que le passage à ce genre de représentation s'effectue au prix d'un sacrifice important souvent critiqué : l'ordre des mots. En effet, dans un espace vectoriel, l'ordre des vecteurs n'a aucune signification et une permutation n'a aucun impact sur les algorithmes (une représentation tabulaire comme dans l'exemple ci-dessus induit en erreur à cause de l'ordre gauche-droite des valeurs affichées). Ce sacrifice a été longtemps pointé du doigt dans le domaine du traitement automatique des langues à travers le terme de sac de mots (*bag of words*). Les mots seraient ainsi mélangés et ne refléteraient en aucun cas les spécificités de la langue naturelle. Or l'ordre des mots dans un texte est bien entendu prépondérant à sa compréhension.
Il existe également un écart important en terme de performance entre les exemples pédagogiques et le passage à l'échelle. Comme nous l'avions mentionné précédemment, les mots peuvent être également augmentés par des méta-informations. En prenant en compte ces données supplémentaires, ajoutés au vocabulaire assez conséquent dans un corpus moyen, les dimensions de l'espace vectoriel peuvent poser des problèmes insolubles du point de vue du temps de calcul. Il n'est pas rare d'avoir affaire à des espace à plusieurs milliers de dimensions. Un apprentissage par un algorithme statistique basé sur des vecteurs peut alors se compter en jours.
D'autres paramètres peuvent également agir sur la modélisation dans un espace vectoriel. Dans l'exemple que nous avons montré, les vecteurs produisaient une matrice binaire, composé seulement avec 2 valeurs possibles. 1 lorsque le vecteur contenait le mot, 0 sinon. Mais il est possible de pondérer

ces valeurs. Les mesures les plus classiques essaient de traduire la notion d'importance d'un mot. Cela peut se matérialiser par un simple comptage du nombre de fois où un même mot apparaît. Mais certaines mesures dans leur formule, comme le *tf.idf*, se basent plutôt sur le rapport entre ce comptage et le nombre total de mots dans le corpus.

Modèles d'apprentisage

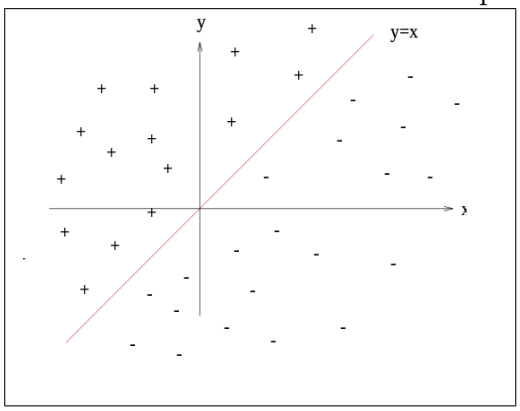
Il existe aujourd'hui plusieurs algorithmes pour apposer une étiquette à une séquence de mots. Une des approches les plus populaires est de considérer la séquence à annoter comme étant indépendante de son voisinage. La représentation dans un espace vectoriel apparaît donc comme le choix privilégié pour modéliser les données d'entraînement. Les machines à vecteurs de support (*Support Vector Machine*, SVM) restent aujourd'hui un algorithme de choix pour de tels tâches de catégorisation.
Cependant, d'autres techniques tentent de capturer les dépendances entre les étiquettes de mots adjacents. Les meilleurs résultats de l'état de l'art sont aujourd'hui obtenus par ces algorithmes, notamment avec les champs conditionnels aléatoirs (*Conditionnal Random Fields, CRF*).
Nous allons aborder brièvement ces deux techniques dans les paragraphes qui suivent.

SVM Les machines à vecteurs de support (ou séparateurs à vaste marge) sont une technique d'apprentissage supervisé, visant à résoudre les problèmes de discrimination (à quelle classe appartient un échantillon ?) et de régression (prédire la valeur d'une variable). Développées dans les années 1990 à partir de la théorie de Vladimir Vapnik (Boser et al., 1992), les SVM ont été largement utilisés pour leur capacité à agir sur des espaces à grande dimension et le faible nombre d'hyper-paramètres que l'utilisateur doit fixer (Zelenko et al., 2003; Zhao et Grishman, 2005). Les SVM sont utilisés aussi dans de nombreux domaines comme l'extraction d'information, mais aussi dans les finances ou dans la bio-informatique.
Pour expliquer le principe général de ce classifieur, nous allons nous placer dans un espace à 2 dimensions (plan). Chaque point a pour coordonnée un couple de valeur (x, y) et peut être positionné sur ce plan. Il s'agit de trouver la droite (si les classes sont linéairement séparables) qui sépare les échantillons appartenant à chacune des classes. On utilisera ensuite cette droite pour la généralisation à de nouveaux échantillons non classifiées dans un problème de discrimination à deux catégories (catégorisation binaire).

Certaines répartitions d'échantillons sont plus complexes et ne sont pas linéai-

rement séparables. Un des apports de SVM est de transformer l'espace des données en un espace de plus grandes dimensions dans lequel la probabilité de trouver un séparateur lineaire est plus importante. Cette transformation est réalisée avec une fonction noyau (Zelenko et al., 2003; Zhao et Grishman, 2005). L'un des grands avantages de leur utilisation réside dans le gain de performance computationnel, en réduisant le calcul de produits scalaires dans un espace de trop grandes dimensions (ce qui est extrêmement coûteux), à une évaluation simple d'une fonction.

FIGURE 5.3 – Problème linéairement séparable



CRF L'utilisation des techniques comme les SVM implique de considérer les évènements comme étant indépendants les uns par rapport aux autres. Mais d'autres modèles tentent de capturer la dépendance entre étiquettes de mots adjacents. Une des approches populaires est d'utiliser le modèle de Markov caché (*Hidden Markov Model*, HMM) ou son dérivé le modèle de Markov caché à entropie maximale (*Maximal-Entropy Marov Model*, MEMM (McCallum et al., 2000a)). Les HMM et MEMM ont été majoritairement employés jusque dans les années 2000. Aujourd'hui, les champs condtionnels aléatoires (*Conditionnal Random Fields, CRF* (McCallum et Li, 2003)) relâchent la forte présomption d'indépendance des évènements des HMM et MEMM, en offrant des mécaniques flexibles et puissantes afin d'étiqueter une observation conjointement avec les étiquettes des observations voisines, au moins du point de vue théorique (Li et al., 2011; Kolya, 2010). Mais pour des raisons de performances, seuls des unigrammes ou des bigrammes d'étiquettes sont utilisés pour l'entrainement.
Le processus de labélisation de nouvelles observations selon le modèle entrainé par les CRF s'appelle l'inférence. Avec un nombre d'étiquette e et une séquence de mots de longueur n, l'ordre de complexité des possibilité de

labélisation est de $O(e^n)$. Ce qui représente potentiellement une tâche très longue avec le matériel informatique d'aujourd'hui. Heureusement, cet ordre de complexité peut être ramené à $O(ne^2)$ grâce à l'algorithme de Viterbi (Viterbi, 2006) basé sur la programmation dynamique.
Les CRF obtiennent de meilleurs résultats que les HMM ou MEMM pour les tâches d'extraction d'informations (McCallum et Li, 2003).

Méthodes à base d'apprentissage statistique : conclusion

Les techniques que nous avons décrites obtiennent des résultats satisfaisants pour la détection d'entités nommées. Cependant, le problème des relations pose encore de très nombreux défis. Les performances des systèmes de détection de relations oscillent entre 50 et 70% de précision au sein d'un programme comme ACE dans lequel les données sont précises et homogènes. Nous avons vu également que les tâches d'enrichissement ontologique définies par TAC sont également un échec avec des performances qui peinent à dépasser les 10%. En réalité, dans un domaine comme le web, chaque système ajoute ses propres démarches spécifique à un cas particulier, les rendant hermétiques et peu portables dès lors que l'on change de domaine applicatif.

Chapitre 6

Description du contexte industriel

6.1 Introduction

La difficulté inhérente à ce travail de recherche et aux réalisations auxquelles il donne lieu, tient à ce qu'il renvoie à la fois à des problématiques théoriques du domaine du Traitement Automatique des Langues, ainsi qu'à des problématiques industrielles. En effet, le contexte de la thèse, qui s'est déroulée dans le cadre d'un doctorat sous convention CIFRE, nous a conduit à devoir satisfaire au double objectif d'un travail pris entre une exigence de qualité académique et un impératif d'industrialisation des résultats. Partant, il nous semble indispensable d'introduire brièvement les principaux axes qui ont motivé plusieurs des choix opérés au cours des travaux présentés dans la suite de ce chapitre.
En premier lieu, nous nous intéressons à la tâche de reconnaissance des entités nommées dans des textes. Plus particulièrement, nous nous intéressons à l'un des nouveaux défis proposés par les Text Analysis Conferences (TAC), à savoir la capitalisation des entités nommées pertinentes, extraites par les systèmes automatiques en vue d'alimenter des bases de connaissances ontologiques. Comme on l'a rappelé plus haut en section 2.4, les performances peu satisfaisantes des systèmes évalués dans le cadre des campagnes TAC questionnent directement la maturité des techniques implémentées dans les applications actuelles. Ce constat nous mène tout d'abord à reconsidérer le rôle et la constitution des ressources ontologiques exploitées par ce type de systèmes d'extraction automatique, dans des formats et à des échelles variées.
En second lieu, nous nous intéressons à la tâche de détection et d'extraction de relations entre entités nommées. Différentes approches existent pour la

mener à bien, les unes appartenant au paradigme des méthodes statistiques, les autres à celui des systèmes à base de règles symboliques. Pour autant, les besoins existants dans l'industrie (notamment, la performance, l'industrialisation des méthodes, la nécessité de répondre à des demandes de clients divers, d'où la nécessité de pouvoir s'adapter à différents domaines d'application) imposent nécessairement des contraintes sur les choix technologiques opérés pour la mise en œuvre des moteurs d'analyse. C'est dans ce contexte que nous nous positionnons, nos travaux héritant de ces contraintes, qui induisent des orientations spécifiques sur ces derniers. Nous entendons y apporter des éléments de réflexion, en soutenant l'idée qu'améliorer les systèmes de reconnaissance des entités nommées dans des textes, en vue de les intégrer dans une ontologie, a un impact positif sur les relations qu'il sera ensuite possible de détecter entre ces mêmes entités nommées ; idée que nous nous attacherons à légitimer dans les chapitres suivants du présent mémoire.

6.2 Capitalisation des connaissances dans une ontologie

La capitalisation des connaissances, c'est-à-dire la capacité pour les systèmes d'extraction de connaissances (comme par exemple les entités nommées), à accumuler les informations qu'ils identifient comme pertinentes dans leur procédure d'analyse est l'une des principales applications de l'EI en milieu industriel à ce jour. L'entreprise dans laquelle cette thèse s'est déroulée le confirme : la demande client s'accroît, à tel point qu'en trois ans, la moitié des projets se sont, à notre connaissance, tournés vers cet objectif de capitalisation. Cet état de fait s'accompagne d'un double constat :

- Les systèmes de gestion des connaissances déployés en entreprise sont confrontés à un nombre toujours plus important d'informations métier à structurer ;
- Du côté de la recherche académique, les travaux se confrontent, quant à eux, au coût important induit par le développement de connaissances structurées établies pour la détection des entités nommées et des relations qui existent entre elles.

Ce sont ces deux points particuliers dont nous introduisons ici les enjeux, en vue de fournir des éléments de contribution plus substantiels au chapitre 7.

6.2.1 Pourquoi s'intéresser à la capitalisation des connaissances dans une ontologie ?

La capitalisation des connaissances dans une ontologie permet :

- d'améliorer la reconnaissance des entités nommées
- d'améliorer la détection des relations entre entités nommées
- d'attribuer une référence aux entités nommées reconnues

Avant tout, rappelons les motivations qui, selon nous, peuvent amener à se questionner sur l'intérêt même de chercher à capitaliser des connaissances dans une ontologie. En premier lieu, l'accumulation d'informations pertinentes telles que les entités nommées, dans une ressource structurée et dans un format informatiquement exploitable, permet d'améliorer la performance des systèmes qui prennent en charge leur détection automatisée, dans de grands ensembles de textes numérisés. Partant, nous l'avons dit, cela a un effet positif sur la performance de ces mêmes systèmes pour structurer l'information, en particulier lorsqu'il s'agit pour eux de détecter les relations entre différentes entités nommées. Enfin (nous y reviendrons par la suite), cela permet de gagner en qualité lorsque ces systèmes associent un référent du monde à chacune des entités qu'ils détectent. De manière générale, la capitalisation des connaissances peut être définie comme recouvrant l'ensemble des techniques de création d'une mémoire des connaissances acquises et détenues par les utilisateurs dans la pratique quotidienne de leurs activités. Typiquement, il s'agit d'expliciter et de formaliser leurs connaissances, afin de les intégrer dans un système informatique de gestion des connaissances, dont le rôle est d'en permettre l'accès et l'utilisation par l'ensemble des utilisateurs.

Historiquement, dans l'entreprise, cette tâche était l'apanage d'ingénieurs de la connaissance et de cogniticiens, chargés de recueillir l'information métier via des interviews d'experts, afin de les structurer en connaissances diffusables et exploitables, le résultat de cette démarche étant *in fine* validé par les experts eux-mêmes. De nos jours, deux impératifs majeurs jouent en faveur de l'informatisation de cette démarche : d'une part, un nombre toujours plus important d'entreprises accroissent leur nombre de collaborateurs et doivent gérer un savoir-faire expert plus étendu et changeant ; d'autre part, il est crucial pour ces entreprises d'assurer la transmission de ces connaissances et savoir-faire d'experts lors des cycles de remplacement de collaborateurs, pour garantir la pérennité des compétences. On comprend donc mieux combien les

techniques d'extraction d'informations adossées aux technologies de Traitement Automatique des Langues revêtent un caractère essentiel et pourquoi elles sont de plus en plus sollicitées.

Du point de vue de la recherche académique, les travaux récents font état de problématiques analogues. En particulier, la première cause identifiée des faibles performances des systèmes évalués, notamment dans le cadre des conférences TAC, est celle du manque de ressources ontologiques qui sont coûteuses à élaborer. En outre, l'un des principaux verrous auxquels les chercheurs sont confrontés dans la phase actuelle est l'appariement entre la forme de référence d'une entité nommée, autrement dit, ce à quoi l'entité nommée renvoie parmi les choses du monde et l'ensemble de ses matérialisations textuelles possibles.

Dans ce contexte complexe, la matérialité textuelle se laisse difficilement rationaliser à des fins d'automatisation. Le recours à l'interprétation par l'humain est bien souvent indispensable pour garantir la qualité des résultats fournis, ce qui rend les ontologies d'autant plus coûteuses à élaborer, exploiter et maintenir. Il apparaît donc légitime de se demander quelles sont les pistes possibles pour élaborer des ressources ontologiques dont les coûts et délais d'élaboration seraient mieux maîtrisés, sans en obérer la performance. C'est ce que nous nous proposons de discuter en abordant la question de la constitution automatique ou semi-automatique des ressources ontologiques.

6.2.2 La constitution automatique ou semi-automatique des ressources ontologiques

Etant donné que la capitalisation des connaissances passe par l'exploitation d'ontologies et que ces ontologies sont coûteuses à développer, comment développer des ontologies à moindre coût ? Ces deux dernières décennies, la démocratisation de l'accès aux outils de micro-informatique a permis aux plus grand nombre de produire et de diffuser des documents numériques, dont le nombre suit une courbe que l'on peut, sans abus de langage, qualifier d'exponentielle. Cette tendance est renforcée, à l'heure actuelle, par la généralisation de l'usage d'Internet en particulier. Dans une société qui se qualifie elle-même "de l'information" ou "de la connaissance", la demande civile, commerciale, académique et militaire, de structuration des informations devient de plus en plus prégnante et les domaines qui les produisent, sont toujours plus variés et nombreux. Partant, force est de constater que le développement des ressources nécessaires au traitement et à la gestion de cette masse de

données peut s'avérer long et coûteux. Cette demande s'accompagne d'un impératif : celui de pouvoir construire des modèles de représentation des connaissances, qui augmentent les informations d'une couche sémantique, formalisée de manière claire, intelligible, idéalement non ambiguë et exploitable par un système informatique. En réponse à ce contexte complexe, les langages artificiels, désormais consolidés, du Web sémantique permettent la constitution de bases de connaissances, notamment afin de structurer la masse d'informations circulant sur Internet. Il faut ici souligner le rôle de pivot occupé par le domaine de l'ingénierie des connaissances, qui fournit à celui du Traitement Automatique des Langues des instruments théoriques adéquats pour offrir aux technologies du Web sémantique une assise formalisée pour la modélisation et la représentation de l'information.

A cet égard, la notion de "Linked Data", encore émergente aujourd'hui, tend à fournir une trame claire et intelligible entre ce que nous nous proposons d'appeler le "web des documents", système où les documents qui peuplent l'Internet sont régis et identifiés au moyen d'hyperliens, au "web des objets", qui sont autant de références aux objets du monde, formellement décrits par les langages artificiels du Web sémantique. Récemment, plusieurs initiatives ont vu le jour, visant à proposer la publication d'informations structurées en accès libre, identifiées et formalisées de manière standard, afin de les rendre réutilisables par chacun : c'est le phénomène des "Open Data" . En accès privé ou public, gratuit ou payant, la question de l'exploitation de ces grandes masses de données par un système d'extraction de l'information est loin d'être une tâche triviale. En particulier, elle induit deux aspects problématiques qui revêtent pour nous un intérêt particulier :

- Comment parvenir à filtrer ces informations ? Quand bien même de nombreuses bases sont librement accessibles, force est de constater que les contenus qu'elles agrègent ne sont pas toujours pertinents pour toutes les applications qu'on souhaiterait pouvoir en tirer.
- Comment garantir la congruence de la structure d'une base spécifique, établie à partir d'un système donné de représentation des connaissances, avec un système différent ? La question de la compatibilité entre les systèmes de structuration et de représentation des connaissances apparaît ici comme un problème loin d'être anodin.

6.3 Contexte et choix de l'approche utilisée pour l'extraction

Le chapitre 8 présente un système d'extraction d'entités nommées basé sur une méthode symbolique, qui satisfait aux contraintes précitées. Nous avons précédemment, au chapitre 5, introduit les deux approches existantes présidant à la mise en œuvre de tout système automatisé dédié à l'extraction d'informations : les méthodes symboliques à base de règles et les méthodes statistiques. Si ces dernières peuvent se prévaloir d'une efficacité certaine sur la tâche de reconnaissance des entités nommées avec des F-mesure supérieure à 0,90 sur les catégories principales, leurs performances sur la tâche d'extraction de relations entre entités nommées en sont loin, les systèmes peinant à dépasser la barre des 70% en précision. Ces approches montrent donc aujourd'hui leurs limites, de façon encore plus manifeste sur des tâches plus ardues telles que celles proposées par les conférences TAC, notamment en termes de capitalisation des extractions dans les bases de connaissances. En outre, elles sont confrontées au manque de données annotées, dont elles ont pourtant besoin en grande quantité pour pouvoir fonctionner et alimenter leurs modèles de connaissance. Si les corpus annotés en entités nommées sont de plus en plus nombreux, ce n'est pas le cas des relations entre les entités elles-mêmes ; notamment parce que le nombre de relations qu'il est possible d'identifier entre des entités nommées est bien moindre que ces dernières, ce qui implique mécaniquement de disposer de corpus en volume bien plus important si l'on souhaite traiter un nombre statistiquement pertinent de relations. A cela, il faut ajouter la difficulté supplémentaire de s'accorder sur les standards d'annotation, aspect que nous avons abordé plus en détail dans la première partie de ce mémoire. Enfin, dernier argument en défaveur des méthodes statistiques : les modèles qu'elles produisent sont généralement difficilement lisibles et exploitables par l'être humain, ce qui leur vaut fréquemment le surnom de "boîtes noires". En ce sens, les procédures adossées à ces méthodes ne laissent que très peu de marge à la modification si des erreurs sont éventuellement constatées dans les sorties qu'elles produisent, ce qui nuit à la possibilité de les maintenir facilement.

Les approches symboliques exploitant des règles souffrent quant à elles d'un défaut majeur : leur coût de fabrication. En effet, les grammaires exploitent des dictionnaires, lesquels doivent être alimentés par une quantité suffisante d'entrées afin d'en assurer une couverture satisfaisante. De surcroît, le processus d'élaboration de grammaires locales peut, lui-même, être long et fastidieux. Si des travaux s'orientent vers une automatisation partielle de ces

tâches coûteuses, ils tendent à tomber eux-aussi dans l'écueil de lisibilité et d'intelligibilité par l'humain, évoqué plus haut au sujet des méthodes statistiques, dont découle nécessairement la difficulté à les maintenir. La maintenance est pourtant un point primordial, en contexte académique et surtout en contexte industriel. En effet, les clients d'une entreprise comme Arisem demandent fréquemment des ajustements et des modifications dans les ressources, suivant leurs besoins. Ils attendent un service rapide, donc une capacité à réagir à leurs retours dans des délais les plus courts possible. Partant, il est indispensable de pouvoir revenir sur certaines analyses du système, afin d'y apporter des révisions et, si besoin, d'en corriger les erreurs. En ce sens, la maintenance et la révision des ressources apparaissent comme deux impératifs majeurs, lorsqu'il s'agit de choisir quelle méthode (statistique ou à base de règles) est la plus adéquate à implémenter dans une perspective d'industrialisation. Selon nous, les méthodes à base de règles sont les mieux indiquées dans ce contexte. C'est pourquoi nous positionnons nos travaux dans le sillon de ces méthodes, en tentant d'apporter des améliorations aux différentes étapes dans la perspective d'un système devant répondre à des contraintes industrielles :

- du coût afférent à la création de ressources ontologiques ;
- de la lisibilité des grammaires symboliques ;
- de la maintenance globale du système et des ressources qu'il exploite, tout au long de leur cycle de vie.

6.4 Description de l'analyseur visant à établir des relations entre entités nommées

Dans les applications d'Arisem, les ressources dictionnairiques permettent de reconnaître des unités lexicales dans un texte à traiter, ce afin d'attribuer à chaque unité lexicale un ensemble d'informations morphosyntaxiques. Ces ressources dictionnairiques sont exploitables par des grammaires, via l'utilisation d'étiquettes définies par les dictionnaires. Plusieurs types de dictionnaires existent, distingués selon leur fonction et le type d'informations qu'ils indexent.
Les dictionnaires dits de langue générale en sont un premier type. Ce type de dictionnaire tire son nom des mots qui le constituent, issus de la langue "de tous les jours", susceptible d'être utilisée par le locuteur natif tout venant d'une langue donnée. Dans notre cas de figure, le dictionnaire de langue générale mobilisé par le système Arisem est au format DELA, conforme

pour une utilisation avec le logiciel Unitex [1]. Il comprend donc le vocabulaire général d'une langue donnée et indique, pour chacune de ses entrées, une série d'informations, notamment l'étiquette morphosyntaxiques, la forme lemmatisée et l'ensemble des dérivations flexionnelles correspondantes. Voici un exemple d'entrée d'un dictionnaire au format DELA. Le tableau 6.4 spécifie les principales étiquettes morphosyntaxiques, pour le français.

```
travailler,.V+G1
travaille,travailler.V+G1+1PS+Pr
travaillez,travailler.V+G1+2PP+Pr
travailleras,travailler.V+G1+2PS+Futr
```

Code	**Signification**
A	adjectif
ADV	adverbe
CONJC	conjonction de coordination
CONJS	conjonction de subordniation
DET	déterminant
INTJ	interjection
N	nom
PREP	préposition
PRO	pronom
V	verbe

FIGURE 6.1 – Exemple d'étiquette du DELA-1

Le formalisme DELA autorise l'instanciation de nombreuses autres étiquettes, selon le niveau de description linguistique dont on dispose à des fins d'analyse. Nous pensons par exemple à l'intégralité des dérivations flexionnelles pour une langue à flexion riche comme le français ainsi que des informations d'ordre plutôt sémantique, telles que le registre de langage par exemple.

L'analyseur d'Arisem permet également l'exploitation de dictionnaires structurés dans un format propriétaire. A la différence des dictionnaires de langue

1. http ://www-igm.univ-mlv.fr/ unitex/

générale, leur structure est conçue pour permettre d'y enregistrer des informations mixtes, à mi-chemin entre les informations dictionnairiques à proprement parler et les motifs morphosyntaxiques utilisés par une grammaire symbolique. Ils présentent l'avantage d'utiliser des motifs de contextualisation simples pour, par exemple, préciser une étiquette sémantique, comme l'illustre le cas ci-après :

```
<arrêt> <rendre.V:K> -> DecisionJuridiction
```

Le segement « arrêts rendu » déclanchera cette règle et l'étiquette « DecisionJuridiction » y sera accolée. Ce type de structure est typique de cette famille de dictionnaires, que nous appellerons "dictionnaires étendus" : il présente l'avantage de fournir un cadre simple et facilement lisible par l'humain, qui permet l'enrichissement des entrées par l'ajout d'informations utiles à la désambiguïsation par l'environnement lexico-syntaxique gauche et droit d'une entrée donnée. Si ces enrichissements ne relèvent pas directement de l'entrée dictionnairique à proprement parler, elles participent de la description des contextes d'occurrence possible de chaque entrée potentiellement ambiguë, comme l'illustre l'exemple ci-dessous :

```
[<habiter> <+PREP>] Orange -> Ville
Washington [<être> <situer> <+PREP>] -> Ville
```

Intéressons-nous maintenant aux grammaires. Dans le cas d'un analyseur tel que celui d'Arisem, les grammaires d'extraction visent à décrire des patrons lexico-syntaxiques complexes, porteurs d'une valeur sémantique dépendant du contexte. Lorsqu'une séquence textuelle épouse la structure des grammaires, elle est enrichie d'une annotation sémantique dite de "haut niveau". Ce faisant, la grammaire peut s'appuyer sur l'ensemble des traitements précédents opérés par les dictionnaires, en héritant d'informations dites de "bas niveau". Il est donc possible, dans un tel cas de figure, d'exploiter directement les dictionnaires en s'appuyant sur les enrichissements associés à chacune de leurs entrées (catégorie morphosyntaxique, traits sémantiques, règles dérivationnelles, entre autres). La grammaire peut également inclure dans ses motifs des unités lexicales encodées informatiquement sous la forme de chaînes de caractères, des symboles spéciaux ou des motifs prédéfinis.

Historiquement, les grammaires exploitées chez Arisem étaient entièrement calquées sur la structure des grammaires Unitex, dont elles se sont progressivement affranchies pour ajouter des fonctions utiles aux opérateurs humain. Aujourd'hui, Unitex tient principalement le rôle d'interface graphique dédiée à l'édition des grammaires, lesquelles ne sont pas compilées par le moteur Unitex, mais par celui d'Arisem, qui prend en charge la génération d'automates à partir des fichiers de description ainsi établis.

Les grammaires peuvent également exploiter des informations issues d'ontologies, qui s'articulent dans leur structure des classes et des instances de classes, en référence au paradigme de la programmation orientée objet. Une classe est définie comme un groupe d'entités nommées de même type. En outre, les classes sont organisées en système au moyen de relations hiérarchiques, lesquelles régissent l'héritage de propriétés entre elles. Ainsi, si la classe "Organisation" subsume la classe "Société", cette dernière est considérée comme une instance de la première et hérite donc de ses propriétés. Une instance est un item identifié de manière unique et appartenant à la population des items d'une classe. Par exemple, "Paris" est un item défini tel qu'il est une instance de la classe "Lieu". A chaque instance de classe peuvent être associées une ou plusieurs lexicalisations, qui correspondent à l'ensemble des variantes graphiques et synonymiques possibles en langue pour chaque instance. Dans ce cadre, l'instance "Nicolas Sarkozy" appartenant à la classe "Personne", peut être lexicalisée de diverses manières, telles que "N. Sarkozy" ou "Mr Sarkozy".

6.5 Orientations et conclusion

Après avoir présenté les différents types de ressources qu'un analyseur tel que celui d'Arisem peut exploiter, force est de constater qu'il recourt essentiellement à de l'analyse syntaxique de surface, matérialisée par l'exploitation de grammaires locales dépendant du contexte. La question, naturelle, qui se pose alors, est liée à l'absence d'une étape d'analyse syntaxique profonde, point qu'il nous paraît indispensable de discuter dès à présent. Selon l'état de l'art, les systèmes d'analyse syntaxique profonde présentent trois inconvénients majeurs pour leur utilisation dans le cadre d'un système industriel :
Le premier inconvénient est d'ordre qualitatif. L'analyse syntaxique profonde intervient, en principe, en amont de l'analyse, ce qui implique nécessairement que chaque erreur potentielle engendre une chaîne de conséquence néfaste : les erreurs se répercutent mécaniquement sur les phases ultérieures de l'ana-

lyse, ce qui représente un facteur de bruit important et superflu, dans un contexte où la précision des analyses prime. L'état de l'art rapporte qu'environ une phrase sur sept est mal analysée pour le français (Abeillé et al., 2003), et encore ce constat est optimiste, vu qu'il ne prend en compte que des phrases simples et de taille réduite. De surcroît, de tels analyseurs mobilisent des méthodes d'apprentissage statistique qu'il est difficile d'améliorer directement en cas d'erreur constatée.
Le deuxième inconvénient est d'ordre quantitatif : c'est un processus coûteux en ressources machine, qui représente un véritable frein dans un contexte d'industrialisation. De façon générale, il est fortement indiqué d'avoir recours à l'exploitation d'algorithmes dont l'ordre de complexité est de $O(n)$ pour l'analyse de grands volumes de documents. Or, l'analyse syntaxique profonde repose, la plupart du temps, sur des algorithmes dont la complexité est d'un niveau $O(n^2)$ ou supérieur, ce qui est incompatible avec le traitement de grandes masses de données. Pour se donner un ordre d'idée, un système de moteur de recherche, opérant à échelle industrielle, doit être capable d'indexer 40 000 documents par heure tout en tenant une cadence de 10 000 requêtes par minute sur une seule machine.
Le troisième inconvénient est celui de la dépendance à la langue inhérente aux procédures d'analyse syntaxique profonde ; il faut un analyseur spécifique pour chaque langue visée. S'il est aujourd'hui aisé de se doter d'un dictionnaire de langue générale pour une large variété de langues, ce n'est pas le cas des analyseurs syntaxiques profonds.

Nous voici au terme de l'examen des contraintes liées au milieu industriel de notre thèse. Il nous paraît important de souligner ici qu'outre les bénéfices que nous avons évoqués au sujet des méthodes symboliques, notre démarche applicative prend en compte l'intervention d'utilisateurs experts que sont les linguistes, qui ont maille à partir avec le système industriel Arisem dans une posture qui va bien au-delà que celle de simples validateurs. Dans ce cadre, nous poursuivons le présent exposé en deux temps :

- tout d'abord, nous commencerons par décrire les étapes de travail inhérentes à la constitution et à l'enrichissement des différents types de ressources décrits plus haut, en particulier les ontologies ;
- ensuite, nous décrirons la méthode proposée pour mener à bien la tâche de génération de grammaires locales pour l'extraction de relations, à l'aune des contraintes induites par les différents choix présentés au cours de ce chapitre.

Chapitre 7

Constitution de ressources pour un système d'extraction d'information

7.1 Introduction

L'hypothèse de travail qui motive l'orientation des travaux présentés ci-après, telle que nous l'avons énoncée au précédent chapitre, procède de l'idée qu'améliorer les systèmes de reconnaissance des entités nommées dans des textes, en vue de les intégrer dans une ontologie, a un impact positif sur les relations qu'il sera ensuite possible de détecter entre ces mêmes entités nommées. A cet égard, le processus de capitalisation des connaissances mérite une attention particulière : les bases de connaissances formalisées en ontologies mise à disposition en accès libre et gratuit sur Internet, posent paradoxalement le problème de leur manque de standardisation, ce qui est un verrou à leur exploitation en vue de peupler de nouvelles ontologies. Deux questions en découlent, d'un point de vue pratique :

- comment filtrer les données enregistrées dans des structures non homogènes d'une base à l'autre ?
- Est-il possible de rendre leur structure exploitable afin de les intégrer dans un autre modèle de représentation des connaissances et quels sont les problèmes rencontrés pour mener à bien cet objectif ?

Nous nous proposons d'aborder ces deux points dans ce chapitre, en tentant d'y apporter des éléments de réponse opératoires. Après avoir présenté le contexte et les enjeux, nous proposerons une méthode de filtrage fondée sur

une heuristique permettant d'estimer, parmi l'ensemble des informations disponibles dans une base de connaissance issue d'Internet, celles qui semblent être des candidats pertinents pour notre objectif. Suite à quoi, nous nous intéresserons au problème spécifique de l'import de connaissances pertinentes dans le modèle que nous exploitons.

7.2 Filtrage des données : un impératif lorsqu'on exploite des ontologies très peuplées

Intéressons nous dans un premier temps aux problèmes de filtrage des données que l'on estime pertinentes, sachant que celles-ci, comme déjà précisé supra, ont un modèle de représentation non standard. Plus généralement, ceci renvoie au problème de la constitution et de l'enrichissement automatiques de ressources dictionnairiques et ontologiques pour l'extraction d'informations. Pratiquement, nous nous intéressons en particulier aux ressources de type *géo-ontologiques*, lesquelles visent le recensement exhaustif des noms de lieux géographiques. Dans le cadre de nos travaux, notre choix se porte sur ce type de ressources car celles-ci peuvent être exploitées dans de nombreuses applications, tels l'analyse de curriculum vitae, d'offres d'emplois, de nouvelles de journaux, ou encore d'informations décisives pour la Défense. Les ressources géo-ontologiques, transversales à plusieurs domaines, nous semblent donc particulièrement utiles de ce point de vue. En particulier, nous choisissons de recourir à l'utilisation de la base GeoNames [1], qui a été constituée dans le cadre d'un projet d'envergure internationale. Nous allons filtrer ces données en vue d'identifier les plus pertinentes. Nous confrontons ensuite les résultats obtenus à la ressource encyclopédique en ligne Wikipédia. Ce choix est motivé par l'objectif d'obtenir, à l'issue de l'étape de filtrage, des connaissances dont la fiabilité est suffisante pour ne pas induire des erreurs dans les étapes ultérieures du traitement. Les difficultés du filtrage découlent essentiellement de l'ambiguïté possible des entrées géo-ontologiques avec des mots de la langue générale ; par exemple, un cas d'ambiguïté est possible lorsque le système rencontre l'entrée *Orange* et doit décider s'il s'agit de la ville ou du fruit, notre but ici étant qu'il parvienne à sélectionner l'occurrence correspondant au référent dont le signifié renvoie à la ville.

1. http ://www.geonames.org/

7.2.1 Description de l'ontologie GeoNames

GeoNames est une base de connaissances géographique, maintenue par une communauté internationale de contributeurs. Ainsi, GeoNames est mise à jour en continu, ce qui évite que les informations qu'elle contient de devenir obsolètes. A ce jour, ce sont plus de 6.5 millions [2] d'entrées qui sont renseignées.
Ces entités se répartissent en 9 catégories, elles-mêmes segmentées en 645 sous-catégories, une granularité fine dans la représentation des connaissances. En effet, GeoNames permet de distinguer aussi bien entre les pays, régions, villes et villages, qu'entre les lieux-dits, parkings ou routes. Chacune de ces entrées contient des informations précises – par exemple, la population, les coordonnées GPS, l'altitude, les subdivisions administratives, le code postal, le tout, en plusieurs langues.

Catégories	**Sous-catégories**
A	pays, état, région...
H	lac, court d'eau...
L	aire, parking...
P	ville, village...
R	route, chemins ferrés...
S	bâtiment, ferme...
T	montagne, colline, rocher...
U	sous-marin
V	forêt, bois...

FIGURE 7.1 – Typologies des 9 grandes catégories de Geonames

Cette base de connaissances est interrogeable de différentes façons ; les données qu'elle contient étant informatiquement enregistrées sous la forme d'une base de données relationnelle, il est possible de recourir aux requêtes SQL – moyen dont nous nous sommes servi ; un service web avec une interface utilisateur riche est également disponible. Les connaissances que la base recèle, enregistrées informatiquement dans un format dont la structure est propriétaire, constituent des ressources mobilisables par le moteur d'analyse d'un système automatisé d'extraction d'informations. De notre point de vue, dans le cadre

2. Chiffre de septembre 2010

d'un tel système, le statut de ces ressources est comparable à celui de ressources linguistiques, tels les dictionnaires ou les ontologies, par exemple.

7.2.2 Stragégie de filtrage proposée

Pour satisfaire aux objectifs précités quant à la sélection d'entités nommées estimées les plus pertinentes, il est nécessaire de filtrer cette ressource. En effet, cette étape de sélection doit être stricte, étant donnée la taille importante de la base de connaissances géo-ontologiques que nous souhaitons exploiter ; ce qui implique mécaniquement un nombre conséquent de candidats ambigus. A cette fin, nous nous proposons d'appliquer une stratégie de filtrage en trois étapes successives :

(i) la première s'appuie sur des informations présentes dans les entrées de la base GeoNames ; en particulier, l'effectif de la population (E) pour une ville et l'altitude (A) pour une montagne ; ici, le principe de filtrage est le suivant : (E) et (A) doivent être respectivement supérieurs à des seuils, fixés de façon arbitraire et qui peuvent être modifiés si besoin ;

(ii) ensuite, la deuxième étape de filtrage exploite les informations de l'encyclopédie en ligne Wikipédia ; on postule ici que l'importance d'un lieu procède de la quantité et de la variété des connaissances encyclopédiques qui lui sont associées. Par exemple l'entrée « BarnHill » peut correspondre à plusieurs lieux différents (BarnHill aux Etats-Unis ou en Ecosse). Wikipédia expose des articles sur chacun des lieux, mais ils sont relativement courts et en seulement trois langues. C'est donc une entrée ambiguë qui peut être source de bruit, et qui ne semble pas relever d'une importance capitale pour Wikipedia. En pratique, les paramètres à partir desquels le système évalue l'importance d'un lieu sont liés au nombre de traductions existant pour un article de l'encyclopédie : on postule ici une corrélation entre ce nombre et la probabilité que l'entité soit pertinente. Aussi simple soit-elle, la mesure que nous proposons prend en compte deux paramètres : la longueur, en nombre de mots, de l'article Wikipédia correspondant à une entrée GeoNames donnée, noté n_m et le nombre de traductions de l'article, noté n_t ; la formule appliquée est la suivante : $F(x) = n_m \left(1 - \frac{1}{\log n_t}\right)$.

(iii) enfin, la troisième étape du filtrage consiste à identifier des entrées potentiellement ambiguës, pour les exclure de la ressource (si on cherche à être précis dans l'extraction) ou pour en reporter la désambiguïsation à des étapes ultérieures de la chaîne de traitement ; à cette fin, nous utilisons un diction-

naire de langue générale et une ontologie d'entités nommées.

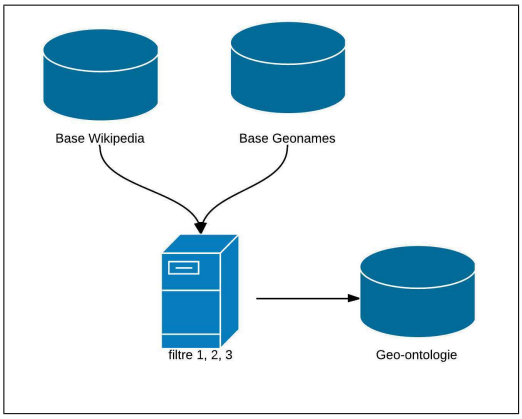


FIGURE 7.2 – Schéma général pour l'import de données de Geonames

7.2.3 Résultats et Discussion

Au terme de notre expérimentation, nous obtenons les résultats présentés dans la table 7.1 :

TABLE 7.1 – Présentation des résultats des données concernant les noms de villes de plus de 1000 habitants et des lieux autres que des villes ; croisement des données entre les bases GeoNames et Wikipédia - les entrées pertinentes sont indiquées dans la colonne « Retenus » et résultent d'un filtrage par mesure d'importance.

	Présents dans Wikipédia	**Ambigus dans Wikipédia**	**Retenus après filtrage**
Villes ($\geqslant 1000 hab.$)	76 007	19 691	35 533
Lieux Autres	75 017	25 147	28 815

Nous avons retenu, pour cette expérimentation, les ensembles de données suivants, à partir de la ressource GeoNames :

– toutes les villes de plus de 1 000 habitants ;
– une sélection aléatoire de 10% des lieux autres que les villes.

Nous avons ensuite effectué un filtrage de ces données par une mise en correspondance des entrées retenues avec celles de l'encyclopdie en ligne Wikipédia. Premier constat : 67,73% des villes présentes dans GeoNames le sont également dans Wikipédia ; ce taux chute à 14,11% quant aux lieux autres que les villes. Second constat : 17,55% des noms de villes sont ambigus dans Wikipédia, c'est-à-dire que l'encyclopédie en ligne indique par exemple les homographes - pour la ville de Rennes, ayant une seule entrée dans GeoNames, Wikipédia présentera une ambiguïté avec les animaux du même nom ; ce taux est de 4,73% pour les lieux autres que les villes. A l'issue du filtrage par mesure d'importance, sur les 76 007 villes présentes dans Wikipédia, 46.75% sont retenues ; sur les 75 017 lieux autres, 38.41% sont retenus. Pour indicatifs qu'ils soient, ces résultats tendent à indiquer que s'adosser à Wikipédia pour le peuplement d'une ontologie de noms de lieux est une stratégie qui peut s'avérer pertinente si l'on s'intéresse aux villes, mais l'est moins pour les autres types de lieux.

Ces résultats nous donnent l'occasion de mettre en exergue les limites de l'heuristique proposée pour la stratégie de filtrage présentée plus haut. Concernant le premier critère de filtrage (i), l'argument immédiatement opposable est son caractère arbitraire : rien ne dit en effet qu'un lieu faiblement peuplé, ne constituera pas à l'avenir un centre d'intérêt majeur lié à des enjeux importants.

Concernant le deuxième critère de filtrage (ii), deux réserves sont à apporter. Tout d'abord, le nombre de traductions d'un article dans Wikipédia n'est pas, en lui-même, un critère qualitatif. En effet, tout comme les articles eux-mêmes, les traductions réalisées par les contributeurs de l'encyclopédie doivent être validées par la communauté – sur le mode de la *peer review*, bien connue dans le monde de la recherche – lorsqu'elles sont proposées ; si elles ne le sont pas, une mention explicite en est faite à même l'article [3] : cela constituerait, selon nous, un paramètre automatiquement détectable dont l'effet serait potentiellement positif quant à la qualité de la procédure de filtrage que nous avançons. Ensuite, les spécialistes de la traduction nous opposeront sans doute, à raison, le fait que la longueur, en nombre de mots, de la version traduite d'un article, ne peut s'estimer de façon solide sans prendre en

3. Un exemple en anglais : « this article provides insufficient context for those unfamiliar with the subject. Please help improve the article with a good introductory style. » trouvé le 21.04.2013 à 18h36 pour http ://en.wikipedia.org/wiki/Kernel_Function

compte le paramètre du coefficient de foisonnement – différence de longueur entre le texte en langue source et sa traduction en langue cible.

Concernant le troisième critère de filtrage (iii), un argument, selon nous majeur, est opposable : le fait de reporter la désambiguïsation ne fait évidemment que déplacer le problème. L'une des pistes que nous aimerions poursuivre à cet égard, dans les travaux consécutifs à ce travail de thèse, serait la suivante : à partir de ressources dictionnairiques et ontologiques, il s'agirait d'attribuer à chaque entité dont l'ambiguïté est probable, un score traduisant cette probabilité d'ambiguïté.

Pour la suite, il paraît indispensable de passer par une remise en contexte de l'expérimentation présentée dans ce chapitre, en s'attachant à décrire la spécificité du problème que pose notre objectif de représentation des connaissances.

7.3 Peuplement d'ontologies : un problème de représentation des connaissances

Nous exploitons les langages du Web sémantique, en vue de développer des ontologies, pour représenter la sémantique de secteurs d'activité bien identifiés par l'industrie, en mobilisant les outils de la logique de description. De notre point de vue, exploiter des entités issues de bases de connaissances disponibles sur Internet, en vue d'en isoler celles qui peuvent être estimées pertinentes, revient à obtenir un ensemble d'entités dont le statut est analogue à des ressources linguistiques adéquates pour nourrir un moteur d'analyse intelligent.
C'est pourquoi il nous paraît indispensable de discuter la procédure de représentation des connaissances en elle-même, censée fournir des outils pour modéliser des connaissances dans des domaines divers et variés, dont les concepts ne portent pas nécessairement la même sémantique d'un modèle à l'autre.

7.3.1 Sur la modélisation des connaissances à structurer dans une ontologie

Parmi les langages du Web sémantique utilisés pour la représentation et la modélisation des connaissances, le standard de référence est le *Web*

Ontology Language, couramment abrégé par l'acronyme OWL [4] . Issu des travaux de recherche dans le domaine de la logique de description, ce langage permet de construire des ontologies par la définition de concepts et des propriétés qui leur sont afférentes, illustrant des domaines concrets. OWL inclut trois systèmes de représentation, qui se distinguent par leur degré d'expressivité, que nous énonçons ci-après, du moins expressif au plus expressif : OWL-Lite, OWL-DL et OWL-Full . Du point de vue de l'algorithmique, l'expressivité d'un langage a une influence directe sur sa décidabilité, c'est-à-dire sa capacité à pouvoir résoudre un problème en un nombre fini d'étapes avec un algorithme. En effet, si de tels algorithmes peuvent être conçus à partir d'OWL-Lite, d'importants problèmes d'inférence existent dans les algorithmes élaborés à partir d'OWL-DL, nécessitant un temps de résolution exponentiel ; il n'existe, à notre connaissance, aucune garantie de parvenir à résoudre un problème avec un algorithme exploitant des données structurées avec OWL-Full.

Dans le contexte industriel où se situent nos travaux, notre choix s'est porté sur l'utilisation d'OWL-DL, d'une part, parce que ce langage présente une puissance d'expressivité suffisante en regard de nos objectifs, d'autre part, parce qu'il est décidable. L'une des contraintes que l'on se donne en contexte industriel pour la modélisation des ontologies, est la disjonction entre classes, c'est-à-dire qu'une entité ne peut appartenir qu'à une seule classe : en principe, une entité ne peut appartenir en même temps à la classe *Personne* et à la classe *Lieu* (La ville de Paris et Paris Hilton par exemple). En pratique, cependant, une entité peut appartenir à deux sous-classes distinctes d'une même classe : une *Personne* peut à la fois appartenir à la sous-classe *Acteur* et à la sous-classe *Personnalité politique*, comme c'est par exemple le cas pour l'entité *Arnold Schwarzenegger*. Un tel modèle matérialise donc l'intuition selon laquelle il est possible de catégoriser exclusivement les objets du monde, sous réserve de déclarations explicites : cela peut être décrit de façon systématique en OWL-DL. Par expérience, ce cas de figure est plus ou moins fréquent en fonction des domaines, c'est pourquoi il est préférable de recourir à la déclaration de propriétés ou de contraintes explicites d'une sous-classe dans une hiérarchie de concepts. Concrètement, pour le cas d'une personne célèbre, il est possible de déclarer les propriétés *est acteur* et *est un homme politique*, pour les lier à la classe *Personne*, plutôt que de définir les sous-classes *Acteur* et *Politicien* comme subsumées par la classe *Personne*. Nous sommes donc dans un cas à nous cherchons toujours à définir une hiérarchie de classes la plus générique possible, que l'on peut enrichir au besoin de

4. http ://www.w3.org/2004/OWL/

contraintes et de propriétés explicites, lorsque le besoin applicatif l'impose.

Nous avons constaté, dans le cadre de notre pratique du terrain, que la description des propriétés joue un rôle essentiel lorsque l'on souhaite modéliser des entités nommées et des relations entre ces entités. Avec OWL, notamment OWL-DL, il est possible de définir des propriétés en tant que relations binaires entre deux entités. Cependant, ce formalisme ne prévoit pas de définition précise des relations n-aires, comme par exemple *Jacques Chirac habite à Paris depuis 2007*. Pour contribuer à pallier ce manque, nous proposons un mécanisme de représentation de ce type de relations n-aires, en appliquant un processus dit de réification. On peut définir la réification comme la transposition d'une abstraction composite, en un tout représenté par un objet concret. En informatique par exemple, nous pouvons définir la réification comme suit :

Soit une instance p de la classe *Point*, contenant deux entiers 4 et 5 dans son état. Le principe de réification consiste à considérer les valeurs 4 et 5 comme un couple, donc un objet unique et manipulable de sorte à pouvoir l'interpréter, par exemple, comme une coordonnée dans un espace à deux dimensions. Ainsi, p est une réification du point de coordonnée (4,5) et donne une existence aux deux valeurs ainsi reliées.

Dans le cadre de notre expérimentation, notre proposition de formalisation est la suivante : nous recourons, pour rendre opératoire la réification, à l'utilisation d'un nœud de réification anonyme – dit anonyme parce qu'il est instancié sans propriété qui lui est propre – lequel relie toutes les instances incluses dans la relation que l'on cherche à matérialiser. Pour illustrer notre propos, prenons l'exemple suivant. Soit la phrase *Chirac habite à Paris depuis 2007*. Dans ce cas, illustré ci-après par la figure 7.3.1, on dit que le nœud de réification anonyme sera rattaché à la classe de la relation *habiter à* et qu'il relie les instances *Chirac*, *Paris* et *2007* - respectivement, instances de la classe *Personne*, de la classe *Lieu* et de la classe *Date*.

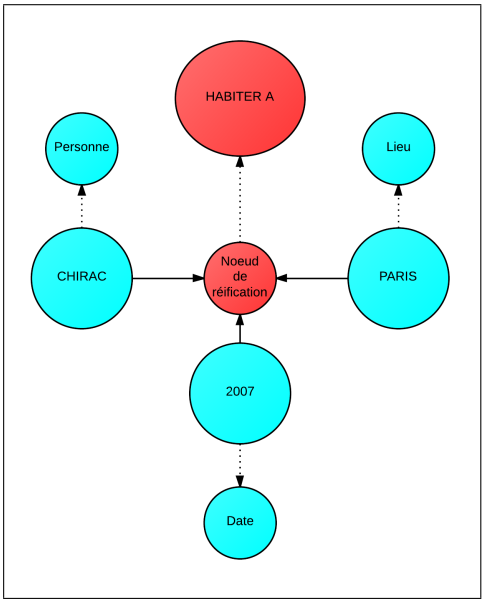


FIGURE 7.3 – Schéma du concept de réification des relations

7.3.2 Contexte de l'expérimentation

De notre point de vue, applicatif, l'objectif général des recherches dans lesquelles se situe la présente expérimentation touche à l'automatisation partielle du processus dit de peuplement d'ontologie. La mise en œuvre de ce processus passe, en l'occurrence, par l'utilisation du langage OWL-DL, présenté plus haut et comporte deux étapes particulièrement importantes qu'il convient de rappeler ici :

- la modélisation des connaissances, qui implique la définition des classes, ainsi que des relations hiérarchiques et non hiérarchiques qui participent à leur organisation au sein du modèle de l'ontologie ;
- le peuplement de l'ontologie, autrement dit l'instanciation des individus appartenant à chacune des classes et sous-classes définies

Ces deux étapes doivent pouvoir être gérées par un système automatisé, capable de :

- comprendre et gérer l'organisation du modèle de l'ontologie ;
- l'alimenter avec des données considérées pertinentes, provenant d'un autre modèle d'ontologie.

La notion de *mapping* ou mise en correspondance – c'est-à-dire la procédure de recherche d'équivalences ou de similarités entre classes et propriétés appartenant à des modèles distincts de représentation des connaissances – prend ici toute son importance. Ceci est dû au fait qu'il est nécessaire, pour répondre à l'objectif de capitalisation des connaissances, comme on l'a évoqué précédemment, d'établir l'ensemble des correspondances qu'il est possible de définir entre deux modèles donnés, pour un ensemble de classes donné. Le caractère paramétrable de l'établissement de ces correspondances est un point central pour nous.

Concrètement, la présente expérimentation s'est attachée à trois domaines différents, qu'il nous a été donné d'aborder dans le cadre de notre mission en entreprise : celui des jeux vidéo, celui des organisations pour la défense militaire nationale, et enfin celui de l'énergie. Dans ce cadre, nous avons utilisé DBpedia [5] pour tenter d'alimenter les ontologies sur lesquelles nous avons travaillé.

7.3.3 De la difficulté d'homogénéiser les structures de représentation des connaissances

Nous sommes ici confronté au manque d'homogénéité entre les modèles mobilisés pour structurer la représentation des connaissances. A cette fin, il s'avère indispensable de disposer d'un instrument pour réaliser le *mapping* entre deux modèles distincts, qui permette d'assurer la cohérence conceptuelle d'une structure de représentation avec une autre. En pratique, la probabilité de rencontrer des écarts importants entre deux modèles est conditionnée par leur dépendance respective à des domaines spécifiques : si l'on se trouve dans le besoin d'alimenter une ontologie technique avec des connaissances provenant d'une ontologie généraliste, le risque de rencontrer des incohérences est non négligeable. Cet état de fait constitue, selon nous, une alerte méthodologique, une question qu'il convient de systématiquement se poser pour établir une stratégie de *mapping* adéquate. Pour illustrer les différents degrés auxquels peut se poser cette question, nous exposons ci-après trois cas basés sur notre retour d'expérience.

Jeux vidéo : un domaine généraliste, peu problématique pour le mapping Notre première illustration concerne le domaine des jeux vidéo. En l'occurrence, nous montrons le cas d'une harmonisation du concept de jeu

5. http ://dbpedia.org/About

vidéo entre deux ontologies : la première issue d'un projet nommé Doxa[6] ; la seconde issue de la ressource DBPedia, évoquée plus haut. Nous avons constaté, après avoir exploré la structure des classes de ces deux ontologies, que le concept de jeu vidéo[7] y est représenté de façon similaire. En effet, l'on trouve une correspondance explicite entre la classe *doxa : VideoGame* et la classe *dbpedia-owl : VideoGame*, ce qui assure une cohérence lorsque l'on cherche à réaliser un *mapping* entre elles.
Il s'agit là d'un cas relativement particulier : la mise en regard des deux ontologies précitées pour d'autres classes de concepts ne permet pas de généraliser la cohérence entre les concepts qu'elles contiennent.

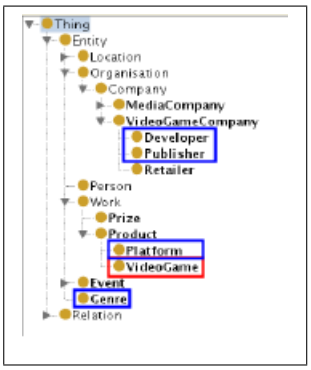


FIGURE 7.4 – Ontologie doxa video-games, sous-ensemble de la hiérarchie des classes

Défense : un domaine semi-spécialisé, partiellement problématique pour le mapping Avançons d'un pas dans la difficulté, pour nous tourner vers le domaine de la défense civile. Nous avons eu l'opportunité de travailler sur le projet CAHORS[8] , lié à l'analyse d'informations extraites d'Internet en vue d'associer à chacune d'elles, de la façon la plus fiable possible, un degré de confiance. Cet aspect est en effet fondamental dans les systèmes d'aide à la décision exploités pour le renseignement et la sécurité intérieure. Lorsque nous avons cherché à opérer une mise en correspondance du concept d'*arme militaire*[9] entre l'ontologie CAHORS et l'ontologie DBPedia, nous avons constaté que la première contenait le concept *Cahors :*

6. Projet sur l'extraction d'opinion lancé par Cap Digital entre 2009 et 2012 auquel nous avons participé. L'ontologie « jeux video » a été conçue spécialement pour ce projet
7. « video game » en anglais
8. CAHORS est un projet ANR de trois ans portant sur la défense et a débuté en 2009
9. Traduit par « military weapon » en anglais

MilitaryWeapon subsumé par la classe *Cahors : Weapon*. Du côté de DBPedia, ontologie généraliste, seul le concept d' *arme* est présent, matérialisé sous l'entrée *dbpedia-owl : Weapon*. La cohérence n'étant pas assurée ici, nous avons résolu le problème en recourant à la spécification d'une relation sémantique telle que : *arme* est l'hyperonyme de *arme militaire*. Dans ce cas, en pratique, nous faisons le choix d'effectuer la mise en correspondance entre *dbpedia-owl : Weapon* et *Cahors : MilitaryWeapon*. Ainsi, force est de constater que la granularité de l'ontologie généraliste s'avère insuffisante pour peupler l'ontologie semi-spécialisée.

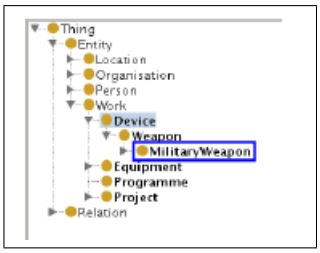


FIGURE 7.5 – Ontologie défense, sous ensemble de la hiérarchie des classes

EDF : un domaine métier, très problématique pour le mapping Abordons maintenant le cas de figure le plus complexe. L'entreprise nous a permis d'avoir accès à une ontologie (elle aussi issue du projet Doxa) couvrant un domaine métier, en l'occurrence l'ontologie décrivant les services de la société EDF. Comme le laissent pressentir nos remarques sur les précédents cas de figure, on peut d'ores et déjà anticiper que la granularité de l'ontologie généraliste DBPedia sera insuffisante pour peupler une base de connaissances métier, a fortiori aussi spécifique que celle d'un domaine tel que celui de l'énergie. Concrètement, on remarque que le concept *doxa : EDFAgency*[10] est spécifique au domaine métier ; il l'est dans une latitude trop élevée pour que l'ontologie Doxa puisse être peuplée par DBPedia, dans laquelle la correspondance la plus proche identifiée est *dbpedia-owl : Company*[11]. A cet égard, il est clair que faire la mise en correspondance depuis DBPedia vers Doxa engendrerait une perte drastique en terme de granularité, laquelle est pourtant essentielle lorsque l'on cherche à alimenter des systèmes automatisés intelligents dont les résultats doivent être précis. La seule issue possible est ici d'envisager le recours à une autre ontologie que DBPedia, dont la structure

10. L'anglais « agency » signifiant « agence » en français.
11. L'anglais « company » signifiant « entreprise » en français.

soit suffisamment fine pour ne pas engendrer une telle perte de précision dans le transfert des connaissances d'une ontologie à l'autre.

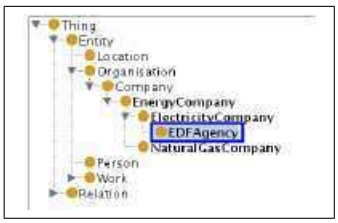


FIGURE 7.6 – Ontologie doxa edf, sous-ensemble de la hiérarchie des classes

En conclusion, nous avons constaté que trois niveaux de compatibilité existent, lorsque l'on cherche à peupler – par le procédé de mise en correspondance – une ontologie cible à partir d'autres :

(i) la compatibilité partielle ; dans ce cas, la cohérence entre les concepts de deux ontologies n'est pas tout le temps possible, mais une exploration détaillée du modèle de représentation des connaissances permet d'y pallier ;

(ii) l'incompatibilité partielle ; en l'occurrence, la différence de grain – générique versus spécifique – entre les deux modèles donnés peut poser problème, mais des solutions satisfaisantes peuvent être avancées ;

(iii) l'incompatibilité totale ; ici, nous n'avons pu identifier de solution adéquate pour résoudre le problème.

Nous identifions donc une problématique directement liée à la compatibilité des modèles de représentation des connaissances. C'est, selon nous, le premier verrou pour la capitalisation des connaissances, qui passe en particulier par le format dans lesquelles sont structurées les connaissances. Nous nous proposons donc de discuter ce point dans la fin de ce chapitre.

7.4 Vers un format pour opérationnaliser la capitalisation des connaissances

Nos travaux visent à permettre l'extraction d'un ensemble d'informations dans des textes par un système automatique, afin de les injecter dans une ontologie cible. De notre point de vue, la cohérence entre la structure du modèle mobilisé par le système d'extraction et les ressources utilisées, est

nécessaire au processus de capitalisation qui nous préoccupe ici. La structuration de données extraites de masses d'information est un levier qui nous semble décisif quant à l'amélioration des ressources linguistiques utilisées par les systèmes. Améliorer ces ressources a, selon nous, un impact direct et positif sur les performances de ces systèmes, lesquels, peuvent à leur tour enrichir ces mêmes ressources, suivant le principe du cercle vertueux. Notre propos sera donc de tenter de rénover la conception actuelle de la représentation des connaissances : aller au-delà de l'annotation des textes au moyen d'étiquettes, pour les identifier structurellement dans une représentation du monde.

7.4.1 L'intérêt des formats d'annotation automatique structurés et exploitable

Revenons tout d'abord sur le type de sorties fournies par les systèmes d'extraction traditionnels. Le plus souvent, il s'agit d'annotations embarquées dans le texte, associées de façon linéaire à chacun des segments analysés, comme l'illustre l'exemple suivant [12] :

```
Mr. <enamex type=person>Donner</enamex> met with
<enamex type=person>Martin Puris</enamex>, president and chief
executive of <enamex type=organization>Ammirati&Puris</enamex>
```

Cet exemple est représentatif d'un style d'annotation structuré par occurrence, c'est-à-dire que, pour une phrase donnée en entrée du système d'extraction d'entités nommées, chacune de ces entités est encadrée de balises dans la phrase analysée en sortie du système. Le texte est restitué tel quel, les balises qui marquent le texte comportant des informations lisibles et interprétables aussi bien par une machine que par un humain. Ce type d'annotation est courant et très largement utilisé dans le domaine du Traitement Automatique des Langues et de l'Ingénierie Linguistique, aussi bien dans les travaux de recherche que dans l'industrie. D'une part, il est facile à prendre en charge dans le cadre de langages artificiels de structuration de l'information, dont *XML* [13] est le standard le plus répandu de nos jours. D'autre part, il existe des langages informatiques pouvant interpréter de telles annotations *XML* – notamment *DOM* [14] – ou opérer des requêtes sur elles – comme *XPATH* [15]. Aussi facile à appréhender qu'il soit, ce style d'enrichissement textuel est, selon nous, un véritable verrou à une exploitation riche des

12. Exemple tiré des conférences MUC
13. Pour plus d'informations sur XML, voir http ://www.w3.org/XML/
14. Pour plus d'informations sur DOM, voir http ://www.w3.org/DOM/
15. Pour plus d'informations sur XPATH, voir http ://www.w3.org/TR/xpath20/

textes ainsi traités. En effet, ce format d'annotation a deux défauts majeurs :

1. il ne permet pas de rendre compte de la hiérarchie de classe lorsqu'une entité est détectée, sans avoir recours à une modélisation externe au résultat ;
2. ce qui le rend, en conséquence, difficilement exploitable par des systèmes intelligents, capables d'associer à chaque instance d'entité nommée détectée un ensemble de savoirs sur le monde, pour faire de l'inférence par exemple (Paris est une ville, une ville est un lieu, donc Paris est un lieu).

Comme l'ont souligné des travaux de recherche antérieurs (Grouin et al., 2011) sur les systèmes d'extraction d'information conçus dans l'objectif de constituer des bases de connaissances, les entités nommées ont un rôle de pivot. Mécaniquement, les relations qu'elles entretiennent ont un statut qui nous paraît essentiel : les relations entre entités peuvent, en toute logique, être exploitées pour la désambiguïsation des entités elles-mêmes. Bien que nous n'ayons pas eu le temps de creuser raisonnablement cet aspect, nos travaux ultérieurs viseront à consolider cette idée, dont nous présentons ici le principe. Selon nous, les relations sont, sur le plan de la matérialité linguistique, le ciment syntaxique qui fournit le contexte local adéquat à la désambiguïsation des entités nommées. Prenons l'exemple de l'entrée ambiguë *Orange* : comment savoir s'il s'agit d'une instance de la classe *Lieu* ou de la classe *Organisation* ? Le contexte local fourni par la présence d'une relation déjà connue par le système permet de résoudre une telle ambiguïté. En effet, si *Orange* est inclus dans une relation de type *aller à*, il s'agit d'une instance de la classe *Lieu* ; à l'inverse, si ce segment de texte appartient à une relation de type *travailler pour*, il s'agit d'une instance de la classe *Organisation*. De notre point de vue, souligner dès à présent cet aspect permet de recentrer la problématique de l'extraction d'information sur le fait que ce sont les relations, et non les entités seules, qui ont le rôle de pivot dans la démarche de constitution de bases de connaissances. Conséquemment, la structuration de bases de connaissances, notamment ontologiques, dans un format permettant à la fois de représenter structurellement les entités et les relations dans lesquelles elles peuvent s'inscrire, a un impact nécessairement positif sur la performance des systèmes d'extraction automatisés, en particulier du point de vue de leur performance dans la qualité, précise, des sorties qu'ils produisent. C'est avec cette idée en tête que nous avançons maintenant la proposition d'un modèle de structuration d'ontologies par instance

de classes.

7.4.2 Pour aller vers la structuration d'ontologies par instance de classes

Afin de poursuivre notre exposé, il paraît pertinent de rappeler que la modélisation des ontologies, dans le cadre du Web sémantique, s'inspire fortement des principes de structuration issus de la programmation orientée objet – couramment abrégée POO. C'est notamment le cas du principe de modélisation distinguant classes et instances de classes. Pour illustrer cette différence, prenons l'exemple suivant : Barack Obama. A la lecture de ce nom propre, le lecteur informé de l'actualité politique conviendra avec nous qu'il s'agit d'une personne de sexe masculin. Le lecteur conviendra également qu'une personne de sexe masculin est également une personne. Dans ce cas précis, *Personne* est une classe, laquelle entretient des relations avec d'autres classes, notamment la classe *Personne sexe masculin*, qui est son hyponyme. On dira que cette dernière classe est peuplée d'instances : *Barack Obama* est une instance de la classe *Personne sexe masculin*. Cette instance peut se distinguer d'autres par le fait qu'elle est caractérisée par des propriétés – comme le fait d'avoir un titre honorifique, par exemple. On dira que cette distinction entre classes et instances résulte d'une méta-modélisation, c'est-à-dire la représentation d'un point de vue particulier exprimé par un formalisme approprié, capable de rendre compte de l'usage afférent à une instance donnée. Une telle représentation des connaissances permet de décrire de manière naturelle et intuitive les connaissances d'un domaine spécifique ; en particulier, une telle représentation permet au système d'opérer l'inférence sémantique appropriée, telle que : si *Barack Obama* est une *Personne sexe masculin*, alors c'est une *Personne* avec des propriétés spécifiques.

Pour permettre l'exploitation des connaissances extraites d'un ensemble d'informations par un système automatique, afin de les injecter dans une ontologie cible, nous proposons un nouveau format de résultat d'extraction qui s'inspire de la méta-modélisation utilisée dans les langages du Web sémantique. L'objectif est de regrouper toutes les mentions et variantes d'un nom d'entité dont tous les éléments ont un référent unique [16]. La représentation de *Barack Obama* dans le format que nous proposons est la suivante :

16. A cet égard, nous rappelons l'importance accordée par les conférences ACE et TAC à la tâche de résolution d'anaphores, qui permettent entre autres d'obtenir ce genre de résultat.

```
<entities class=Person>
   <entity class=ArisemBase:PersonGenderMan
   uri="PersonGenderMan:BarackObama"
   form="Barack Obama">
      <atts>
          <val name="FirstName" value="Barack">
          <val name="LastName" value="Obama">
      </atts>
      <occs>
          <occ offset="199" length="12">Barack Obama</occ>
          <occ offset="367" length="15">President Obama</occ>
      </occs>
   </entity>
</entities>
```

Cet exemple montre comment il est possible de rendre compte de la hiérarchie entre les classes dont les mentions apparaissent dans le texte : la classe *Personne* subsume la classe *Personne sexe masculin*. Ceci est matérialisé ici par l'élément *entity class=ArisemBase :PersonGenderMan* positionné en tant que sous-élément de *entities class=Person*. Ce format indique également la position des occurrences (élément *occ*, avec la position et la longueur en nombre de caractères) rencontrées dans le texte, information qui peut être exploitée à d'autres fins, comme par exemple pour le surlignage des entités nommées dans un texte. L'instance est, en outre, caractérisée par un identifiant unique et une forme normalisée (l'attibut *uri="PersonGenderMan :BarackObama"*), cette dernière étant reprise de l'ontologie source exploitée par le système. Ce principe permet d'apparier structurellement des occurrences différentes , comme *Barack Obama* et *President Obama*.

Concernant les relations, nous proposons un format congruent avec celui proposé pour les entités nommées. En effet, il est possible d'expliciter, dans la structure de connaissances, le type de la relation, son identifiant unique, sa nature, les actants qu'elle implique, ainsi que la phrase dans laquelle elle a été identifiée par le système. Ainsi, pour la phrase *Barack Obama became that day the president of the US.*, la relation *est président de* est identifiée comme reliant deux entités nommées, que sont le nom du président *Barack Obama* et le nom du pays qu'il préside *US*. Voici comment cette relation se matérialise dans la structure XML exploitée par le système que nous avons élaboré :

```
<relations class=StaticRelation>
   <relation class=ArisemBase:PresidentOf
   uri="PersonGenderMan:BarackObama|Country:UnitedStates">
      <args>
          <entity uri="PersonGenderMan:BarackObama">
          <entity uri="Country:UnitedStates">
      </args>
      <occs>
          <occ offset="1265" length="52">Barack Obama became
          that day the president of the US
          </occ>
      </occs>
   </entity>
</entities>
```

Précisons qu'il est nécessaire de recourir à certaines heuristiques pour apparier différentes formes. Considérons pour exemple les deux phrases suivantes :

1. Thierry Derez, patron de la GMF, a rencontré le ministre de l'économie Christine Lagarde lundi 24, pour une concertation sur les taux d'intérêts des plans d'épargne.
2. M. Derez a rencontré Christine Lagarde lors de la conférence nationale du contrôle bancaire jeudi dernier.

Ces deux phrases renvoient à des évènements différents ; en effet, celui de la phrase 1 s'est tenu « lundi 24 », alors que celui de la phrase 2 a eu lieu « jeudi dernier » : la date de l'évènement est donc distinctive. Comme on l'a défini au chapitre 1, nous distinguons entre deux types d'instances de relations :

– les relations statiques, type de relation impliquant des arguments obligatoires – les actants – et dont la valeur de vérité est vraie en tout points du temps, étant donnée une période définie.
– les évènements, type de relation définie par la présence d'arguments obligatoires et d'arguments optionnels – les circonstants, qui ne sont pas toujours présents – et dont la valeur de vérité n'est valable que de façon incidente.

En conséquence, nous composons deux règles d'appariement d'occurrences de relations, selon le type d'instance de relation rencontré. Dans le cas d'un

évènement, il n'y a d'appariement entre deux occurrences de la relation que si les arguments obligatoires et les arguments optionnels sont identiques. Dans le cas d'une relation statique, il n'y a d'appariement entre deux occurrences de la relation que si les arguments obligatoires sont les mêmes.
Une telle heuristique n'est possible que dans le cadre d'un formalisme, tel que celui que nous proposons, où les formes de référence des entités nommées sont spécifiquement distinguées par un identifiant unique. Nos travaux de recherches ultérieurs nous mèneront à consolider la mise en place de mécanismes spécifiquement dédiés aux entités encore inconnues dans une ontologie, pour en gérer l'appariement.

7.5 Conclusion et perspectives

Nous avons abordé dans ce chapitre la constitution d'ontologies à partir de données existantes, en nous posant la double question du filtrage des données enregistrées dans une ontologie source, pour les structurer de façon à les rendre exploitables et intégrables à un autre modèle de représentation des connaissances. Une telle démarche, si elle est permise par les technologies du Web sémantique – qui fournissent des formalismes adéquats pour la structuration des connaissances et les processus d'import de données, dans le cadre de dispositifs partiellement automatiques – est loin d'être triviale. En effet, les standards du Web sémantique ne sont pas encore largement adoptés sur Internet : on est en prise avec une prolifération des modèles de représentation, le plus souvent définis de façon ad hoc pour une application spécifique. Force est de constater que l'on est encore loin d'une véritable adoption de ces standards, qui permettraient pourtant une exploitation en masse des données produites et mises en ligne sur Internet aujourd'hui. Néanmoins, les objectifs d'exploitation industrielle de ces données sont, eux, bien actuels et requièrent des solutions pratiques et robustes. C'est pourquoi nous avons proposé, en premier lieu, une stratégie de filtrage, opérationnelle, qui permet de répondre aux impératifs industriels d'extraction des connaissances pertinentes, tout en évitant au système de fournir des sorties bruitées, car cela nuirait à ses performances. En second lieu, nous avons dressé un panorama des problèmes liés au peuplement d'ontologies, lorsqu'il s'agit d'extraire des connaissances d'une base sources pour les injecter dans une ontologie cible. Les premiers éléments de réponse que nous avons tenté d'y apporter, ainsi que la mise en exergue des limites qui en découlent, sont relatifs à la structuration des connaissances elles-mêmes. Nous avons évoqué le fait que ce sont les relations, et non les entités seules, qui ont le rôle de pivot dans la démarche de constitution de bases de connaissances. Nous faisons un premier pas dans ce sens au cours

du chapitre suivant, au cours duquel nous présentons une démarche pour la génération de grammaires locales afin d'extraire semi-automatiquement des relations entre entités nommées.

Chapitre 8

Génération de grammaires locales pour l'extraction de relations entre entités nommées

Nous avons proposé, dans le chapitre 5, un panorama des différentes méthodes impliquées dans un système d'extraction d'information. Après avoir rappelé, dans le chapitre 6, certaines limites des travaux précédents ainsi que les impératifs liés au contexte industriel de nos travaux, nous avons, au chapitre 7, proposé une nouvelle procédure de constitution de ressources linguistiques. Notre parcours nous a mené à considérer les relations entre entités nommées – et non ces seules entités – comme pivot dans la démarche de constitution de bases de connaissances. Sur le plan industriel, cette démarche, en tant que telle, est loin d'être triviale, parce qu'elle est soumise à une double tension :

1. La demande croissante en ressources linguistiques, adaptées à des domaines toujours plus divers et en constante évolution. Ces ressources doivent en outre être lisibles pour être maintenables ;
2. La contrainte industrielle de performance et de qualité des systèmes, qui exclut le recours aux procédures d'analyse syntaxique profonde, comme précisé plus haut au chapitre 6.

Afin de proposer des réalisations en réponse aux problématiques déjà évoquées, ce chapitre présente une démarche pour la génération de grammaires locales afin d'extraire automatiquement des relations entre entités nommées. En effet, après avoir donné une vue d'ensemble des processus d'ingénierie linguistique mis en œuvre pour la génération de grammaires locales, nous

en décrirons plus particulièrement les entrées et présenterons la façon dont des occurrences de relations entre entités nommées sont identifiées, enrichies, puis organisées par le linguiste, pour être exploitées automatiquement à travers des grammaires locales. Nous soulignons d'emblée l'attention toute particulière apportée à la place du linguiste dans la chaîne de traitement présentée. C'est l'expert linguiste qui apporte au système sa connaissance, qu'il est ensuite possible d'étendre en recourant à des algorithmes d'apprentissage automatique.

8.1 Corpus utilisé

Avant d'aborder plus en détail le processus de génération de grammaires qui fait l'objet de ce chapitre, arrêtons-nous dans un premier temps sur le corpus utilisé pour l'expérimenter. A partir du corpus C17, qui rassemble les articles parus dans le journal Le Monde en 2007 – présenté plus haut au chapitre 4 – nous avons établi un corpus de travail par sélection aléatoire afin que nos données d'expérimentation ne soient pas trop fortement dépendantes de la saisonnalité de l'actualité médiatique. Dans la même logique, le corpus d'apprentissage et le corpus d'évaluation ont été eux aussi construits par sélection aléatoire, correspondant respectivement à neuf dixièmes et à un dixième du corpus de travail [1].

TABLE 8.1 – Principales caractéristiques du corpus

	Nombres d'articles	Nombre de mots	Taille (Mo)
Corpus C17	44 290	17 621 487	277
Corpus de travail	18 069	7 207 138	86.1
Corpus d'apprentissage	16 207	6 454 707	77.4
Corpus d'évaluation	1 862	752 431	8.7

1. Pour garantir la bonne conduite de l'expérimentation, une procédure stricte de vérification a été mise en œuvre pour s'assurer qu'aucun article du corpus d'apprentissage ne se retrouve dans le corpus d'évaluation.

8.2 Approche générale : structuration des ressources, modélisation et chaîne de traitement

L'approche générale dans laquelle nous nous situons s'accompagne donc de deux problématiques spécifiques, que nous avons souhaité aborder à partir d'un corpus de presse :

- la capacité à obtenir une grammaire la plus lisible possible, pour en favoriser la maintenance par un opérateur humain,
- permettre au linguiste d'intergair avec un système automatisé.

Pour tenter d'y apporter des éléments de réponse, nous formulons des suggestions sur la conception et la structuration des ressources linguistiques dans le cadre d'un système logiciel utilisé à échelle industrielle.

8.2.1 Les ressources linguistiques vues comme un code informatique

A bien des égards, la démarche de développement de ressources linguistiques en contexte industriel est analogue à celle mise en œuvre dans le cadre du développement informatique. La première similarité que nous voyons est celui du cycle de vie [2] : tout comme le code d'un programme, le développement de ressources implique le suivi d'un processus spécifique, des vérifications et validations avant la mise en production, ainsi qu'un suivi de maintenance et de mise à jour. Les mécanismes impliqués sont donc suffisamment proches pour les mettre en parallèle, en particulier : la définition des objectifs ; la méthode de conception générale ; le codage ; les tests unitaires et de non-régression [3] ; la maintenance. Dans le monde informatique industriel, le codage et la maintenance sont régies par des règles de bonnes pratiques, qui ont notamment pour objectif de permettre la révision des programmes dans des délais courts, et donc à coût maîtrisé. Ce sont des préoccupations qui ne sont pas au premier plan dans la tradition académique de l'exploita-

2. Le cycle de vie d'un logiciel (en anglais *software lifecycle*), désigne toutes les étapes du développement d'un logiciel, de sa conception à sa disparition

3. Les tests de non-régression d'un programme préalablement testé, après une modification, s'assurent que des défauts n'ont pas été introduits ou découverts dans des parties non modifiées du logiciel. Les tests unitaires désignent l'ensemble des procédures de vérification partielle d'un logiciel

tion des ressources informatiques – aussi bien programmes que ressources linguistiques ; en effet, les protocoles d'expérimentation et d'évaluation ne s'attachent, à notre connaissance, qu'aux performances des ressources linguistiques en termes de rappel et de précision. Aucune démarche, selon ce que nous savons de l'état de l'art, ne prend en compte des critères tels que :

- la qualité de l'organisation d'une ressource linguistique à l'aune de sa capacité à être maintenue et révisée facilement, dans des délais courts ;
- la congruence d'une ressource linguistique avec les différentes étapes de son cycle de vie pour un objectif applicatif donné ;
- le temps passé à la mise à jour et à la révision d'une ressource, mesuré en tant que tel, qui pourrait selon nous former un critère d'évaluation tout à fait pertinent quant à la performance d'une ressource linguistique donnée.

Le contexte industriel, dans lequel cette thèse a été réalisée, nous a maintes fois rappelé à quel point de telles préoccupations sont essentielles et ne doivent pas être écartées de la problématique que nous abordons. En effet, plusieurs constats se sont imposés à nous, dans le cadre des différents projets auxquels nous avons participé :

- le développement et la maintenance d'une seule ressource linguistique requiert l'attention et le travail de plusieurs linguistes ;
- le cycle de développement et d'exploitation de cette ressource est généralement itératif ;
- la ressource devient de plus en plus complexe au cours du développement.

A partir de cet état de fait, il est selon nous essentiel de pouvoir créer des ressources lisibles et faciles à maintenir, afin que tous les linguistes qui y contribuent voient leur tâche facilitée lorsqu'ils doivent l'élaborer ou la réviser. Certains principes de "réusinage[4]"– aussi refactorisation – issus de la programmation informatique, peuvent répondre à ce besoin. De façon générale, il s'agit d'organiser les ressources de sorte à les rendre plus lisibles, sans pour autant en modifier la qualité et la performance : autrement dit, rendre une ressource linguistique plus facile à manipuler, sans influer sur la nature des informations linguistiques qu'elle renferme. Nous voyons deux avantages majeurs à s'inspirer de telles techniques :

4. De l'anglais *refactoring*

1. la maintenabilité [5], définie telle que les ressources linguistiques, parce qu'elles sont plus accessibles, lisibles, documentées et structurées, sont en conséquence plus simples à appréhender, à réviser et à mettre à niveau ;
2. l'extensibilité, définie telle que les ressources linguistiques, parce qu'elles sont organisées et structurées selon des principes de "réusinage", ont la capacité à faire face à des charges d'utilisation variables. Il est possible, facilement, d'en étendre les capacités, notamment la couverture via des mises à jour ou l'ajout de modules.

Le principe de "réusinage" implique en particulier de scinder un programme informatique en plusieurs ensembles et sous-ensembles de fonctions, chaque procédure devant être exprimée de la façon la plus concise possible et nommée selon une convention qui en reflète l'usage. Nous choisissons de transposer ce principe à la constitution de ressources linguistiques : dans notre approche, les grammaires sont organisées en ensembles de sous-grammaires, nommées, autant que faire se peut, selon la réalité d'un phénomène linguistique. Ces aspects sont précisés plus bas dans ce chapitre, lorsque nous abordons la chaîne de traitement. Avant cela, il nous paraît pertinent de détailler nos propositions pour une organisation des grammaires exploitées en contexte industriel par les systèmes d'extraction de relations entre entités nommées.

8.2.2 Un cadre méthodologique pour l'organisation des grammaires locales en cascade

Si nous introduisons le principe de « réusinage » dans notre réflexion sur la constitution de ressources linguistiques, c'est non seulement par souci pratique quant à l'élaboration de ressources exploitées par des systèmes industriels, mais aussi – devrions-nous dire surtout – afin de faciliter, littéralement, la tâche du linguiste. En effet, celui-ci a la charge d'élaborer, renseigner et maintenir la ressource, tout au long de son cycle de vie. Venons-en donc à la description du cadre méthodologique que nous proposons pour organiser les grammaires locales en cascades. Nous avons évoqué plus haut une organisation en cascade des grammaires ; ce principe n'est pas nouveau : des outils tels qu'Intex ou Unitex [6] implémentent déjà l'imbrication de sous-grammaires

5. maintenabilité, notion issue de la terminologie industrielle et largement utilisée dans le domaine informatique, correspond à la capacité d'un système à être rapidement et simplement maintenu et/ou réparé, afin de diminuer le coût en temps et en argent de chaque intervention.

6. www.nyu.edu/pages/linguistics/intex et www-igm.univ-mlv.fr/ unitex

dans des grammaires de niveau supérieur. Cependant, il n'existe pas, à notre connaissance, de cadre méthodologique permettant de guider l'élaboration des grammaires locales en cascade, c'est-à-dire de principe énonçant la façon dont ces grammaires doivent être imbriquées, en vue d'en faciliter l'exploitation et la maintenabilité. C'est sur ce point précis que nous souhaitons proposer une contribution, guidés par le principe de « réusinage ». Le cadre méthodologique d'organisation pour la construction de grammaires locales en cascade proposé, se répartit en quatre niveaux, introduits ci-après :

Le premier niveau relève d'une séparation en sous-grammaires, fondée sur l'identification par le linguiste de différents types de prédicat. Par exemple, pour un évènement comme le rachat d'une entreprisespar une autre, les occurrences réalisées par un prédicat nominal et celles réalisées par un prédicat verbal sont décrites par deux sous-grammaires différentes. Ainsi, les relations qui font intervenir les prédicats « racheter, acquérir » seront dans une sous-grammaire distincte de celles qui font intervenir les prédicats « rachat » ou « acquisition ».

Le deuxième niveau de profondeur est fonction de la position des arguments – donc des entités nommées – par rapport au prédicat auquel ils sont associés dans une même relation. Par exemple, une sous-grammaire décrivant les relations dans lesquelles le prédicat est encadré par ses arguments – "X a racheté Y", qui peut être décrit par la suite : "[Entité nommée 1] [Prédicat] [Entité nommée 2]"– sera distincte de celle décrivant les relations dans lesquelles le prédicat est antéposé à ses arguments – "[Prédicat] [Entité nommée 1], [Entité nommée 2] [...] ". Il s'agit donc à ce niveau, pour le linguiste, d'esquisser une typologie de la position des arguments, pour chacun des types de prédicats identifiés en corpus.

Le troisième niveau de profondeur correspond à l'analyse des segments éventuellement présents entre les arguments et le prédicat de chaque séquence formant une relation. Concrètement, il s'agit, via une démarche automatisée, de fournir une analyse des insertions, par exemple dans une séquence telle que : "[Entité nommée 1] [insertion] [Prédicat] [insertion] [Entité nommée 2]". Il s'agit donc, à ce niveau, de produire autant de sous-grammaires que de phénomènes différents identifiés dans les insertions. Nous pensons que ces zones de texte correspondent à des phénomènes de nature linguistique comme la mise en apposition ou les subordonnées relatives par exemple.

Enfin, le quatrième et dernier niveau fournit une description symbolique, générée automatiquement, des phénomènes du niveau précédent ; ladite description est exprimée au moyen de symboles lexico-syntaxiques.

La méthode de structuration en cascade à quatre niveaux que nous venons d'aborder, présente selon nous trois avantages majeurs pour un linguiste :

1. elle contribue à faciliter la gestion de la ressource car le linguiste n'est plus confronté à une imbrication importante de modules de grammaires à manipuler. Nous supposons que cela lui permet de travailler plus efficacement lors des étapes de construction et de maintenance de la ressource ;
2. elle permet une intervention facilitée du linguiste sur différents niveaux de granularité ;
3. elle articule, selon nous à bon escient, les étapes d'intervention du linguiste et de la machine, le premier projetant des connaissances de la langue et du monde qui seraient coûteuses et complexes à intégrer autrement au système, le second prenant en charge les étapes fastidieuses et chronophages pour un humain.

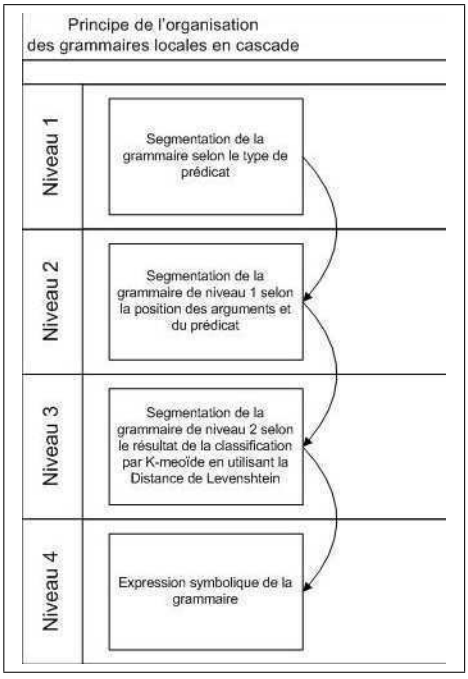


FIGURE 8.1 – Principe général du cadre méthodologique pour l'organisation des grammaires locales en cascade

Des tests de cas d'utilisation, présentés ultérieurement dans ce chapitre, permettront de discuter ces propositions. Avant d'y arriver, nous nous proposons d'aborder la chaîne de traitement mise en place dans le cadre de nos recherches.

8.2.3 Chaîne de traitement proposée

Le but de la chaîne de traitement que nous proposons est, à partir d'un corpus donné en entrée, de produire en sortie un ensemble de grammaires et de sous-grammaires locales pour l'extraction de relations entre entités nommées. Pour ce faire, une succession d'opérations est nécessaire, qu'elle soit mise en œuvre par un utilisateur humain expert – le linguiste – ou par une procédure automatisée. Nous les présentons succinctement ci-après, avant d'en fournir une description plus détaillée dans la suite de ce chapitre. Voici donc les quatre principales étapes de la chaîne de traitement en question (cf. figure 8.2) :

Extraction d'instances de relations entre entités nommées à partir du corpus en entrée : ici, la partie automatisée consiste en l'application d'un algorithme exploitant le calcul de cooccurrences à partir d'un corpus, pour produire un ensemble de phrases candidates contenant potentiellement des occurrences de relations ; le corpus est ici vu simplement, comme un ensemble de phrases prédicatives ; les relations candidates identifiées automatiquement sont soumises au linguiste, dont le rôle est à la fois de valider ou d'écarter certaines d'entre elles, et de compléter les connaissances du système en l'alimentant d'informations spécifiques au domaine.

Catégorisation supervisée des instances de relations validées pour la génération de grammaires : une fois les relations non valides écartées par le linguiste, celui-ci élabore un ensemble de règles de catégorisation permettant de distinguer entre les différents types de relations qu'il a identifiés, en particulier en fonction de la position syntagmatique des différents prédicats – verbaux ou nominaux – et des actants qu'ils impliquent ; ces règles sont intégrées au système, qui procède ensuite, en fonction d'elles, à une catégorisation automatique de l'ensemble des relations validées. Par exemple, une catégorie sera créée pour les relations dont le prédicats est placé entre les deux entités nommées. Une autre sera dédiée pour les relations dont le prédicat est placé avant les deux entités.

Affinage automatique de la granularité des relations pour la génération de sous-grammaires : pour chaque catégorie de relations ainsi obtenue, une seconde étape de segmentation systématique est appliquée par le système, de façon complètement automatisée ; le but est de préciser les règles de description permettant de rendre compte des syntagmes intermédiaires positionnés entre les entités nommées et le prédicat ; cela permet au système de construire des sous-grammaires ; à cette fin, nous exploitons l'algorithme d'apprentissage non-supervisé dit des k-médoïdes, qui s'appuie sur la distance de Levenshtein.

Génération de grammaire par exploitation de la « mémoire de Levenshtein » : pour pouvoir générer la grammaire qui sera exploitable par le système d'extraction, le programme s'appuie sur des connaissances préalablement constituées au fil des étapes en amont, et s'appuie en particulier sur certaines données mémorisées dans le calcul de la distance de Levenshtein, au moyen d'un procédé que l'on nommera mémoire de Levenshtein [7].

Nous détaillons dans les sections qui suivent chacun des points que nous venons d'esquisser rapidement.

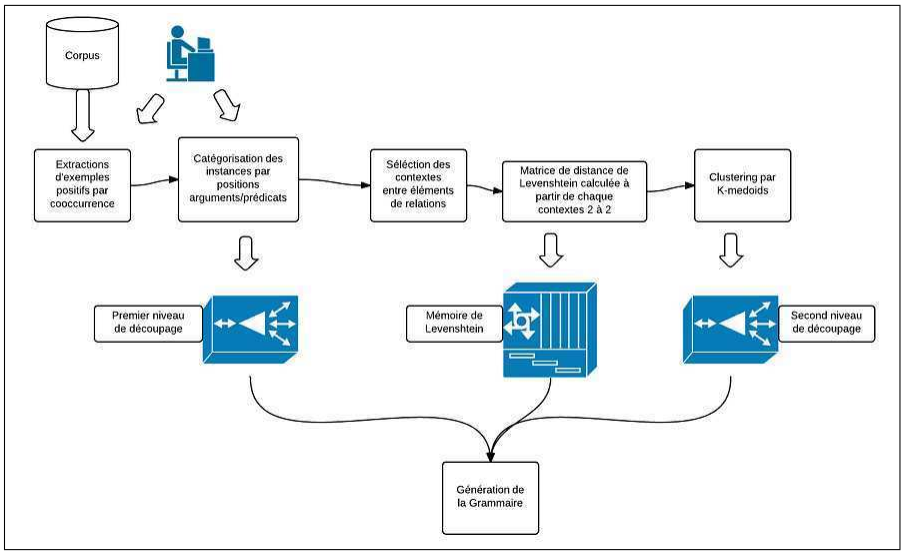


FIGURE 8.2 – schéma général de la procédure implémentée, en regard du cadre méthodologique proposé

7. L'algorithme du calcul de distance de Levenshtein comporte des résultats intermédiaires que nous allons réutiliser

8.3 Extraction d'instances de relations assistée par le linguiste

Nous commençons d'abord par préciser la façon dont le contexte de nos travaux de recherche [8] a influé sur la réflexion que nous avons menée au cours de cette thèse. En effet, il faut souligner, pour comprendre le cheminement que nous avons suivi, qu'au début de nos recherches, la réflexion sur les applications liées à l'extraction de relations entre entités nommées était tout juste naissante chez Arisem. Conséquemment, les besoins identifiés par l'industrie et auxquels la recherche apportait de premiers éléments de réponse, se situaient dans la lignée des applications traditionnelles de l'extraction d'information, notamment pour pouvoir fournir une couche sémantique exploitable par les moteurs de recherche industriels (navigation par facettes par exemple). Le temps passant, de nouveaux besoins se sont fait jour, parmi lesquels la veille d'information, intimement liée à la question de la capitalisation des connaissances : ces questions apparaissent aujourd'hui comme les principaux enjeux liés à l'extraction de relations entre entités nommées.

C'est donc autour de ces nouveaux besoins que nos travaux ont évolué ; à la question "Quelle sont les relations intéressantes à extraire dans un corpus ?", s'est, au fil du temps, ajoutée la question "Comment extraire une relation déjà connue en amont dans ce même corpus ?". Nous nous sommes donc retrouvés en situation de devoir proposer une méthode qui permette à la fois d'explorer efficacement un corpus pour y identifier les relations qu'il serait pertinent d'extraire, et aussi de sélectionner des occurrences de relations préalablement connues ou définies dans la base de connaissances.

Cette première étape d'extraction d'instances de relations entre entités nommées à partir du corpus en entrée, a donc un double objectif :

1. en premier lieu, il s'agit de permettre au linguiste un accès facilité au corpus, pour avoir rapidement une vue globale des données ;
2. en second lieu, il est question pour le linguiste de valider les choix qui lui sont soumis par le système, à savoir les relations entre entités présentes dans le corpus ; ces choix, une fois validés par le linguiste, ont

8. Nous rappelons que notre thèse s'est déroulée dans le cadre d'un contrat CIFRE, nos travaux ont donc naturellement suivi l'inflexion du cadre applicatif dans lequel ils se sont déroulés.

vocation à alimenter la base de connaissances exploitée par les modules présentés dans les étapes ultérieures de notre chaîne de traitement.

Pour permettre au système automatisé de proposer des fonctionnalités d'exploration efficaces et de soumettre des propositions de qualité au linguiste, nous avons fait le choix d'y intégrer un algorithme statistique d'analyse de segments textuels par cooccurrence pour sa rapidité de mise en œuvre et de résultats. L'acception courante de la cooccurrence, qui nous est donnée par Dubois et al. (2001), dit qu'il s'agit d'un "groupe de mots apparaissant fréquemment ensemble". Cette acception exprime ce que Lebart et Salem (1994) définit, de façon plus technique, comme la "présence simultanée, mais non forcément contiguë, dans un fragment de texte (séquence, phrase, paragraphe, voisinage d'une occurrence, partie du corpus, etc.) des occurrences de deux formes données". Il s'agit donc bien d'une combinaison lexicale qui, sur l'axe syntagmatique, n'est pas orientée ou normée. De plus, la cooccurrence associe un nombre d'éléments qui n'est pas prédéfini, dont la contiguïté n'est pas obligatoire et qui n'entretiennent pas nécessairement une relation d'ordre sémantique ; c'est un phénomène statistiquement identifiable sur lequel le linguiste peut projeter sa connaissance.

Les recherches issues du domaine de la statistique textuelle ont fourni, sur la dernière décennie, des résultats expérimentaux qui mènent à soutenir l'idée que des segements textuels cooccurrents entretiennent souvent un lien sémantique, même si ce n'est pas toujours le cas. Nous renvoyons, notamment, aux travaux de Martinez (2000), qui traitent de la mise en évidence de rapports synonymiques par la méthode des cooccurrences. Par ailleurs, nous retenons des travaux antécédents la nécessité de pouvoir retourner facilement au texte pour vérifier et valider les relations identifiées automatiquement à partir de corpus.

Ainsi, nous nous appuyons sur l'hypothèse de l'existence d'une relation sémantique entre deux entités nommées cooccurrentes dans une fenêtre donnée du corpus. Du point de vue de la chaîne de traitement que nous proposons, nous escomptons qu'elle parvienne à extraire des instances de relations sémantiques existant entre des entités nommées cooccurrentes. Du point de vue de l'implémentation informatique qui en découle, nous intégrons un mécanisme permettant un retour au texte, afin de permettre au linguiste de contextualiser la connaissance qu'il projette sur les phénomènes identifiés.

8.3.1 Exploration du corpus pour l'identification d'une classe de relations pertinentes sans connaissances préexistantes

"Quelles sont les relations intéressantes à extraire dans un corpus donné ?" est donc la première question à laquelle nous avons été confronté dans le cadre de nos travaux ; autrement dit proposer une méthode permettant, en premier lieu, d'explorer efficacement un corpus pour y identifier les relations qu'il serait pertinent d'en extraire. Le but visé était de permettre au linguiste, face à un nouveau domaine, sans disposer de connaissances formalisées au préalable, d'identifier de manière non supervisée des relations potentiellement pertinentes. La méthode proposée pour y parvenir consiste en une analyse des cooccurrences présentes dans une fenêtre correspondant à la phrase. Trois raisons ont motivé ce choix de fenêtrage. Tout d'abord, le but est d'extraire des entités cooccurrentes dans le cadre de structure prédicatives, lesquelles se retrouvent dans des empans textuels correspondant à la proposition ou à la phrase. Ensuite, le recours à une solution permettant la résolution de coréférences n'est pas indiqué au regard de nos objectifs applicatifs car la question de la résolution de la coréférence est encore, à l'heure actuelle, un verrou que les solutions existantes ne parviennent pas à lever de façon suffisamment satisfaisante pour justifier leur intégration dans une chaîne de traitement industrialisable. En effet, selon nos expérimentations (Ezzat, 2010), environ 30% des relations entre entités nommées apparaissant dans des textes français incluent des coréférences. Enfin, nous sommes guidé par l'intuition linguistique que la probabilité pour que deux entités nommées entretiennent une relation est plus importante si la fenêtre de cooccurrence est restreinte à la phrase par rapport à une fenêtre plus large, comme celle du paragraphe par exemple.

Concrètement, voici la façon dont nous procédons :

L'analyseur d'Arisem est exploité pour la segmentation des textes du corpus en phrases ; ce même analyseur est mobilisé pour la détection de l'ensemble des entités nommées présentes dans chacune des phrases. A cette étape, plusieurs points sont à noter quant au paramétrage de notre algorithme, afin de le rendre le plus adaptable possible. Tout d'abord, le nombre d'entités nommées par phrase qu'il est possible de détecter n'est pas limité à un maximum et il doit y avoir au minimum deux entités nommées dans la même phrase, pour que le résultat soit pris en compte pour l'étape ultérieure. Ensuite, le type des entités nommées n'est pas contraint, ce qui permet de considérer toutes les

relations possibles entre elles. Nous considérons qu'il existe une cooccurrence entre deux entités nommées si et seulement si ce même couple apparaît dans deux phrases distinctes, disons la phrase A et la phrase B.

Les cooccurrences d'entités nommées détectées dans la fenêtre de la phrase, sont ensuite classées selon un score de cooccurrence. Là encore, apportons quelques précisions. Nous posons que deux cooccurrences détectées dans une phrase A et dans une phrase B sont identiques si et seulement si :

- les entités nommées sont les mêmes dans A et dans B – c'est-à-dire que chacune des entités de la phrase A est strictement équivalente, du point de vue de la chaîne de caractères qui la constitue et de l'étiquette d'annotation qui l'accompagne, à chacune des entités nommées de la phrase B.
- elles comportent exactement le même nombre d'entités ; Si « Microsoft » et « Google » sont présentes dans la phrase A et B, mais si B contient en plus « Apple », alors nous considérons que ce sont deux cooccurrences distinctes.

Le score de cooccurrence est défini par la fréquence absolue des cooccurrences identiques identifiées dans un même corpus.

Enfin, un mécanisme de retour au texte permet au linguiste d'observer chaque cooccurrence d'entités nommées dans le contexte local de la phrase dans laquelle elle apparaît, ainsi que dans le contexte global de l'article contenant ladite phrase ; ce mécanisme permet de déterminer si une relation sémantique existe entre les entités en relation de cooccurrence.

Pour illustrer sur un exemple concret les limites et les apports de la procédure proposée, nous l'avons appliquée sur un extrait de corpus dont les données nous ont été accessibles dans le cadre du projet Infom@gic [9]. Il s'agit d'un ensemble de dépêches émises par l'Agence France Presse, portant sur la crise politique en Côte d'Ivoire ; le corpus utilisé pour cette expérimentation comporte 13 570 dépêches, pour un volume de 6 586 044 mots, ce qui représente une taille de 69,1 Mo. Après avoir segmenté le corpus en phrases et détecté les entités nommées s'y trouvant, nous avons appliqué la méthode d'affecta-

9. Infom@gic est un projet de Recherche et Développement dans le domaine de l'Analyse de l'Information, visant la mise en place d'un laboratoire industriel de sélection, de tests, d'intégration et de validation d'applications opérationnelles des meilleures technologies franciliennes dans le domaine de l'ingénierie des connaissances. Il avait vocation à produire une plate-forme commune d'interopérabilité, laquelle doit faciliter l'intégration de composants d'analyse de l'information dans les domaines de la recherche, de l'indexation, l'extraction de connaissances et la fusion d'informations multimédias.

tion de scores aux cooccurrences d'entités identifiées. 67 388 cooccurrences d'entités nommées ont été trouvées par le système. Nous présentons, dans le tableau 8.2, les résultats obtenus, en indiquant, pour chaque niveau de score, le nombre de cooccurrences correspondant.

TABLE 8.2 – Résultats de l'expérimentation préliminaire réalisée - corpus Infom@gic sur la crise en Côte d'Ivoire

Nombre total de cooccurrences : 67 388	
Score	**Nombre de cooccurrences**
1	54 490
2	8 561
3	2 275
4	809
5	385
6	247
7	126
8	81
9	66
10	41
11 et supérieur	307

TABLE 8.3 – Exemples de phrases issues du corpus Infom@gic sur la crise en Côte d'Ivoire

	Plage de score	Phrase dans laquelle la cooccurrence est identifiée
Exemple 1	Score supérieur à 10	Le MPCI (mouvement patriotique de Côte d'Ivoire) appelle à une réaction.
Exemple 2	Score de 5 à 10	Laurent Gbagbo arrêté lundi par les forces d'Alassane Ouattara, ce n'est pas une première.
Exemple 3	Score de 5 à 10	Alors qu'Alassane Ouattara crée la carte de séjour pour les travailleurs étrange, le chef de l'opposition Laurent Gbagbo est condamné à deux ans de prison.

.

Le nombre de phrase avec une combinaison unique d'entités nommées (c'est-à-dire que la même combinaison ne se retrouve nulle part ailleurs dans le corpus) est très important. Il ne faut pas perdre de vue qu'en pratique, le traitement manuel d'un nombre élevé de candidats est une tâche chronophage, qui rend rédhibitoire l'exploration au cas par cas pour le linguiste.

Nous expliquons ce nombre important par deux raisons, selon nous, principales :

1. d'une part, le système d'Arisem ne comporte pas de module prenant en charge l'appariement entre entités nommées ; par exemple, "Chirac" et "Jacques Chirac" sont, dans ce cadre, considérées comme deux entités distinctes ;
2. d'autre part, le fait qu'à des fins exploratoires, la configuration de la procédure se caractérise par une absence de contraintes sur le nombre d'entités nommées cooccurrentes pouvant être détectées dans une phrase, ainsi que sur leur type ; en effet, il n'est pas rare de rencontrer des suites d'entités de types variés – noms de personne, expressions spatio-temporelles, expressions numériques, noms topographiques, par exemple – dans une même phrase. Si le nombre est trop grand, il est peu probable de rencontrer des cooccurrences identiques.

Le tableau 8.3 présente des exemples de phrases selon le nombre de cooccurrences. Le nombre important de phrases incluant 2 à 4 entités nommées est aussi trop important pour en permettre l'exploration détaillée.

Concernant les cas dont les scores sont supérieurs à 10, nous avons constaté qu'ils consistent principalement en des entités nommées qui, si elles sont bien cooccurrentes, ne sont cependant pas reliées par un prédicat. L'exemple 1 du tableau 8.3 illustre d'ailleurs bien ce cas de figure : les entités nommées "MPCI" et "mouvement patriotique de Côte d'Ivoire" sont fréquemment en cooccurrence ; cependant, ces cas de cooccurrences relèvent plus d'une relation définitoire entre l'acronyme et son explicitation, que d'une relation prédicative. A cet égard, ce ne sont pas, de notre point de vue, les cas de figure les plus productifs sur lesquels le linguiste, qui agit dans le but de valider et de typer les relations prédicatives existantes entre des entités nommées cooccurrentes dans une phrase, doit porter son attention.

En résumé, les phrases présentant des combinaisons d'entités nommées apparaissant entre 5 et 10 fois semblent être les plus utiles à fournir au linguiste pour validation. En effet, leur exploration est possible de manière efficace : quelques centaines de cas peuvent être explorés relativement rapidement. Il est indispensable de souligner ici que, toutes choses égales par ailleurs, les relations sémantiques présentes dans ces ensembles d'entités nommées cooccurrentes peuvent être de nature différente. D'après nos observations, il peut

ou non s'agir de cas de cooccurrences impliquant une relation prédicative. Les exemples 2 et 3, présentés ci-dessus, illustrent bien notre propos : on voit bien que, si les entités "Laurent Gbagbo"et "Alassane Ouattara" sont cooccurrentes dans les deux cas, les relations dans lesquelles elles apparaissent dans chacune de ces phrases sont bien distinctes. En effet, l'exemple 2 indique une relation sémantique exprimée par un prédicat, dans laquelle il est question de l'arrestation de l'une des deux entités par la seconde. En regard de quoi, l'exemple 3 juxtapose deux propositions renvoyant à des évènements contemporains, sans contenir de prédicat entre ces deux propositions.

Cette expérimentation préliminaire, dont nous venons de souligner les limites, met cependant en évidence certains apports de la méthode proposée, selon nous non négligeables pour le linguiste. En premier lieu, la méthode offre l'avantage de réduire l'étendue des cas de cooccurrences d'entités nommées sur lesquelles le linguiste doit concentrer son attention. De ce point de vue, nous satisfaisons à l'un de nos objectifs initiaux, celui de permettre au linguiste un accès facilité au corpus, pour en avoir rapidement une vue globale. Dans le cas de ce corpus, l'examen des phrases comportant des combinaisons d'entités nommées apparaissant entre 5 et 10 fois montre que la relation dominante est celle de "contact entre deux personnes". Cela reflète bien la nature du corpus, rapportant de nombreuses prises de contact entre personnes pour essayer de résoudre la crise ouverte en Côte d'Ivoire, sujet du corpus en question. En somme, la méthode proposée fournit des pistes concrètes permettant de :

- fournir un mode opératoire facilitant le travail du linguiste dans le cadre de la validation des choix qui lui sont proposés par le système ;
- fournir des candidats de cooccurrences entre entités nommées qui sont pertinents, afin de découvrir de nouvelles classes de relations de façon endogène et sans connaissances extérieures.

8.3.2 Cooccurrences entre entités nommées avec apport de connaissances supplémentaires

"Comment extraire une relation déjà connue dans ce même corpus ?" est la seconde question à laquelle nous avons été confronté dans le cadre de notre travail chez Arisem. Nous avons vu dans le cadre de la précédente section,

que l'absence de contraintes fortes sur le type des entités nommées lors de la recherche de cooccurrences, a pour principal inconvénient de générer un nombre important d'occurrences candidates, ce qui peut entraîner un temps de validation prohibitif. De plus, il n'est pas toujours nécessaire d'adopter une approche purement exploratoire : on peut avoir, en pratique, un préalable qui, étant donné un corpus et un domaine de connaissance, permet d'avoir une idée a priori des relations pertinentes. Pour répondre à de tels de cas de figure, la méhode que nous proposons peut intégrer des contraintes, sur le type de relations à rechercher notamment. Ces contraintes peuvent être matérialisées par deux critères, que sont le type des entités et les lexicalisations des prédicats, que nous nous proposons de préciser ci-après :

1. le type des entités va souvent de paire avec le type de la relation qui les lie ; dans le domaine de la finance, la relation de rachat implique par exemple des entités de type "Organisation" (c'est-à-dire acquisition d'une organisation par une autre) ; autrement dit, il est possible de contraindre le système en lui indiquant que l'on souhaite rechercher des cooccurrences de relations dont les arguments sont des entités d'un type donné ;
2. la lexicalisation des prédicats procède de la même logique ; restons dans le domaine de la finance : la présence du verbe "acquiert" dans une phrase constitue un indice fort de l'existence d'une relation de rachat ; le linguiste peut donc, éventuellement aidé par un expert du domaine, établir une liste des prédicats représentatifs de la relation ciblée ; précisons ici que le système permet, à partir d'une liste de prédicats lexicalisés sous leur forme lemmatisée, de rapatrier parmi les résultats les formes fléchies correspondant à ces prédicats.

Ces deux critères, fournis en même temps comme contraintes au système, doivent donc en principe permettre de réduire drastiquement le nombre de cas soumis pour validation au linguiste. Pour illustrer cela sur un cas pratique, nous nous proposons, à partir des données du corpus Le Monde – présenté au début de ce chapitre – d'appliquer la méthode proposée en ciblant dans le corpus une relation précise : la relation d'achat d'une entreprise par une autre. A cette fin, la contrainte imposée sur les entités nommées ciblées dans les cooccurrences recherchées est qu'elles soient de type "Organisation". Pour cette expérimentation sur le corpus de travail, une liste initiale de prédicats, contenant des verbes ou des noms déverbaux, a été fournie au système, tels que :

"fusion, achat, rachat, acquisition, acheter, acquérir, OPA". Nous présentons, dans le tableau 8.4, les résultats obtenus, en indiquant, pour chaque niveau de score, le nombre de cooccurrences correspondant. Le talbeau 8.5 montre quelques exemples de phrases candidates.

TABLE 8.4 – Résultats de l'expérimentation réalisée sur le corpus de travail Le Monde

Nombre total de cooccurrences : 2 285	
Score	**Nombre de cooccurrences**
1	1 543
2	234
3	98
4	123
5	112
6	66
7	32
8	13
9	4
10	1
11	1

Premier constat : d'une expérimentation à l'autre, étant donnés des corpus de taille comparable, on remarque d'emblée que le nombre de cooccurrences est drastiquement réduit par l'application des contraintes évoquées plus haut. Deuxième constat : Une revue des cooccurrences uniques montre qu'elles contiennent effectivement le type de relation recherchée – ce qu'illustrent les exemples 1 et 2. D'autres constats peuvent être faits, illustrés par les exemples 3 à 8, qui rendent compte d'écarts constatés par rapport à la catégorisation attendue, que nous nous proposons de commenter. Quatre cas de figure peuvent être distingués :

1. les cas dans lesquels une négation est présente dans la lexicalisation du prédicat ; en particulier, les exemples 3 et 4 montrent des flexions du verbe "renoncer", dont le sens intrinsèque induit une négation ; de notre point de vue, l'écart par rapport à l'attente est mineur : le sens de la phrase ne correspond pas à l'action effective de rachat, cependant le contexte de tentative de rachat d'une entreprise par une autre pourrait être utile dans le cadre d'un objectif d'application tel qu'une veille

TABLE 8.5 – Exemples de phrases issues du corpus Le Monde

	Type de relation	Phrase dans laquelle la cooccurrence est identifiée
Exemple 1	Relation d'achat	Le 5e fabricant mondial d'éoliennes le danois Bonus, finalement racheté par Siemens et dont le prix a, depuis, quintuplé...
Exemple 2	Relation d'achat	Un montant record pour une opération de 44 milliards réalisée par Texas Pacific Group pour l'acquisition du groupe énergétique texan TXU (Le Monde du 27 février).
Exemple 3	Relation avec négation	Après plusieurs mois de plaintes, de recours et de blocages juridiques, la situation est aujourd'hui relativement claire : Gas Natural a renoncé à son OPA et E.ON a la voie libre pour le rachat d'Endesa.
Exemple 4	Relation avec négation	L'allemand E.ON a finalement renoncé à son offre publique d'achat (OPA) sur Endesa de 42,3 milliards d'euros.
Exemple 5	Action de rachat non finalisée	Le groupe énergétique allemand E.ON a relevé, vendredi 2 février, son offre d'achat à 38,75 euros pour chaque action de l'électricien espagnol Endesa, le valorisant à 41 milliards d'euros, contre 36,5 milliards lors de sa précédente proposition.
Exemple 6	Action potentielle	De par sa taille, un rachat d'ABN Amro par Barclays relativiserait les opérations précédentes.
Exemple 7	Sous-spécification du type des arguments	Le chef de l'état annonce "clairement la fusion de l'ANPE et de l'Unedic".
Exemple 8	Absence de relation prédicative	8 mars : "On tourne en rond", s'inquiète l'équipe de campagne du candidat de l'UMP, fragilisé par les révélations du Canard enchaîné sur les conditions d'achat de son duplex à Neuilly..

sur l'actualité financière en général, non exclusivement focalisée sur les rachats effectifs réalisés par des entreprises ;

2. les cas où le prédicat ciblé indique que l'action recherchée n'est pas achevée, comme le montre l'exemple 5, où le rachat d'une entreprise par une autre n'est pas thématisé dans la phrase, mais dont l'interprétation

permet de comprendre qu'une action de rachat est en cours, sans pour autant en préciser l'issue ; ce cas de figure peut être rapproché des cas tels que celui montré dans l'exemple 6, qui n'est que potentiel du fait de l'emploi du conditionnel ;

3. les cas où le type des arguments est sous-spécifié, comme le montre l'exemple 7. Il faut noter que tout type d'achat peut être repéré, même s'il n'est pas accompli par une entreprise au sens propre.
4. les cas où la relation prédicative est absente de la phrase, comme le montre l'exemple 8, qui illustre un cas dans lequel le système produit une erreur malgré les contraintes posées en amont.

Ces deux derniers cas de figure confortent l'idée selon laquelle une validation par le linguiste est indispensable pour produire des connaissances de qualité. Malgré les limites que nous venons d'évoquer, il semble que concrètement, cette démarche présente l'avantage d'accélérer le processus d'exploration du corpus pour identifier des instances de relations pertinentes. C'est donc une méthode qui nous semble particulièrement intéressante en regard de notre problématique générale d'extraction de relations à partir de corpus. Elle facilite en effet le travail du linguiste, en lui permettant de concentrer son attention sur un nombre limité de cooccurrences candidates susceptibles de contenir les relations que l'on cherche à extraire. Cette expérimentation nous dévoile également une piste productive pour la poursuite de nos travaux de recherche. L'une des remarques que l'on pourrait adresser à la méthode proposée est qu'introduire une contrainte ciblant spécifiquement la présence de certaines lexicalisations prédicatives à détecter dans les phrases du corpus restreint de fait, obligatoirement les résultats. En pratique, si la liste de lexicalisations fournie au système n'est pas suffisamment étendue, cela peut induire du silence, en ne détectant pas certaines formulations pourtant pertinentes. Une piste intéressante pour la suite serait de s'inspirer du co-apprentissage itératif proposé par Hearst (1992) Il s'agirait :

1. à partir d'un ensemble de phrases thématisant la relation recherchée, d'extraire les entités nommées cooccurrentes qui s'y trouvent ;
2. ensuite, une procédure permettrait d'extraire les éléments porteurs de la relation ;
3. après les avoir soumises au linguiste, le sous-ensemble des lexicalisations prédicatives validées serviraient d'entrée pour une nouvelle itération de la recherche de cooccurrences d'entités dans le corpus.

Une telle démarche pourrait, à terme, révéler de nouvelles instances de relations pertinentes entre entités nommées.

8.4 Déterminer des règles pour un typage structurel des relations

Cette étape prend en entrée les phrases extraites du corpus à l'issue de l'étape précédente. Le but est ensuite de procéder à un typage structurel indiquant la structure employée pour exprimer la relation prédicative. Cela va permettre au système de produire un niveau supplémentaire de la grammaire – le deuxième niveau de profondeur évoqué plus haut, lequel est fonction de la position des arguments et du prédicat dans la phrase.

A l'issue de cette étape, chacune configuration de positions identifiée donnera lieu à l'écriture d'une sous-grammaire particulière. Avant de poursuivre, notons que les connaissances linguistiques jouent ici un rôle de premier plan. La position du prédicat – qu'il soit verbal ou nominal – par rapport aux arguments qu'il lie, peut varier, d'une langue à l'autre, mais aussi dans une même langue [10]. Nous donnons dans le tableau 8.6 des exemples de configuration variées de relations entre entités nommées pour le français. On voit que plusieurs positions sont possibles pour le prédicat, qu'il soit verbal ou nominal. Le prédicat peut-être encadré par ses arguments, antéposé ou postposé à ceux-ci. Cette connaissance de la langue et la capacité à la modéliser relèvent des compétences du linguiste qui intervient dans la procédure de modélisation des grammaires locales.

Dans l'expérience menée, nous nous intéressons spécifiquement aux occurrences de la relation d'achat, dont les arguments sont des entités nommées de type « Organisation ». Rappelons que nous prenons en compte les cooccurrences d'entités dont le score est médian. ces cas semblent être les plus propices pour une étude détaillée réalisée par le linguiste. Ainsi, nous avons choisi de sélectionner, dans les résultats obtenus à partir du corpus C17, le sous-ensemble de phrases présentant des cooccurrences d'entités dont le score s'étend de 3 à 7 (c'est-à-dire que chaque cooccurrence se trouve dans trois à sept phrases différentes dans le corpus). Ce sous-ensemble contient 431 phrases, parmi lesquelles 175 présentent exactement la relation prédicative re-

10. En japonais par exemple, la prédicat est en fin de phrase.

TABLE 8.6 – Les différents types de positions du prédicat et des arguments qu'il lie, identifiées dans un corpus français

Positions possibles du prédicat et de ses arguments	**Exemple extrait du corpus analysé**
[Argument1 Prédicat-verbal Argument2]	Partygaming a racheté en 2007 le groupe spécialisé dans le poker Empire Online.
[Prédicat-verbal Argument1 Argument2]	Racheté par Betclic en 2008, Bet-at-home (335 millions d'euros), [...]
[Argument1 Argument2 Prédicat-verbal]	L'ex-procureur de Nanterre Philippe Courroye et l'ex-président Nicolas Sarkozy se sont rencontrés au moins huit fois à des dates proches de moments-clés de l'affaire Bettencourt.
[Argument1 Prédicat-nominal Argument2]	Fin avril, Safran a annoncé le rachat d'une participation de 81capital de GE Homeland Protection.
[Prédicat-nominal Argument1 Argument2]	Le rachat de l'américain Terminal en 1996 aura couté la bagatelle de 20 millions de dollars au géant de Redmond, Microsoft.
[Argument 1 Argument 2 Prédicat-nominal]	Sarkozy – Fillon : les rencontres non officielles

cherchée – celle de l'achat d'une organisation par une autre – ce qui représente 40,8% du sous-ensemble considéré. Ces 175 phrases sont analysées manuellement, afin de typer la nature du prédicat et d'en identifier les arguments. Nous obtenons finalement 121 phrases dont le prédicat est verbal et 54 dont le prédicat est nominal.

La relation d'achat induit une asymétrie dans la relation de ses arguments : autrement dit, l'une des entités de la relation en est l'agent, quand l'autre entité la subit. En conséquence, on s'attend en français à trouver des réalisations du type « X a racheté Y », « Y a été racheté par X » ou « le rachat de X par Y » ; mais par contre, on ne s'attend pas à rencontrer des cas tels que « X et Y se sont rachetés » : cette dernière formulation peut être qualifiée de non-acceptable. Ainsi, la voix verbale – active ou passive – a une incidence sur la position des arguments par rapport au prédicat. Nous notons donc qu'il est important à ce niveau de distinguer agent et patient, c'est-à-dire d'introduire l'annotation des rôles sémantiques.

Ces informations sur les rôles sémantiques et le prédicats sont fournies au module prenant en charge la catégorisation des occurrences de relations. A l'issue de cette étape, le programme type structurellement les occurrences de relations traitées. Nous présentons ci-après la répartition des 175 phrases analysées, en fonction des catégories en question. L'examen manuel de ces résultats confirme la pertinence des contraintes linguistiques fournies au programme : en effet, aucune mauvaise catégorisation n'est constatée à ce stade. On note que le type du corpus a une incidence sur la fréquence des types de catégories identifiés : dans les données provenant d'articles de presse, la voix active est la plus fréquente comme le montre le tableau 8.7

TABLE 8.7 – Répartition des phrases analysées en fonction des catégories identifiées par le programme informatique mis en place – indication du rôle sémantique des arguments dans la relation d'achat

Catégories identifiés – indication du rôle sémantique des arguments	**Fréquence**
[Agent Prédicat-verbal Patient]	72
[Patient Prédicat-verbal Agent]	23
[Prédicat-verbal Agent Patient]	19
[Agent Prédicat-nominal Patient]	41
[Prédicat-nominal Patient Agent]	17
[Prédicat-nominal Agent Patient]	3

Cette étape permet donc au linguiste, à partir des résultats fournis par le programme, de pouvoir développer un niveau ultérieur de la grammaire, le niveau 2 (figure 8.4). Les connaissances mobilisées par le système d'extraction de relations entre entités nommées, s'adosse ainsi à des connaissances fournies par le linguiste, tout en permettant à celui-ci de gagner un temps précieux pour la mise en place de la grammaire.

A la lumière de ces résultats, nous pensons que des développements ultérieurs pourraient, en toute logique, permettre d'établir une typologie étendue des occurrences de relations entre entités nommées, selon leur type et en fonction des catégories dans lesquelles elles sont thématisées et des rôles sémantiques des arguments qu'elles impliquent. Il est probable que différentes relations soient instanciées par des catégories similaires. A cet égard, des ressources telles que FrameNet pour l'anglais seraient vraisemblablement en mesure de fournir une base exploitable, ces conjectures appelant nécessairement d'être confirmées ou infirmées par des travaux applicatifs ultérieurs.

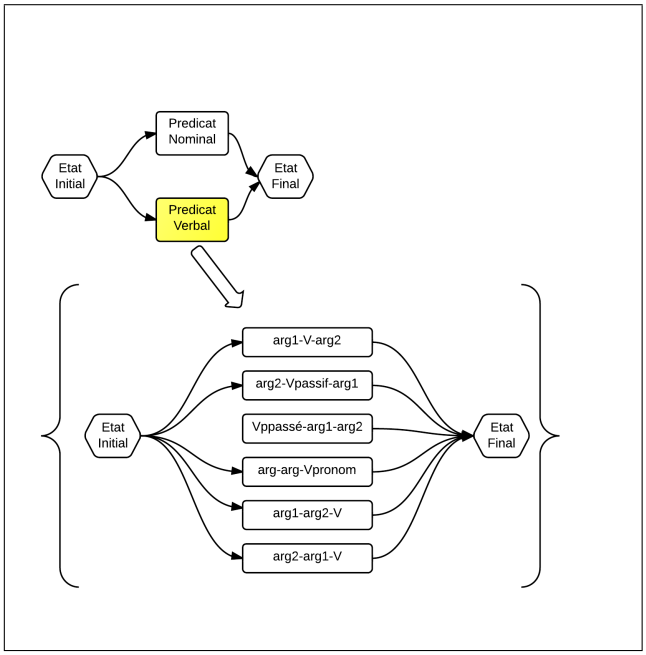


FIGURE 8.3 – Schéma présentant le principe de l'organisation de la grammaire de niveau 2, en fonction de la position du prédicat et de ses arguments

8.5 Représenter les insertions entre le prédicat et ses arguments

L'étape précédente a permis d'identifier différentes catégories de positions, établies à partir des configurations possibles entre un prédicat et les arguments qu'il lie, dans les occurrences de relations entre entités nommées identifiées en corpus. Plutôt que de proposer un descriptif général, il nous semble plus clair de nous concentrer sur l'une d'entre elles – la plus fréquente, [Agent Prédicat-verbal Patient], qui correspond à un énoncé à la voix active – afin de poursuivre l'explication de la démarche méthodologique proposée pour l'élaboration de grammaires locales en cascade. En évoquant les configurations de position présentées plus haut, telles que [Agent Prédicat-verbal Patient], nous avons volontairement réduit la complexité donnée à voir dans les textes. Concrètement, en corpus, l'on se trouve confronté à des segments textuels positionnés entre les arguments et le prédicat – différents types de propositions, comme les subordonnées relatives par exemple – que nous appellerons insertions. Ces insertions constituent, selon nous, la plus grande difficulté pour la représentation des connaissances à des fins d'extraction de relations entre entités nommées. En effet, leur modélisation est complexe et

coûteuse à mettre en œuvre pour le linguiste, a fortiori lorsque le système n'inclut pas de module d'analyse syntaxique profonde, comme c'est ici le cas pour les raisons évoquées dans le chapitre 6.

Pour pallier ce problème, nous proposons de passer par une classification automatique des insertions, fondée sur les propriétés morphosyntaxiques des segments textuels qui les constituent, afin de faire gagner du temps au linguiste dans sa tâche d'élaboration de la grammaire. Celui-ci a en effet la possibilité de valider, modifier et ajouter de l'information à chacune des représentations candidates qui lui sont indiquées par le système. L'hypothèse sous-jacente est la suivante : la classification donne une vue globale au linguiste. Il peut donc s'appuyer sur cela pour produire une description plus cohérente des différentes constructions ce qui confère une meilleure lisibilité à la grammaire en cours de construction, par rapport à une description réalisée intégralement à la main.

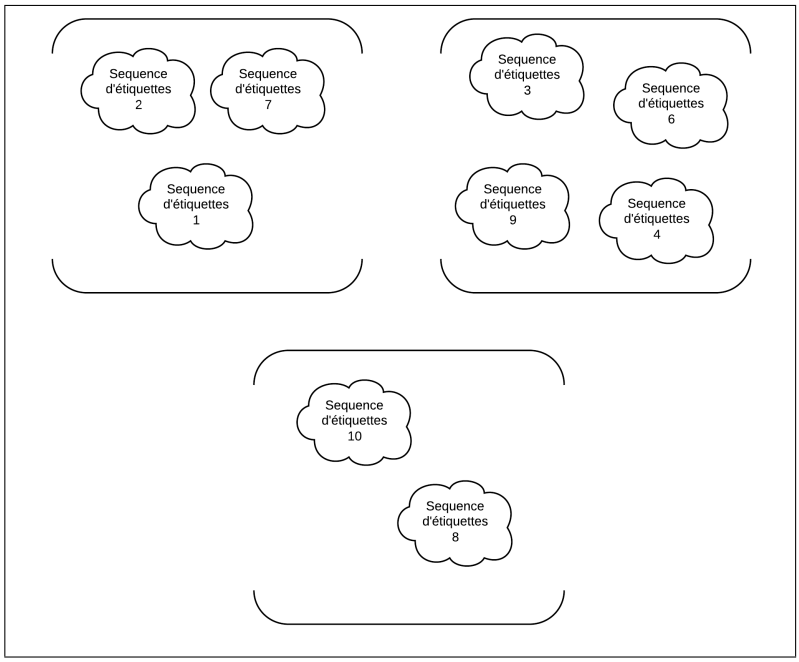


FIGURE 8.4 – Classification des séquences d'étiquettes

Précisons que l'algorithme de classification mis en œuvre est un dérivé du calcul des k-moyennes (MacQuenn, 1967), appelé k-medoïdes, lequel s'appuie sur une mesure de distance, nommée la distance de Levenshtein. Il convient, avant de poursuivre, d'expliquer le fonctionnement de chacune de ces mesures.

La Distance de Levenshtein, aussi appelée distance d'édition, est utilisée pour mesurer la similarité entre deux chaînes de caractères. Cette mesure correspond au nombre minimal de caractères qu'il est nécessaire de supprimer, insérer ou remplacer pour passer d'une chaîne à l'autre ; plus le nombre d'opérations nécessaires est important, plus la mesure de distance est élevée. Plus précisément, à chacune de ces opérations est associé un coût, toujours égal à 1. Issue de la programmation dynamique, cette mesure utilise une matrice de dimension (n+1) x (m+1), où n et m représentent les longueurs respectives de deux chaînes N et M comparées. Concrètement :

- Soit D la Distance de Levenshtein ;
- Si N = « matin » et M = « matin », alors D = 0, car aucune opération n'est nécessaire pour passer d'une chaîne à l'autre ;
- Si N = « matin » et M = « marin », alors D = 1, car une opération de remplacement – du caractère « t » par le caractère « r » – est nécessaire pour passer d'une chaîne à l'autre.

Dans le cadre de notre expérience, nous utilisons ce calcul afin d'obtenir une matrice de distance permettant d'alimenter un algorithme de classification. Pour pouvoir effectuer une classification automatique sur la matrice de distance obtenue avec le calcul de la Distance de Levenshtein, nous exploitons l'algorithme des k-médoïdes (Kaufman et Rousseuuw, 1987). Comme les k-moyennes, cet algorithme itératif permet de partitionner un ensemble d'éléments en plusieurs classes. Il tire son nom de la notion statistique de médoïde, qui réfère au représentant le plus central d'une classe donnée et est défini par une valeur arbitraire de distance par rapport à tous les autres représentants de cette classe. Cette méthode est réputée plus robuste de celle des k-moyennes, du fait de sa capacité à opérer sur des matrices de distance arbitraires entre représentants d'une classe, plutôt que dans un espace vectoriel. Les éléments de classe sont ici représentés par les séquences d'étiquettes morphosyntaxiques, obtenues par ailleurs dans le cadre de l'expérience. Le principe est le suivant :

- Soit k [11] le nombre de représentants – les médoïdes de départ – choisis aléatoirement parmi un ensemble d'éléments ;
- Soit i un nombre d'itérations donné ;

11. Il faut noter que k ne représente pas nécessairement le nombre de classes créées à la fin des itérations. En effet, n'étant pas dans un espace vectoriel, l'algorithme peut confondre le même représentant en un seul médoïde pour effectuer le partitionnement, ce qui résultera en une fusion de ces deux classes dans l'itération courante de l'algorithme.

– L'algorithme associe tous les éléments au médoïde le plus proche, créant ainsi un premier partitionnement ; ici, la proximité utilisée pour déterminer le médoïde le plus proche est la distance de Levenshtein ;
– l'algorithme calcule ensuite un coût, déterminé par la somme des distances entre le médoïde courant et tous les autres éléments, pour chaque élément de l'ensemble ;
– l'algorithme sélectionne un nouveau médoïde, si et seulement si un médoïde dont le coût de distance est plus bas que le médoïde courant, est identifié ;
– l'agorithme itère les deux précédentes étapes, soit jusqu'à stabilisation, soit jusqu'à i fois.

Le résultat est matérialisé par un ensemble de classes d'étiquettes morphosyntaxiques, réputées proches en fonction de la Distance de Levenshtein. Par exemple, les insertions « qui est rentré en bourse » et « qui est rentré hier en bourse » auront une distance de Levensthein réduite et seront regroupées ensemble dans le résultat. On espère ainsi obtenir des classes pertinentes à soumettre au linguiste, pour faciliter l'observation et la modélisation des séquences d'étiquettes morphosyntaxiques obtenues à l'issue de l'analyse des insertions considérées par l'algorithme.

Le choix de s'appuyer sur ces techniques de classification et de calcul de distance, loin d'être choisies au hasard, est motivé par le fait que ces techniques :

– ne sous-tendent pas une représentation du texte comme sac de mots ; au contraire, le texte est conçu comme une séquence de mots ;
– et donnent accès à des données utiles pour la génération de la grammaire locale.

Revenons-en à la configuration la plus fréquente dans notre corpus entre un prédicat et ses arguments, à savoir [Agent Prédicat-verbal Patient], pour expliquer la façon dont agit l'algorithme proposé. Le contenu textuel de chaque insertion identifiée dans cette configuration, est analysé et représenté sous la forme d'une séquence d'étiquettes morphosyntaxiques ; nous avons, pour cette étape, utilisé l'analyseur morphosyntaxique Arisem. Ces séquences d'étiquettes sont données en entrée de l'algorithme de classification, dont chaque classe produite en sortie représente une branche dérivée de la grammaire locale en cascade. En l'occurrence, l'algorithme fait émerger 7 classes à partir des 72 phrases appartenant à la catégorie [Agent Prédicat-verbal Patient] comme le montre le tableau 8.8.

TABLE 8.8 – Répartition des insertions analysées au terme de la procédure de classification des séquences d'étiquettes morphosyntaxiques qui les composent

Identifiant de la classe	**Nombre de séquences regroupées dans la classe**
Classe 1	6
Classe 2	6
Classe 3	17
Classe 4	7
Classe 5	14
Classe 6	18
Classe 7	4

TABLE 8.9 – Répartition des insertions analysées au terme de la procédure de classification des séquences d'étiquettes morphosyntaxiques qui les composent

Classe	**Exemple de séquences**	**Commentaire**
6	a) PUN PRO-REL VER-pres ADV b) PUN PRO-REL VER-imp ADV c) PUN PRO-REL VER-imp6	Les séquences d'étiquettes correspondent à des subordonnées relatives (exemple a : « , qui a récemment »)
7	a) PUN VER-pres VER-pper PUN NOM NUM NOM PUN b) PUN VER-pres PRO-PER PUN VER-pper c) PUN VER-pres VER-pper PUN NOM PUN	Les séquences d'étiquettes correspondent à des incises, dont les fonctions sont variées (exemple a : « , a annoncé, vendredi 13 avril, »)

Outre la disparité de population entre les différentes classes que fait émerger l'algorithme, qui ne semble pas, selon nous, être en elle-même un phénomène interprétable, il est intéressant de remarquer que ces résultats, bien que difficilement généralisables en l'état, correspondent cependant a des séquences linguistiques cohérentes, comme des propositions relatives ou des incises, comme on le voit dans le tableau 8.9.

A cette étape, les connaissances acquises à partir du corpus, bien qu'elles demanderaient des travaux de description linguistiques complémentaires, permettent de disposer d'informations pour permettre de développer le niveau 3 de la grammaire.

Le résultat idéal pour cette procédure serait que l'algorithme regroupe

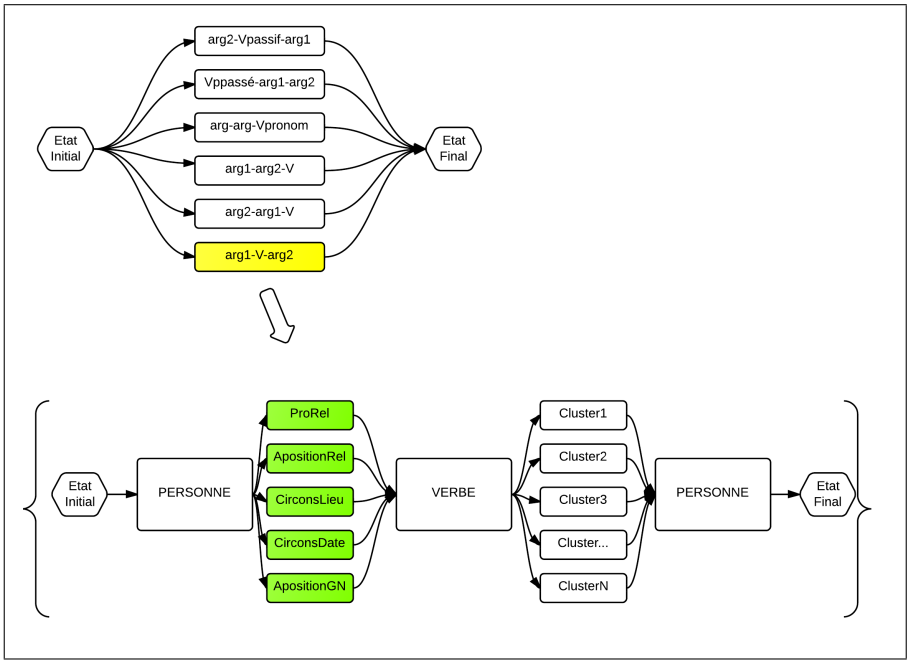


FIGURE 8.5 – Schéma présentant le principe de l'organisation de la grammaire de niveau 3, en fonction des séquences d'étiquettes morphosyntaxiques constituant les insertions identifiées

dans une même classe des phénomènes linguistiques de même nature. Cependant, cela demanderait à être évalué sur des échantillons plus conséquents. Néanmoins, ces résultats fournissent des pistes concrètes pour avancer dans l'élaboration de la grammaire, à partir de données endogènes au corpus : c'est, selon nous, un résultat encourageant pour poursuivre cette piste de recherche dans le cadre de travaux ultérieurs. En ce sens, différentes pistes de développement sont, selon nous, immédiatement identifiables. La première est liée à la lisibilité des résultats produits, qui pourraient être améliorée par un étiquetage linguistiquement fiable des classes que l'algorithme fait émerger. En effet, en l'état, il est nécessaire de passer par une observation fine du contenu de chaque classe pour identifier précisément les phénomènes linguistiques auxquels elle correspond. L'une des façons de parvenir à les étiqueter de façon consistante serait de mener une campagne impliquant l'intervention de plusieurs experts linguistes, à qui l'on proposerait de définir une étiquette pour chaque classe. La seconde piste d'amélioration tient à l'évaluation, sur un corpus à plus grande échelle, de la robustesse de l'algorithme utilisé, sensible à la longueur des séquences textuelles qui forment les insertions.

8.6 Génération et factorisation des symboles terminaux de la grammaire

Nous arrivons à la dernière étape de la procédure, c'est-à-dire la génération des symboles terminaux de la grammaire. Pour chaque classe issue de l'étape précédente, une branche de la grammaire est crée. Et pour chacune de ces branches, nous allons générer les symboles terminaux qui reconnaissent les insertions dans les textes. Pour ce faire, on exploite les données fournies par le calcul de la distance de Levenshtein (préalablement réalisée dans l'étape précédente) pour identifier les opérations précises à faire à chaque étape de la génération de la grammaire. Chaque opération d'édition de la distance de Levenshtein implique une opération de génération précise de symboles terminaux. Nous les détaillons ci-après :

- Soient une séquence d'étiquettes A et une séquence d'étiquettes B ;
- Si l'opération de substitution est nécessaire pour passer de A à B, alors l'algorithme opère une union des symboles dans la grammaire ;
- Si l'opération d'ajout est nécessaire pour passer de A à B, alors l'algorithme opère une bifurcation à partir du symbole ajouté dans la grammaire ;
- Si l'opération de suppression est nécessaire pour passer de A à B, alors l'algorithme opère un ajout de l'étiquette présent dans B, dans la grammaire ;
- Si aucune opération n'est nécessaire pour passer de A à B, alors l'algorithme fusionne l'étiquette en un seul état de la grammaire – il procède donc à une factorisation.

Nous conservons donc la « mémoire de Levenshtein » utilisée pour le calcul de distance nécessaire à la classification par k-médoïde. C'est-à-dire que dans une classe de séquences d'étiquette données, nous avons conservé l'ensemble de la suite des opérations effectuées pour passer d'une séquence à l'autre. Nous illustrons un cas d'analyse de deux séquences d'étiquettes complet dans la figure 8.6.

Ce principe est appliqué à toutes les étiquettes de l'ensemble en entrée de l'algorithme [12]. A terme, la grammaire générée est donc supposée plus lisible ;

12. Il faut souligner un problème inhérent à l'algorithme : si celui-ci opère deux unions à partir de la même étiquette, vers une même étiquette, alors l'algorithme produit une bifurcation vers deux états différents portant la même étiquette, rendant la grammaire non déterministe. Pour pallier ce problème, on a recours à un algorithme destiné à fusionner les deux étiquettes identiques en une seule, rendant ainsi la grammaire finalement produite déterministe.

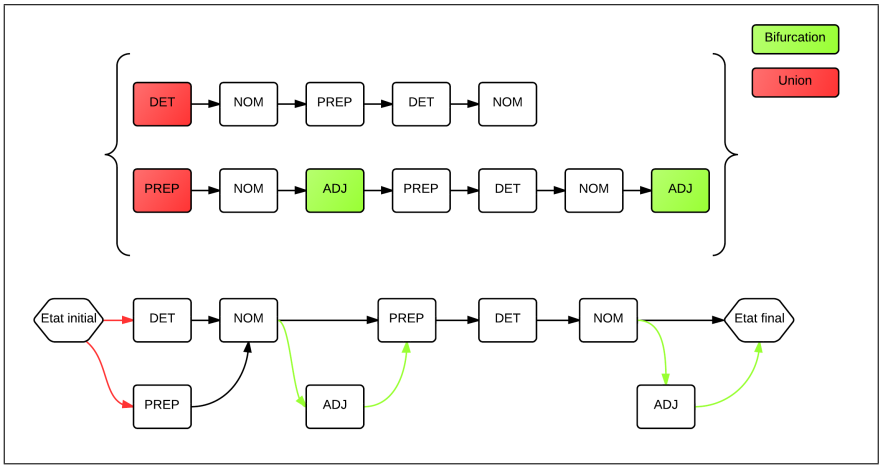


FIGURE 8.6 – Exemple de génération entre deux séquences d'étiquettes - les opérations de substitution sont marquées en rouge, les ajouts sont marqués en vert

de plus, elle correspond à l'ordre syntagmatique du langage naturel.

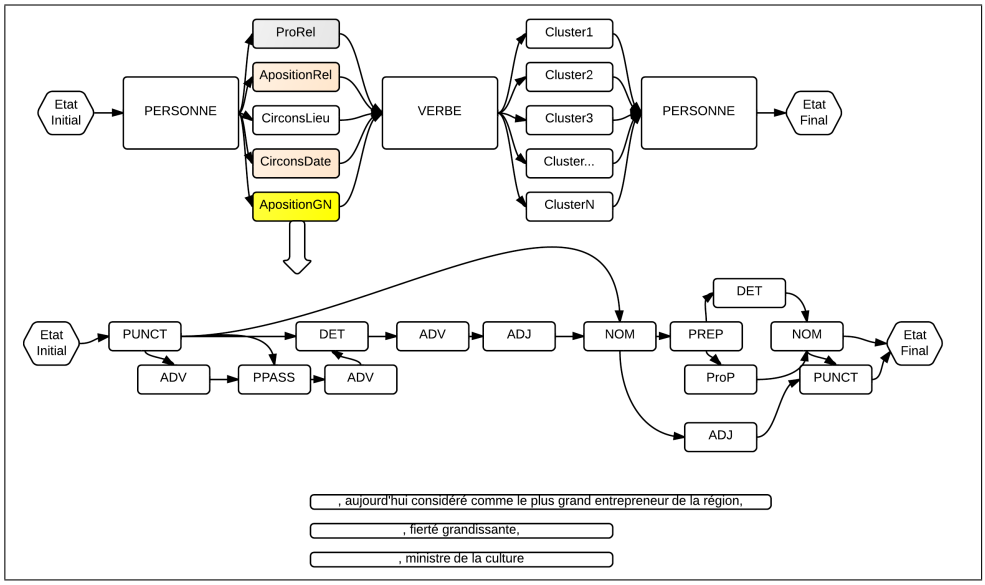


FIGURE 8.7 – Schéma présentant le principe de l'organisation de la grammaire de niveau 4, en fonction du résultat de la classification par les k-médoïdes

8.7 Evaluation du cadre méthodologique proposé

Au cours de ce chapitre, nous avons présenté les différentes étapes d'un cadre méthodologique pour l'organisation des grammaires locales en cascade, en mettant au cœur de notre réflexion les principes de « réusinage », de lisibilité et de maintenabilité des ressources linguistiques. Autrement dit, à partir des problématiques auxquelles nous nous sommes retrouvé confronté sur le terrain, nous avons cherché une réponse opérationnelle. En effet, les ressources linguistiques telles que les grammaires locales, sont aujourd'hui encore largement, en contexte industriel, constituées manuellement. Outre le fait que cela représente un coût important, la constitution manuelle des ressources linguistiques pose le problème de la maintenabilité.

Afin d'estimer l'apport de la méthode proposée, nous procédons maintenant à son évaluation, sous deux aspects : d'une part, en termes de précision et de rappel par rapport à une grammaire élaborée manuellement ; d'autre part, en termes de maintenabilité et de lisibilité, en proposant à trois linguistes de corriger des phrases présentant des erreurs et en chronométrant le temps qu'ils ont mis pour le faire. Nous avons ainsi souhaité utiliser des critères qui soient les plus objectifs et reproductibles possible.

8.7.1 Résultats de l'évaluation en précision et en rappel

Dans un premier temps, notre objectif est de mesurer les performances de la grammaire générée au terme de la démarche décrite au cours de ce chapitre et ce, sans prendre en compte des facteurs tels que la lisibilité, la maintenabilité ou le temps nécessaire à l'élaborer, volets qui feront l'objet de la seconde étape de l'évaluation décrite dans cette section. Présentons avant tout le protocole utilisé pour évaluer les performances de la grammaire produite, en termes de précision et de rappel. L'évaluation s'appuie sur le corpus d'évaluation établi à partir du corpus C17, présenté au début de ce chapitre, dans lequel 232 occurrences de la relation de type « Achat » impliquant deux entités nommées de type « Organisation » et marquées par un prédicat verbal, ont été annotées manuellement. Nous comparons ici le résultat de la grammaire générée semi-automatiquement, dans le cadre de la procédure proposée dans ce chapitre, avec celle produite manuellement par un linguiste, au terme d'un travail approfondi fondé sur son expertise du domaine. Avant d'en venir aux résultats, rappelons brièvement en quoi

consistent les métriques utilisées :

- L'évaluation en rappel, classiquement, se fonde sur le rapport entre le nombre de relations correctement détectées par la grammaire et le nombre total relations annotées ; le score de rappel résultant est exprimé en pourcentage. Le silence est le complémentaire du rappel.
- La précision, elle, se fonde sur le nombre de relations correctement détectées par la grammaire par rapport au nombre total d'occurrences détectées par la grammaire. On parle de bruit généré par la grammaire si celle-ci détecte des occurrences qui ne correspondent pas à la relation recherchée.

Type de grammaire	***Rappel***	***Précision***
Grammaire générée semi-automatiquement	45.6%	96.2%
Grammaire élaboré manuellement	47.2%	83.3%

TABLE 8.10 – *Estimation du rappel et de la précision*

En termes de rappel, les deux grammaires ont des performances similaires et produisent toutes deux un silence important : Plus de la moitié des occurrences ne sont pas détectées par la grammaire, mise au point manuellement ou automatiquement. Il faut souligner que l'analyseur Arisem utilisé en amont de l'application des grammaires, ne résout pas les coréférences d'entités nommées, ce qui a un impact négatif sur la performance.

En termes de précision, la grammaire élaborée manuellement obtient 12,9 points de moins que celle générée semi-automatiquement. Cela est directement imputable au fait que la grammaire élaborée par le linguiste autorise des retours en arrière : si ces derniers permettent de factoriser certaines parties de la grammaire, ils augmentent par contre la combinatoire des segments reconnus par la grammaire, d'où la perte relativement importante du point de vue de la précision. Sur ce point, il faut préciser que, dans le cadre des applications industrielles visées – en particulier le peuplement d'ontologies, les moteurs de recherche et la veille d'information – la présence de retours en arrière dans les grammaires exploitées génèrent un silence qui se répercute sur les étapes suivantes s'appuiyant sur le résulat de l'extraction de relations. C'est pourquoi, en contexte industriel, la précision est généralement favorisée sur le rappel dans l'évaluation des systèmes automatisés mis en place. En outre, pour anecdotique qu'il soit, le paramètre du coût en temps passé a aussi une incidence d'un point de vue industriel ; sur ce point, il est intéressant de noter que le linguiste a passé 10 jours pour mettre en place une version finalisée

de la grammaire, quand la grammaire générée semi-automatiquement a mis 1 jour à être établie pour un usage opérationnel.

8.7.2 Evaluation de la maintenabilité et de la lisibilité

Les résultats, plutôt encourageants, obtenus par la grammaire générée semi-automatiquement en termes de précision, ne permettent pas d'en estimer la lisibilité ou la maintenabilité. Afin d'évaluer en situation cet aspect, nous avons appliqué le protocole suivant :

- Trois phrases dont l'analyse est erronée par les deux grammaires précédemment évaluées sont sélectionnées – deux une erreur de rappel, donc du silence et une produisant une erreur de précision, donc du bruit ;
- Chacune des phrases est proposée à trois linguistes différents ;
- Il leur est demandé de procéder, s'ils l'estiment possible, à la correction de chaque erreur identifiée ;
- S'ils choisissent de corriger l'erreur, alors le temps qu'ils mettent est chronométré à la seconde près ;
- S'ils choisissent de ne pas corriger l'erreur, alors on indique un cas d'abandon, en leur ayant préalablement précisé qu'il est possible de décider de ne pas corriger l'erreur avant le début de l'expérimentation s'ils estiment que la correction prendra beaucoup de temps.

La tâche a été proposée à chaque évaluateur séparément, afin qu'ils ne puissent pas discuter entre eux lors de la réalisation de celle-ci. Voici les trois phrases qui leur ont été soumises :

- Erreur 1 - « PTC a également profité de sa conférence utilisateurs pour annoncer le rachat de Relex. » (cas de silence : la phrase aurait dû être détectée par les grammaires)
- Erreur 2 - « Lufthansa soumet une nouvelle offre rapide pour le rachat d'Austrian. » (cas de silence : la phrase aurait dû être détectée par les grammaires)
- Erreur 3 - « Anheuser-Busch profite de ce rachat pour attaquer son conccurrent Inbev. » (cas de bruit : la phrase n'aurait pas dû être détectées par les grammaires)

Le tableau 8.11 synthétise le résultat de l'evaluation. On observe que le temps passé à corriger la grammaire est systématiquement plus important dans le cas de celle produite manuellement que pour la contrepartie générée semi-automatiquement. Toutes les modifications apportées par les évaluateurs sur la grammaire générée semi-automatiquement ont permis de

TABLE 8.11 – Résultats de l'évaluation du temps mis pour la révision d'une grammaire par trois évaluateurs linguistes

	Grammaire générée semi-automatiquement	**Grammaire élaboré manuellement**
Linguiste A	4m56s correcte 3m42s correcte 3m15s correcte	10m30s incorrecte 7m31s incorrecte 7m56s incorrecte
Linguiste B	4m39s correcte 3m38s correcte 4m00s correcte	8m56s correcte 9m23s correcte 8m32s incorrecte
Linguiste C	3m46s correcte 4m16s correcte 2m34s correcte	7m18s abandon 8m48s incorrecte 6m35s abandon

corriger les erreurs. L'observation des statuts des corrections apportées sur la grammaire produite manuellement montre que les abandons et les cas d'échec de corrections sont plus fréquents. Cela indique que la grammaire élaborée manuellement est moins lisible et moins facile à maintenir. Ce qui est, en soi, un résultat encourageant quant à la démarche que nous avons proposée dans ce chapitre, qui visait à produire une grammaire lisible et facilement modifiable si nécessaire.

8.8 Bilan du chapitre

Nous avons, au cours de ce chapitre, esquissé un cadre méthodologique pour l'organisation des grammaires locales en cascade, dans le but de proposer des éléments de réponse à différents problèmes identifiés pour les besoins de l'entreprise où nous avons fait cette thèse. En premier lieu, nous avons cherché à proposer une démarche permettant de faciliter la tâche du linguiste, qui a la charge d'élaborer et de maintenir des ressources au fil du temps. En second lieu, nous nous sommes fixé l'objectif de pouvoir produire, de façon semi-automatisée, des ressources linguistiques lisibles par l'humain et opérationnelles, c'est-à-dire qu'elles peuvent être facilement intégrées dans une chaîne de traitement industrielle. Au terme d'une double évaluation, en termes de performance, d'une part, en termes de lisibilité et de maintenabilité, d'autre part, les résultats obtenus sont encourageants et ouvrent des perspectives qui semblent pertinentes pour la poursuite de nos travaux.

Conclusion générale

Résumé et synthèse

Cette thèse a permis de présenter un vaste panorama de recherche autour de la notion de relation entre entités. Notre démarche fut double :

- proposer une définition de la notion de relation en présentant son évolution à travers diverses conférences, tout en pointant les difficultés qui ne sont aujourd'hui pas encore traitées dans la communauté de recherche.
- Elaborer un système semi-automatique pour l'apprentissage de règles d'extraction de relations basées sur une collaboration entre un processus informatique et un ou plusieurs linguistes.

Ainsi, nous avons scindé notre exposé en deux parties. La première pose la question du manque de ponts avec la linguistique dans les communautés d'extraction d'informations. A l'instar des entités nommées, nous avons montré comment les relations ont pu émerger à partir de diverses conférences guidées par les applications. La seconde partie de notre exposé est pragmatique. Etant donné un contexte industriel, nous nous sommes attelé à réaliser un système d'apprentissage de règles d'extraction de relations qui utilise les connaissances apportés par un linguiste car nous pensons qu'une telle collaboration est prometteuse.

Dans un premier temps, nous avons rapproché la notion de relation des prédicats chez Tesnière pour ancrer nos travaux sur une base théorique solide. Puis nous avons examiné, à travers un historique des conférences, comment cet objet a évolué jusqu'à maintenant, Ce qui nous a permis de mettre en exergue les difficultés, notamment du point de vue de l'annotation manuelle, lorsqu'il s'agit d'identifier de manière précise les mentions de relations dans les textes. Ces difficultés relèvent de deux ordres :

– L'ingénierie des connaissances, qui doit être capable de fournir des consignes claires quant à la définition des classes des relations entre entités nommées
– La linguistique, qui montre que la manifestation des relations dans les textes est sujette à des phénomènes rarement pris en compte en TAL, comme la modalité.

Nous nous sommes par ailleurs attachés à l'annotation manuelle de la modalité : nous avons ainsi mené une expérience montrant la validité du modèle, mesuré grâce à un taux d'accord inter-annotateur important.

Dans un second temps, nous avons mis au point un système d'acquisition de règles à partir de corpus, pour l'extraction de relations entre entités nommées dans les textes. Après avoir présenté le contexte industriel dans lequel nos travaux se sont déroulés, nous avons décrit des expériences permettant de constituer des ressources ontologiques nécessaires à la détection des entités nommées. Nous avons ensuite présenté une chaîne de traitements permettant de générer des règles d'extraction de relations à partir d'un corpus. Notre méthode s'appuie sur la présence du linguiste, non seulement comme validateur, mais aussi comme fournisseur de connaissances. L'autre objectif de notre approche est d'offrir une capacité de révision accrue des ressources générées : les systèmes produisent en général des résultats difficiles à maintenir et il est crucial d'avoir une réelle capacité de révision, notamment pour le linguiste chargé du développement et de la maintenance des ressources.

Nous avons évalué nos résultats avec les mesures classiques de rappel et de précision et obtenu des résultats très encourageants. Nous avons aussi mené une évaluation portant sur la maintenabilité des ressources générés, en confrontant nos résultats avec des avis d'experts du domaine, en particulier des linguistes. Nous avons pu montrer que notre méthode a une valeur ajoutée, notamment sur le plan des possibilités de révision des grammaires obtenues quand on les compare avec des grammaires ayant été développées manuellement.

Améliorations et perspectives

Dans cette thèse, nous avons tenté de répondre à certaines problématiques liées aux relations entre entités nommées. Mais nous considérons que le sujet est loin d'être résolu aujourd'hui, même si des chercheurs se penchent sur la question depuis une vingtaine d'années.

Il n'existe actuellement aucun standard sur les relations, que cela soit au niveau des classes sémantiques ou des notions qu'elles recouvrent. Le constat est très différent pour les entités nommées. Par ailleurs, certains phénomènes comma la modalité sont encore peu souvent traités par les systèmes d'extraction d'information. Au niveau des ontologies, nous avons décrit une modélisation formelle des relations dans les langages du web sémantique, mais il n'existe pas encore de ressources faisant consensus pour exploiter une telle représentation.

Notre méthode permet d'obtenir des résultats encourageants au niveau des relations, pour alimenter des systèmes d'extraction d'information par exemple, mais les performances en termes de rappel peuvent être améliorées. Les concepteurs de systèmes commencent à s'intéresser à la résolution des anaphores, ce qui peut contribuer à une amélioration de la qualité des résultats (jusqu'à 30 points de gagner d'après certaines études). Cette tâche offre des pistes qui pourraient être exploitées pour améliorer le rappel des systèmes d'extraction d'informations.

Enfin, pour conclure, nous pensons que les recherches sur la notion de relation gagneraient à mieux définir les objets manipulés (en traitement automatique en particulier) et qu'elles pourraient aussi s'appuyer davantage sur les travaux des linguistes dans les chaînes de traitement.

Bibliographie

Dbpedia. http ://dbpedia.org/About.

Geonames. http ://www.geonames.org/.

Unitex. http ://www-igm.univ-mlv.fr/ unitex/.

Proceedings of the 6th message understanding conference, 1995.

Polarité, négation et scalarité. Langages 162, 2006.

Proposed task description for knowledge-base population at tac 2011, 2011.

A. Abeillé, L. Clément, et F. Toussenel. Building a french treebank. *Treebanks*, pages 165–187, 2003.

E. Agichtein. *Extracting relations from large text collections.* PhD thesis, Columbia University, 2005.

E. Agichtein et V. Ganti. Mining reference tables for automatic text segmentation. *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 20–29, 2004.

E. Agichtein et L. Gravano. Snowball : Extracting relations from large plain text collections. *Proceedings of the 5th ACM International Conference on Digital Libraries*, pages 85–94, 2000.

R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, et A. I. Verkamo. Fast discovery of association rules. *Advances in Knowledge Discovery and Data Mining*, pages 307–328, 1996.

J. Aitken. Learning information extraction rules : An inductive logic programming approach. *Proceedings of the 15th European Conference on Artificial Intelligence*, pages 355–359, 2002.

D. E. Appelt et D. Martin. Named entity extraction from speech : Approach and results using the textpro system. *In Proceedings of the DARPA Broadcast News Workshop*, pages 51–54, 1999.

D. E. Appelt, J. R. Hobbs, D. J. Israel, et M. Tyson. Fastus : A finite-state processor for information extraction from real-world text. *IJCAI*, pages 1172–1178, 1993.

M. Arrivé. Les éléments de syntaxe structurale de lucien tesnière. *Langue française*, 1 :36–40, 1969.

M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, et O. Etzioni. Open information extraction from the web. *IJCAI*, pages 2670–2676, 2007.

E. Benveniste. *Problèmes de linguistique générale*, volume 1. Gallimard, 1966.

M. Berland et E. Charniak. Finding parts in very large corpora. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 57–64, 1999.

D. Bernhard et A. Ligozat. Analyse automatique de la modalité et du niveau de certitude : application au domaine médical. *TALN 2011*, pages 433–444, 2011.

S. Bethard, H. Yu, A. Thornton, V. Hatzi-Vassiloglou, et D. Jurafsky. Automatic extraction of opinion propositions and their holders. *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text : Theories and Applications*, pages 22–44, 2004.

D. M. Bikel, S. Miller, R. Schwartz, et R. Weischedel. Nymble : A highperformance learning name-finder. *ANLP-97*, pages 194–201, 1997.

K. Bloom, N. Garg, et S. Argamon. Extracting appraisal expressions. *HLT-NAACL*, pages 308–315, 2007.

A Borillo. Exploration automatisée de textes de spécialité : repérage et identification de la relation lexicale d'hypéronymie. *LINX*, 34/35 :113–124, 1996.

V. R. Borkar, K. Deshmukh, et S. Sarawagi. Automatic text segmentation for extracting structured records. *Proceedings of ACM SIGMOD International Conference on Management of Data*, pages 175–186, 2001.

A. Borthwick, J. Sterling, E. Agichtein, et R. Grishman. Exploiting diverse knowledge sources via maximum entropy in named entity recognition. *Sixth Workshop on Very Large Corpora New Brunswick*, pages 152–160, 1998.

B. Boser, I. Guyon, et V. Vapnik. A training algorithm for optimal margin classifiers. *Fifth Annual Workshop on Computational Learning Theory*, pages 144–152, 1992.

D. Bouchard. Les verbes psychologiques. *Langue française*, 105 :6–16, 1995.

D. Bourigault. *Un logiciel d'Extraction de TERminologie. Application à l'acquisition de connaissances à partir de textes*. PhD thesis, EHESS, 1994.

E. Breck et C. Cardie. Playing the telephone game : Determining the hierarchical structure of perspective and speech expressions. *Proceedings of the 20th international conference on Computational Linguistics (COLING)*, (120), 2004.

S.E. Brennan, M.W. Friedman, et J.C. Pollard. A centering approach to pronouns. *Proceedings of the 25th annual meeting on Association for Computational Linguistics*, pages 155–162, 1987.

R. Bunescu et R. Mooney. Learning to extract relations from the web using minimal supervision. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 576–583, 2007.

R. C. Bunescu, R. Ge, R. J. Kate, E. M. Marcotte, R. J. Mooney, A. K. Ramani, et Y. W. Wong. Comparative experiments on learning information extractors for proteins and their interactions. *Artificial Intelligence in Medicine*, 33 :139–155, 2005.

M. J. Cafarella, D. Downey, S. Soderland, et O. Etzioni. Knowitnow : Fast, scalable information extraction from the web.

M. J. Cafarella, C. Re, D. Suciu, et O. Etzioni. Structured querying of web text data : A technical challenge. *CIDR*, pages 225–234, 2007.

M. Califf et R. Mooney. Bottom-up relational learning of pattern matching rules for information extraction, 2003.

M. Califf et R. J. Mooney. Relational learning of pattern-match rules for information extraction. *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)*, pages 328–334, 1999.

C. Cardie et K. Wagstaff. Noun phrase coreference as clustering. *Proceedings of the Joint SIGDAT Conference*, pages 82–89, 1999.

V.T. Chakaravarthy, H. Gupta, P. Roy, et M.K. Mohania. Efficiently linking text documents with relevant structured information. *VLDB*, pages 667–678, 2006.

S. Chakrabarti. *Mining the Web : Discovering Knowledge from Hypertext Data*. Morgan-Kauffman, 2002.

N. A. Chinchor. Overview of muc-7/met-2, 1998a.

N. A. Chinchor. Muc-7 information extraction task definition, 1998b.

Y. Choi, C. Cardie, E. Riloff, et S. Patwardhan. Identifying sources of opinions with conditional random fields and extraction patterns. *the Human Language Technology Conference/Conference on Empirical Methods in Natural Language Processing*, pages 355–362, 2005.

Y. Choi, E. Breck, et C. Cardie. Joint extraction of entities and relations for opinion recognition. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 431–439, 2006.

F. Ciravegna. Adaptive information extraction from text by rule induction and generalisation. *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI2001)*, 2 :1251–1256, 2001.

J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20 :37–46, 1960.

W. W. Cohen et S. Sarawagi. Exploiting dictionaries in named entity extraction : Combining semi-markov extraction processes and data integration methods. *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 89–98, 2004.

M. Collins et Y. Singer. Unsupervised models for named entity classification. *In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 100–110, 1999.

Linguistic Data Consortium. Automatic content extraction (ace) program, 1998-2008.

Linguistic Data Consortium. Ace annotation guidelines for entity detection and tracking, 2004.

Linguistic Data Consortium. Tac kbp slots, 2011.

F. Corblin. Pronoms et mentions. *Bulletin de la Société de Linguistique de Paris*, CII(1) :285–323, 2007.

F. Cornish. Référence anaphorique, référence déictique, et contexte prédicatif et énonciatif. *Sémiotiques*, 8 :31–55, 1995.

C. Cumby et D. Roth. Feature extraction languages for propositionalzed relational learning. *Working Notes of the IJCAI-2003 Workshop on Learning Statistical Models from Relational Data (SRL-2003)*, pages 24–31, 2003.

H. Cunningham. Information extraction, automatic. *Encyclopedia of Language and Linguistics*, 2005.

H. Cunningham, D. Maynard, K. Bontcheva, et V. Tablan. Gate : A framework and graphical development environment for robust nlp tools and applications. *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*, pages 168–175, 2002.

T. Dietterich. Machine learning for sequential data : A review. *Structural, Syntactic and Statistical Pattern Recognition ; Lecture Notes in Computer Science*, pages 15–30, 2002.

D. Downey, O. Etzioni, et S. Soderland. A probabilistic model of redundancy in information extraction. *IJCAI*, pages 1034–1041, 2005.

J. Dubois, M. Giacomo, L. Guespin, C. Marcellesi, J-B. Marcellesi, et J-P. Mevel. *Dictionnaire de linguistique*. Larousse, 2001.

P. Elango. Coreference resolution : A survey. *Project report of the course Advanced natural language processing In computer science departments university of Wisconsin Madison*, 36(8) :33–36, 2006.

O. Etzioni, B. Doorenbos, et D. Weld. A scalable comparison shopping agent for the world-wide web. *Proceedings of the International Conference on Autonomous Agents*, pages 39–48, 1997.

M. Ezzat. Acquision de grammaire locale pour l'extraction de relations entre entités nommées. *RECITAL2010*, 2010.

M. Ezzat et T. Poibeau. A new scheme for annotating semantic relations between named entities in corpora. *Proceedings of the Recent Advances in Natural Language Processing (ranlp 2011)*, pages 275–281, 2011.

C. Fabre et D. Bourigault. Extraction de relations sémantiques entre noms et verbes au-delà des liens morphologiques. *TALN'06*, pages 121–129, 2006.

R. Feldman, B. Rosenfeld, et M. Fresko. Teg-a hybrid approach to information extraction. *Knowledge and Information Systems*, 9 :1–18, 2006.

C. Fellbaum. Wordnet : An electronic lexical database. *Cambridge, MA : MIT Press*, 1998.

C. Fillmore, C. Johnson, et M. Petruck. Background to framenet. *International Journal of Lexicography*, pages 235–250, 2003.

J. Fleiss. *Statistical methods for rates and proportions*. John Wiley, New York, 1981.

M. Fossati. Ontologies owl d'entités nommées, exploitation des langages du web sémantique pour la modélisation de domaines de la connaissance. Master's thesis, INALCO, 2010.

D. Freitag. Multistrategy learning for information extraction. *Proceedings of International Conference on Machine Learning (ICML)*, pages 161–169, 1998.

D. Garcia. *Analyse automatique des textes pour l'organisation causales des actions, Réalisation du système informatique COATIS*. PhD thesis, Paris-Sorbonne, 1998.

N. Ge, J. Hale, et E. Charniak. A statistical approach to anaphora resolution. *Proceedings of the sixth Workshop on Very Large Corpora*, pages 161–170, 1998.

O. Gerbé. Introduction au formalisme des graphes conceptuels. URL `http://zonecours.hec.ca/documents/H2006-1-687571.introductioncgs.pdf`.

L. Gosselin. *Les modalités en français. La validation des représentations*. Rodopi, coll. "Etudes Chronos/Chronos Studies, 2010.

G. Grefenstette. *Explorations in automatic thesaurus discovery*. Kluwer Academic Publishers, 1994.

R. Grishman. Information extraction : Techniques and challenges. *Information Extraction (International Summer School SCIE-97)*, pages 10–27, 1997.

R. Grishman. The knowledge base population task : Challenges for information extraction, 2011.

R. Grishman et B. Sundheim. Message understanding conference-6 : A brief history. *Proceedings of the 16th Conference on Computational Linguistics*, pages 466–471, 1996.

R. Grishman, S. Huttunen, et R. Yangarber. Information extraction for enhanced access to disease outbreak reports. *Journal of Biomedical Informatics*, 35 :236–246, 2002.

C. Grouin, S. Galibert, L Rosset, et P. Zweigenbaum. Mesures d'évaluation pour entités nommées structurées. *Actes du 7è atelier Qualité des Données et des Connaissances – Evaluation des méthodes d'Extraction de Connaissances dans les Données*, 2011.

B. Habert et A. Nazarenko. La syntaxe comme marche-pied de l'acquisition des connaissances : Bilan critique d'une expérience. *Actes des septièmes Journées Acquisition des Connaissances (JAC'96)*, pages 137–148, 1996.

S. Handschuh, S. Staab, et F. Ciravegna. S-cream semi-automatic creation of metadata. *13th International Conference on Knowledge Engineering and Knowledge Management (EKAW)*, pages 358–372, 2002.

M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. *Proceedings of the 14th Conference on Computational Linguistics*, pages 539–545, 1992.

J. Hobbs. Resolving pronoun references. *Readings in natural language processing*, pages 339–352, 1986.

L. Horn. *A Natural History of Negation*. CSLI Publication, 2001.

C.-N. Hsu et M.-T. Dung. Generating finite-state transducers for semistructured data extraction from the web. *Information Systems Special Issue on Semistructured Data,*, 23 :521–538, 1998.

J. Huang, T. Chen, A. Doan, et J. F. Naughton. On the provenance of non-answers to queries over extracted data. *Proceedings of the VLDB Endowment*, 1(1) :736–747, 2008.

A. Jackiewicz. L'expression lexicale de la relation d'ingrédience. *Faits de langues*, 7 :53–62, 1996.

M. Jansche et S.P. Abney. Information extraction from voicemail transcripts. *emnlp'02*, pages 320–327, 2002.

T. S. Jayram, R. Krishnamurthy, S. Raghavan, S. Vaithyanathan, et H. Zhu. Avatar information extraction system. *IEEE Data Engineering Bulletin*, 29 :40–48, 2006.

J. Jiang et C. Zhai. A systematic exploration of the feature space for relation extraction. *Human Language Technologies 2007 : The Conference of the North American Chapter of the Association for Computational Linguistics ; Proceedings of the Main Conference*, pages 113–120, 2007.

C. Jouis. *Contribution à la conceptualisation et à la modélisation des connaissances à partir d'une analyse linguistique de textes. Réalisation d'un prototype : le système SEEK.* PhD thesis, LALIC, 1993.

D. Jurafsky et H. Martin. *Speech and Language Processing : An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics.* Prentice-Hall, 2009.

N. Kambhatla. Combining lexical, syntactic and semantic features with maximum entropy models for information extraction. *The Companion Volume to the Proceedings of 42st Annual Meeting of the Association for Computational Linguistics*, pages 178–181, 2004.

L. Kaufman et P. Rousseuuw. Clustering by means of medoid. *Reports of the Faculty of Mathematics and Informatics, Deft University of Technology*, 1987.

S. Khaitan, G. Ramakrishnan, S. Joshi, et A. Chalamalla. Rad : A scalable framework for annotator development. *ICDE'08*, pages 1624–1627, 2008.

S. Kim et E. Hovy. Extracting opinions, opinion holders, and topics expressed in online news media text. *Proceedings of the ACL Workshop on Sentiment and Subjectivity in Text*, pages 1–8, 2006.

D. Klein et C. D. Manning. Conditional structure versus conditional estimation in nlp models. *EMNLP'02*, pages 9–16, 2002.

P. Kogut et W. Holmes. Aerodaml : Applying information extraction to generate daml annotations from web pages. *First International Conference on Knowledge Capture (K-CAP)*, 2001.

K. Kolya. Event-event relation identification : A crf based approach. *Natural Language Processing and Knowledge Engineering (NLP-KE)*, pages 1–8, 2010.

K Kunz. Investigating nominal coreference in originals and translations. *SPRIK Conference*, 2006.

D. Kushal, S. Lawrence, et D. Pennock. Mining the peanut gallery : Opinion extraction and semantic classifcation of product reviews. *Proceedings of the 12th international conference on World Wide Web*, pages 519–528, 2003.

N. Kushmerick, D. Weld, et R. Doorenbos. Wrapper induction for information extraction. *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 729–737, 1997.

J. Lafferty, A. McCallum, et F. Pereira. Conditional random fields : Probabilistic models for segmenting and labeling sequence data. *Proceedings of the International Conference on Machine Learning (ICML-2001)*, pages 282–289, 2001.

S. Lawrence, C.L. Giles, et K. Bollacker. Digital libraries and autonomous citation indexing. *IEEE Computer*, 32 :67–71, 1999.

L. Lebart et A. Salem. *Statistique textuelle*. Dunod, 1994.

W. Lehnert, J. McCarthy, S. Soderland, E. Riloff, C. Cardie, J. Peterson, F. Feng, C. Dolan, et S. Goldman. Umass/hughes : Description of the circus system used for tipster text. *a Workshop on Held at Fredericksburg, Virginia*, page 241.256, 1993.

Y. Li et K. Bontcheva. Hierarchical, perceptron-like learning for ontology-based information extraction. *Proceedings of the 16th International Conference on World Wide Web (WWW)*, pages 777–786, 2007.

Y. Li, J. Jiang, H. Chieu, et K. Chai. Extracting relation descriptors with conditional random fields. *Asian Federation of Natural Language Processing*, pages 392–400, 2011.

L. Liu, C. Pu, et W. Han. Xwrap : An xml-enabled wrapper construction system for web information sources. *International Conference on Data Engineering (ICDE)*, pages 611–621, 2000.

J. MacQuenn. Some methods for classification and analysis of multivariate observations. *Fifth Berkley Symp. Math. Statistics and Probability*, 1 : 281–296, 1967.

R. Malouf. Markov models for language-independent named entity recognition. *Proceedings of the Sixth Conference on Natural Language Learning (CoNLL-2002)*, page 187, 2002.

C. D. Manning et H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999.

B. Marthi, B. Milch, et S. Russell. First-order probabilistic models for information extraction. *Working Notes of the IJCAI-2003 Workshop on Learning Statistical Models from Relational Data (SRL-2003)*, pages 71–78, 2003.

J. Martin et P. White. *Language Of Evaluation : Appraisal In English*. Palgrave Macmillan, 2005.

W. Martinez. Mise en évidence de rapports synonymiques par la méthode des cooccurrences. *Actes des 5è journées internationales d'analyse des données textuelles*, 1 :85–94, 2000.

Y. Mathieu. Verbes psychologiques et interprétation sémantique. *Langue française*, 105 :98–106, 1995.

Y. Mathieu. Un classement sémantique des verbes psychologiques. *Cahier du CIEL*, 1997.

Y. Mathieu. Les prédicats de sentiment. *Langages*, 136 :41–52, 1999.

Y. Mathieu. Navigation dans un texte a la recheches des sentiments. *Linguisticae Investigationes*, pages 313–322, 2008.

D. Maynard, V. Tablan, C. Ursu, H. Cunningham, et Y. Wilks. Named entity recognition from diverse text types. *Proceedings of the Recent Advances in Natural Language Processing (ranp 2001)*, 2001.

A. McCallum. Information extraction : Distilling structured data from unstructured text. *ACM Queue*, pages 48–57, 2005.

A. McCallum et W. Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. *Proceedings of the Seventh Conference on Natural Language Learning (CoNLL)*, 2003.

A. McCallum et B. Wellner. Toward conditional models of identity uncertainty with application to proper noun coreference. *Neural Information Processing systems (NIPS)*, pages 905–912, 2003.

A. McCallum, D. Freitag, et F. Pereira. Maximum entropy markov models for information extraction and segmentation. *Proceedings of the International Conference on Machine Learning (ICML-2000)*, pages 591–598, 2000a.

A. McCallum, K. Nigam, J. Reed, J. Rennie, et K. Seymore. Cora : Computer science research paper search engine. http ://cora.whizbang.com, 2000b.

A. K. McCallum. Mallet : A machine learning for language toolkit. *http ://mallet.cs.umass.edu*, 2002.

D. McDonald, H. Chen, H. Su, et B. Marshall. Extracting gene pathway relations using a hybrid grammar : The arizona relation parser. *Bioinformatics*, 20 :3370–3378, 2004.

A. Mikheev, M. Moens, et C. Grover. Named entity recognition without gazetteers. *In Proceedings of the European Conference of the Association for Computational Linguistics*, pages 1–8, 1999.

R. Miktov. Robust pronoun resolution with limited knowledge. *n Proceedings of COLING-ACL 1998*, pages 869–875, 1998.

G. Miller. Wordnet : A lexical database for english. *ACM*, 38 :39–41, 1995.

I. Muslea. Extraction patterns for information extraction tasks : A survey. *The AAAI-99 Workshop on Machine Learning for Information Extraction*, pages 1–6, 1999.

I. Muslea, S. Minton, et C. A. Knoblock. A hierarchical approach to wrapper induction. *Proceedings of the Third International Conference on Autonomous Agents*, pages 190–197, 1999.

J-C. Na, C. Khoo, et P.H.J. Wu. Use of negation phrases in automatic sentiment classification of product reviews. *Library Collections, Acquisitions and Technical Services*, 29 :180–191, 2005.

R. Narayanan, B. Liu, et A. Choudhary. Sentiment analysis of conditional sentences. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'09)*, 1 :180–189, 2009.

B. Pang et L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, pages 1–135, 2008.

B. Pang, L. Lee, et S. Vaithyanathan. Thumbs up ? sentiment classification using machine learning techniques. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'02)*, pages 79–86, 2002.

J. Pearson. Comment accéder aux éléments définitoires dans les textes spécialisés. *Terminologies nouvelles*, 19 :21–29, 1999.

F. Peng et A. McCallum. Accurate information extraction from research papers using conditional random fields. *HLT-NAACL*, pages 329–336, 2004.

H. Perdicoyanni-Paléologou. Le concept d'anaphore, de cataphore et de déixis en linguistique française. *Revue québécoise de linguistique*, 29(2) :55–57, 2001.

C. Plake, T. Schiemann, M. Pankalla, J. Hakenberg, et U. Leser. Alibaba : Pubmed as a graph. *Bioinformatics*, 22 :2444–2445, 2006.

T. Poibeau. *Extraction automatique d'information. Du texte brut au web sémantique*. Hermès, 2003.

T. Poibeau. Sur le statut référentiel des entités nommées. *Actes de la conférence Traitement Automatique des Langues Naturelles*, 2005.

A. Popescu et O. Etzioni. Extracting product features and opinions from reviews. *Proceedings of HLT/EMNLP*, pages 339–346, 2005.

R. Prasad, N. Dinesh, A. Lee, A. Joshi, et B. Webber. Attribution and its annotation in the penn discourse treebank. *ACL Workshop on Sentiment and Subjectivity in Text*, 2007.

J. R. Quinlan. Learning logical definitions from relations. *Machine Learning*, 5 :239–266, 1990.

R. Quirk, S. Greenbaum, G. Leech, et J. Svartvik. *A Comprehensive Grammar of the English Language*. Longman, 1985.

G. Ramakrishnan, S. Joshi, S. Balakrishnan, et A. Srinivasan. Using ilp to construct features for information extraction from semi-structured text. *Proceedings of the 17th International Inductive Logic Programming*, pages 211–224, 2007.

A. Ratnaparkhi. Learning to parse natural language with maximum entropy models. *Machine Learning*, 34 :151–176, 1999.

F. Reiss, S. Raghavan, R. Krishnamurthy, H. Zhu, et S. Vaithyanathan. An algebraic approach to rule-based information extraction. *ICDE*, 2008.

E. Riloff. Automatically constructing a dictionary for information extraction tasks. *Proceedings of the Eleventh National Conference on Artificial Intelligence*, pages 811–816, 1993.

E. Riloff et J. Wiebe. Learning extraction patterns for subjective expressions. *Conference on Empirical Methods in Natural Language Processing (EMNLP'03)*, pages 105–112, 2003.

B. Rosenfeld et R. Feldman. Using corpus statistics on entities to improve semi-supervised relation extraction from the web. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 600–607, 2007.

A. Roux-Thomas. Ontoloties géographiques. Master's thesis, LIPN Paris XIII, 2009.

V. Rubin. Stating with certainty or stating with doubt : Intercoder reliability results for manual annotation of epistemically modalized statements. *NAACL-HLT 2007*, pages 141–144, 2007.

V. Rubin. Epistemic modality : From uncertainty to certainty in the context of information seeking as interactions with texts. *Information Processing & Management*, 46 :533–540, 2010.

V. Rubin, E. Liddy, et N. Kando. Certainty identification in texts : Categorization model and manual tagging results. *Computing Attitude and Affect in Text : Theory and Applications, volume 20 of The Information Retrieval Series*, pages 61–76, 2006.

J. Ruppenhofer, S. Somasundaran, et J. Wiebe. Finding the sources and targets of subjective expressions. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*, 2008.

N. Sager, C. Friedman, et M. Lyman. *Medical Language Processing : Computer Management of Narrative Data*. Addison-Wesley, 1987.

S. Sarawagi. The crf project : A java implementation. *http ://crf.sourceforge.net*, 2004.

S. Sarawagi et A. Bharnidipaty. Interactive deduplication using active learning. *Proceedings of the 8th ACM SIGKDD International Conference*, pages 269–278, 2002.

R. Sauri et J. Pustejovsky. Factbank : a corpus annotated with event factuality. *Language Resources and Evaluation*, 43 :227–268, 2009.

P. Seguela. *Construction de modèles de connaissances par analyse linguistique de relations lexicales dans les documents techniques*. PhD thesis, CEA, Société de Service en Ingénierie Informatique Euriware, 2001.

S. Sekine, K. Sudo, et N Chikashi. Extended named entity hierarchy. *In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*, 2002.

K. Seymore, A. McCallum, et R. Rosenfeld. Learning hidden markov model structure for information extraction. *Papers from the AAAI- 99 Workshop on Machine Learning for Information Extraction*, pages 37–42, 1999.

W. Shen, A. Doan, J. F. Naughton, et R. Ramakrishnan. Declarative information extraction using datalog with embedded extraction predicates. *VLDB*, pages 1033–1044, 2007.

M. Shilman, P. Liang, et P. Viola. Learning non-generative grammatical models for document analysis. *Tenth IEEE International Conference on In Computer Vision (ICCV'05)*, pages 962–969, 2005.

Y. Shinyama et S. Sekine. Preemptive information extraction using unrestricted relation discovery. *Proceedings of the Human Language Technology Conference of the NAACL*, 2006.

C. Shirky. Ontology is overrated, 2005. URL `http://www.shirky.com/writings/ontology`$_o$*verrated.html*.

C.L. Sidner. Focusing for interpretation of pronouns. *Computational Linguistics*, 7(4) :217–231, 1981.

F. Smajda. Retrieving collocations from text : Xtract. *Computational Linguistics*, (191) :143–178, 1993.

S. Soderland. Learning information extraction rules for semi-structured and free text. *Machine Learning*, 34(1) :233–272, 1999.

S. Soderland, D. Fisher, J. Aseltine, et W. Lehnert. Crystal : Inducing a conceptual dictionnary. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 1314–1321, 1995.

W.M. Soon, H.T. Ng, et D.C.Y. Lim. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4) : 521–544, 2001.

J. Sowa. *Conceptual Strcutures : Information Processing in Mind and Machine*. Addison-Wesley, 1984.

M. Strube. Never look back : an alternative to centering. *Proceedings of the 17th International conference on Computational linguistics*, pages 1251–1257, 1998.

F. M. Suchanek, G. Ifrim, et G. Weikum. Combining linguistic and statistical analysis to extract relations from web documents. *KDD 06 : Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 712–717, 2006.

F. M. Suchanek, G. Kasneci, et G. Weikum. Yago : A core of semantic knowledge. *WWW 07 : Proceedings of the 16th International Conference on World Wide Web*, pages 697–706, 2007.

B. M. Sundheim. Overview of the third message understanding evaluation and conference. *Proceedings of the Third Message Understanding Conference (MUC-3)*, pages 3–16, 1991.

K. Takeuchi et N. Collier. Use of support vector machines in extended named entity recognition. *The 6th Conference on Natural Language Learning (CoNLL)*, pages 119–125, 2002.

L. Tesnière. *Eléments de syntaxe structurale*. Klinksieck, 1959.

J.R. Tetreault. Analysis of syntax-based pronoun resolution methods. *Proceedings of the 37th annual meeting on Association for Computational Linguistics*, pages 602–605, 1999.

C. Touratier. *La sémantique*. Armand Colin, 2004.

J. Turmo, A. Ageno, et N. Catala. Adaptive information extraction. *ACM Computer Services*, 38 :4, 2006.

P. D. Turney. Expressing implicit semantic relations without supervision. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pages 313–320, 2006.

M. Valette. A quoi servent les lexiques sémantiques généralistes ? discussion et proposition. *Cahiers du Cental*, 5 :43–58, 2008.

K. Van Deemter et R. Kibble. On coreferring : Coreference in muc and related annotation schemes. *Computational Linguistics*, 26 :629–637, 2000.

CJ. Van Rijsbergen. *Information Retrieval*. ButterWorths, 1979.

P. Viola et M. Narasimhan. Learning to extract information from semistructured text using a discriminative context free grammar. *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 330–337, 2005.

A.J. Viterbi. A personal history of the viterbi algorithm. *Signal Processing Magazine, IEEE*, 23(4) :120–142, 2006.

S. Wen-tau Yih et K. Toutanova. Automatic semantic role labeling. *Proceedings of the Human Language Technology Conference of the NAACL*, pages 309–310, 2006.

J. Wiebe, T. Wilson, et C. Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2) :165–210, 2005.

M. Wilmet. Les déterminants du nom en français : essai de synthèse. *Langue française*, 57 :15–33, 1983.

J. Yi, T. Nasukawa, R. Bunescu, et W. Niblack. Sentiment analyzer : Extracting sentiments about a given topic using natural language processing techniques. *Third IEEE International Conference*, pages 427–434, 2003.

D. Zelenko, C. Aone, et A. Richardella. Kernel methods for relation extraction. *Journal of Machine Learning Research*, 3 :1083–1106, 2003.

M. Zhang, J. Zhang, J. Su, et G. Zhou. A composite kernel to extract relations between entities with both flat and structured features. *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 825–832, 2006.

S. Zhao et R. Grishman. Extracting relations with integrated information using kernel methods. *In Proceedings of the annual meeting of ACL*, pages 419–426, 2005.

Appendices

Annexe A

Description des corpus utilisés

Nous présentons ici un descriptif des corpus utilisés dans le cadre de notre travail.

A.1 Corpus de presse généraliste

Nous avons eu accès à deux corpus de presse généraliste dont nous présentons les caractéristiques dans le tableau suivant :

Nom du corpus	Nombre de mots
Le Monde 2007	17 621 487
Corpus « news » Arisem	2 912 232

Le corpus « Le monde 2007 » est une ressource payante achetée à la société Elda (elda.org) et regroupe tous les articles du journal « Le monde » sur l'année 2007.
Le corpus « news » Arisem est un corpus fourni par l'entreprise Arisem et regroupe l'aspiration de plusieurs sites de nouvelles sur Internet.

A.2 Corpus des projets auxquels nous avons participé

Nous présentons ici les corpus issus de différents projets auxquels nous avons eu accès durant nos travaux dans l'entreprise.

A.2.1 Infom@gic

Infom@gic est un projet de Recherche et Développement dans le domaine de l'Analyse de l'Information, visant la mise en place d'un laboratoire industriel de sélection, de tests, d'intégration et de validation d'applications opérationnelles des meilleures technologies franciliennes dans le domaine de l'ingénierie des connaissances. Il a été lancé par le pôle de compétitivité Cap Digital en 2006 pour une durée de 3 ans. Il a vocation à produire une plate-forme commune d'interopérabilité, laquelle doit faciliter l'intégration de composants d'analyse de l'information dans les domaines de la recherche, de l'indexation, l'extraction de connaissances et la fusion d'informations multimédias. Le corpus est composé de 13 570 dépêches de l'Agence France Presse pour un nombre de mots s'élevant à 6 586 044.

A.2.2 DoXa

DoXa (2009 - 2012), projet de recherche et développement du pôle de compétitivité francilien « Cap digital », vise à mettre en place une plateforme de technologies liées au traitement automatique des sentiments et opinions dans des ensembles de données multilingues (français et anglais). Les ensembles traités intègrent de grands volumes de données à la fois non-structurées (issues du web 2.0 : blogs, forums, réseaux sociaux etc.), et de données structurées (provenant de bases de données clients).

Nous avons eu accès en particulier à deux corpus dans le cadre de ce projet. Le premier corpus relève du domaine du jeu vidéo et est composé de posts de forum ou d'article de blogs, totalisant un nombre de mots de 8 918 468.

Le second corpus est issu d'une enquête de satisfaction client (« Baromètre Satisafacion Marché ») en provenance de la base marketing clients EDF, autour d'une problématique de départ à la concurrence. Pour ce corpus, seules les années 2006 à 2008 ont été gardées, ce qui représente un total de 7 408 clients.

A.2.3 Cahors

Le projet Cahors (acronyme de Cotation, Analyse, Hiérarchisation et Ontologies pour le Renseignement et la Sécurité) est un projet financé par l'agence nationale de la recherche (ANR) et s'étend de 2009 à 2012. Cahors a été directement suivi par la DGA et la Police Nationale. Le corpus est

composé de 79 000 articles [1]

A.3 Extraits des corpus

La France n'est pas le seul pays à connaître une embellie durable de l'emploi. Selon les chiffres harmonisés d'Eurostat, le taux de chômage atteignait 7,5 % dans la zone euro fin 2006, contre 8,5 % en France et 7,9 % en Allemagne. C'est le meilleur résultat depuis 1993, date de la création de statistiques harmonisées concernant les pays de la zone euro. Après des années de modération salariale et des réformes, l'Allemagne redécolle. Elle recrée de l'emploi, après en avoir détruit dans la foulée de la réunification. La hausse de la TVA au 1er janvier 2007 ne semble pas affecter la croissance. Mais c'est l'Espagne qui a connu le rétablissement le plus spectaculaire au cours de la dernière décennie : elle a créé plus de 6 millions d'emplois et réduit son taux de chômage, désormais équivalent à celui de la France. Cette performance, accompagnée d'une forte croissance, s'est aussi faite au prix de la multiplication d'emplois précaires ou mal payés. En Italie aussi, le chômage a reculé, en dépit d'une croissance médiocre.

FIGURE A.1 – Extrait du corpus « Le Monde 2007 »

WASHINGTON — Le groupe internet américain Google envisagerait une offre d'achat sur son concurrent Yahoo !, selon des informations publiées samedi par le Wall Street Journal. "Google a discuté avec au moins deux sociétés de capital-investissement de la possibilité que celles-ci l'aident à financer un accord pour acheter le coeur de métier de Yahoo !", écrit le journal sur son site internet, citant "une personne au courant de la situation". Google n'a eu que des entretiens "liminaires" sur le sujet et pourrait "finir par abandonner l'idée d'une offre sur Yahoo !", ajoute le quotidien.

FIGURE A.2 – Extrait du corpus « news Arisem »

1. Nous ne pouvons pas spécifier ici les sources, ni la taille du corpus en mots pour des raisons de confidentialité

Un autre parti s'oppose aux papiers administratifs pour voter aux régionales
L'Union pour la démocratie et la paix en Côte d'Ivoire (UDPCI, opposition) a dénoncé vendredi une décision de la Commission électorale indépendante (CEI) sur les pièces requises pour voter aux élections régionales de juillet prochain, a constaté l'AFP. Lors d'une conférence de presse, le secrétaire général adjoint de l'UDPCI, le député Toikeuse Mabri, a exigé la délivrance "d'attestations d'identité spéciales élections départementales", ou à défaut la reconnaissance des documents qui avaient servi lors des différents scrutins de 2000. La CEI, dans laquelle sont représentés tous les partis, avait récemment annoncé que seules les nouvelles cartes d'identité et les nouvelles attestations d'identité seront acceptées pour voter aux élections régionales du 7 juillet. L'UDPCI rappelle que l'opération d'identification de 1999 n'a permis de donner aux Ivoiriens qu'un récépissé. Plusieurs autres citoyens, a déploré le deputé Mabri, "sont restés dans l'attente de la délivrance du récépissé". Mardi, le Rassemblement des républicains (RDR, principal parti d'opposition) de l'ancien Premier ministre ivoirien Alassane Ouattara, avait déjà dénoncé la décision de la CEI. Selon ce parti, seules 2,5 millions de nouvelles cartes et 325.000 attestations d'identité avaient été délivrées. Il avait estimé que la mesure de la CEI excluait près de la moitié des 5,5 millions d'électeurs inscrits. Une opération "d'identification de la population" ivoirienne et étrangère, qui vise a doter tous les habitants de nouvelles cartes d'identité ou de séjour, a démarré lundi dernier. Pour M. Ouattara, qui avait été exclu des présidentielle et législatives de 2000 pour nationalité "douteuse", cette opération pourrait déboucher sur "un apartheid à l'ivoirienne".
str-omj/jlh/pyj dab

FIGURE A.3 – Extrait du corpus « Infom@gic »

Infos : Unreal Tournament 3 compatible avec les modsE3 2007Unreal Tournament 3 compatible avec les mods — le 12 juillet 2007 à 18h26, par Vincent - Bonne nouvelle pour la durée de vie de Unreal Tournament 3 , la version PlayStation 3 sera également compatible avec les mods de la version PC qu'il faudra certainement télécharger sur le PlayStation Store gratuitement. Attendez-vous donc à un jeu long, très long jouable pendant plusieurs mois et avec clavier et souris si vous le désirez ! Lire sur le forum les 1 commentaires Les derniers posts du forum 12-07-2007 à 23 :27 :01 Ca franchement, c'est une aubaine :) Réagissez sur le forum Unreal Tournament 3 9 / 10 Version à Imprimer Envoyer à un amimédias disponibles47 images 13 artworks 5 vidéosarticles disponiblesLire le test Lire la preview Forum du jeunote de la rédaction9/10 Voir toutes les notes

FIGURE A.4 – Extrait du corpus « Doxa Jeu Vidéo »

pour l'instant j'ai rien à rapprocher, mais si un concurrent me fournit les mêmes prestations à un prix moins cher je pourrais changer, reste à savoir pour combien de temps on prend l'engagement. pour ne pas dire 10 car je n'ai pas de données concrètes pour juger

FIGURE A.5 – Extrait du corpus « Doxa EDF »

Selon nos informations, le décret de 2005 fixant les attributions du chef d'état-major des armées (Cema) devrait être remplacé dans les prochains mois par un nouveau texte. Celui-ci fait l'objet de vives discussions entre le cabinet du ministre de la défense et l'état-major des armées. L'enjeu en est l'étendue des pouvoirs du Cema, qu'Hervé Morin souhaite réduire, en remettant le ministre de la défense au centre des affaires."Nous voulons clarifier les responsabilités avec un texte qui définira la gouvernance du ministère", estime une source proche du dossier. "Nous voulons rappeler que le ministre est à la tête du ministère de la défense". Selon l'idée qui prévaut à l'Hotel de Brienne, le schéma idéal est le suivant : le ministre est le patron, assisté de trois grands responsables : le Cema, le délégué général pour l'armement (DGA) et le secrétaire général pour l'administration (SGA). Ensemble, ils forment déjà le ComEx, le comité éxécutif. Cette organisation devrait être concrétisée par la construction du Pentagone de Balard, où les grands chefs seront réunis sous l'autorité du ministre.Schéma idéal, sauf... qu'il y manque le président de la République. Or, celui-ci est le véritable chef des armées. Dans le domaine des opérations, un lien direct existe entre le chef de l'Etat et le Cema. Le ministre de la défense ne commandant pas aux forces, il se trouve donc marginalisé, voire exclu, du processus de décision opérationnelle.Le nouveau décret ne modifiera pas cette donnée fondamentale du fonctionnement de nos institutions, mais il devrait préciser les domaines dans lesquels le Cema est placé sous l'autorité directe du président de la République. Pour le reste, le Cema sera considéré comme subordonné au ministre. Le diable se nichant dans les détails, la rédaction de cette partie du nouveau décret fait l'objet de discussions animées... L'actuel Cema, le général Georgelin, est l'un des principaux auteurs du décret de 2005 ; il était alors chef d'état-major particulier du président de la République. Il considère que le rôle du ministère de la défense, qu'il rebaptiserait volontiers ministère des armées, est de fournir aux armées les moyens leur permettant d'obéir aux ordres de leur chef, c'est-à-dire du président de la République. Cette vision ne déclenche pas l'enthousiasme du ministre Hervé Morin...Autre point de crispation : la place des chefs d'état-major d'armée (Terre, Air, Mer). Depuis 2005, le Cema "a autorité" sur eux. Le décret en préparation vise là encore à remettre le ministre dans le jeu, en précisant les domaines dans lequel les chefs d'état-major sont sous l'autorité du Cema et ceux pour lesquels ils dépendent directement du ministre.La rédaction de ce nouveau décret intervient sur fond de succession du Cema, qui sera atteint par la limite d'âge (61 ans) fin aout. Son successeur devrait être connu dans les prochaines semaines, l'amiral Edouard Guillaud faisant toujours partie des principaux favoris. Problème pour Hervé Morin : l'amiral vient directement de l'Elysée... comme les ordres que reçoit le Cema.

FIGURE A.6 – Extrait du corpus « Cahors »

Annexe B

Exemples de grammaire

Afin de donner un aperçu de la forme des grammaires, nous donnons ici deux exemples. La première est construite manuellement par un linguiste. La seconde est le fruit du sysème mis en place et décrit dans notre travail au chapitre 8.

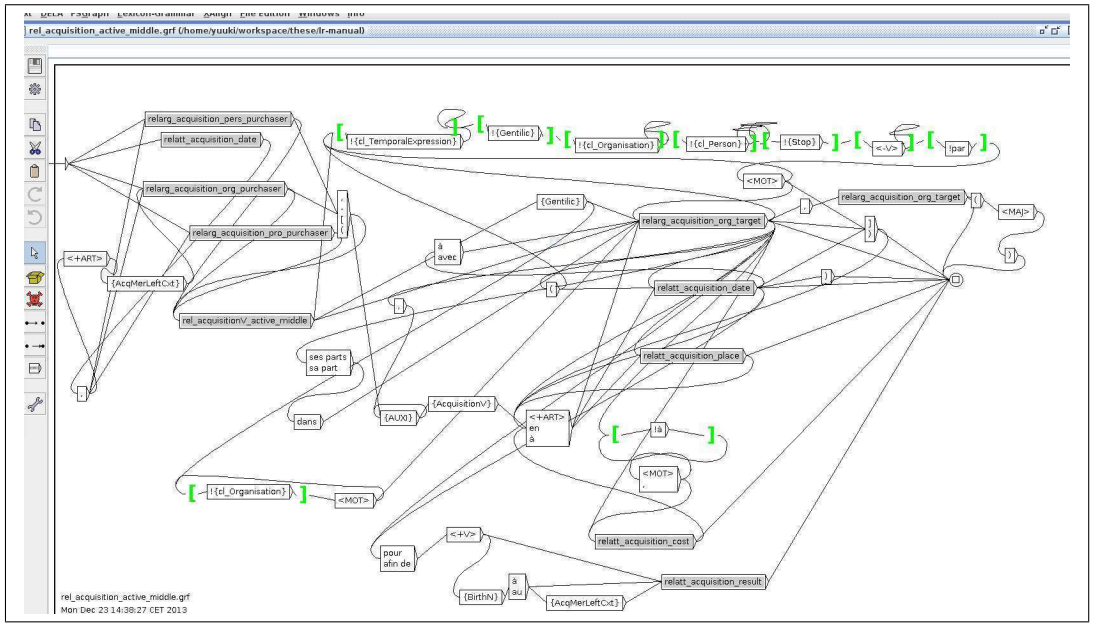


FIGURE B.1 – Exemple de grammaire construite manuellement pour la relation d'achat d'une entreprise par une autre où le prédicat est placé entre les deux arguments

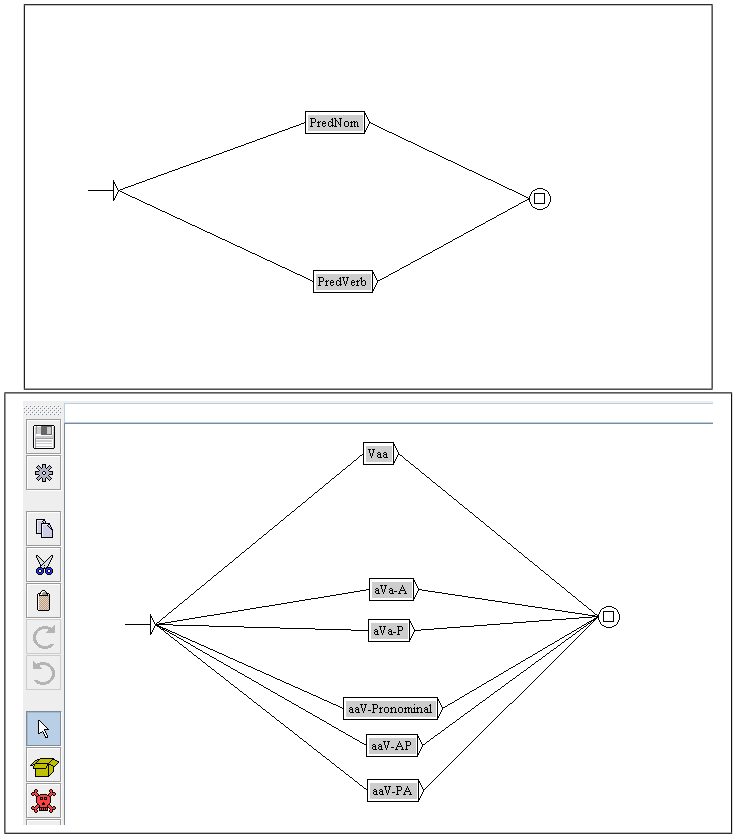
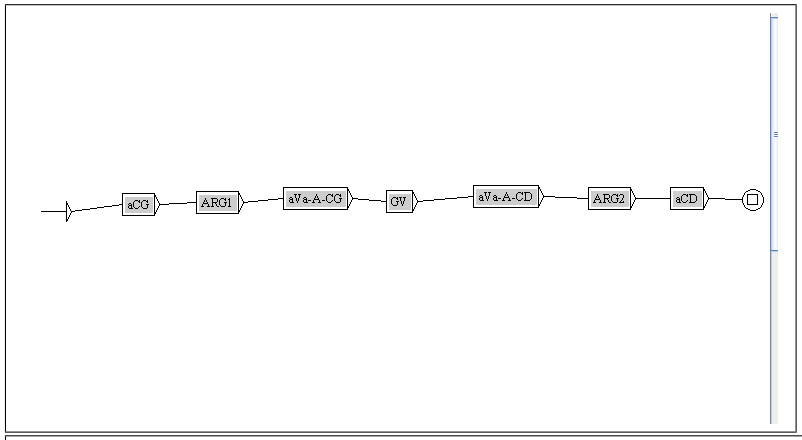
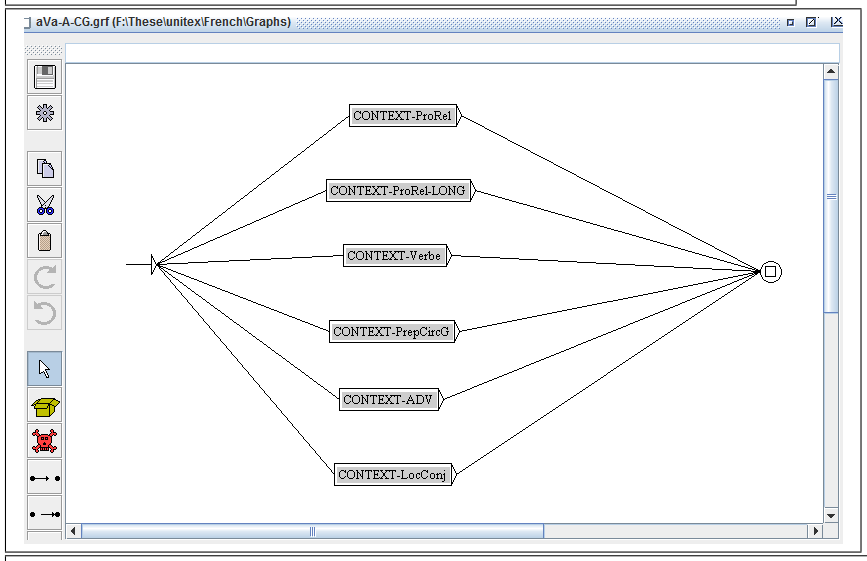
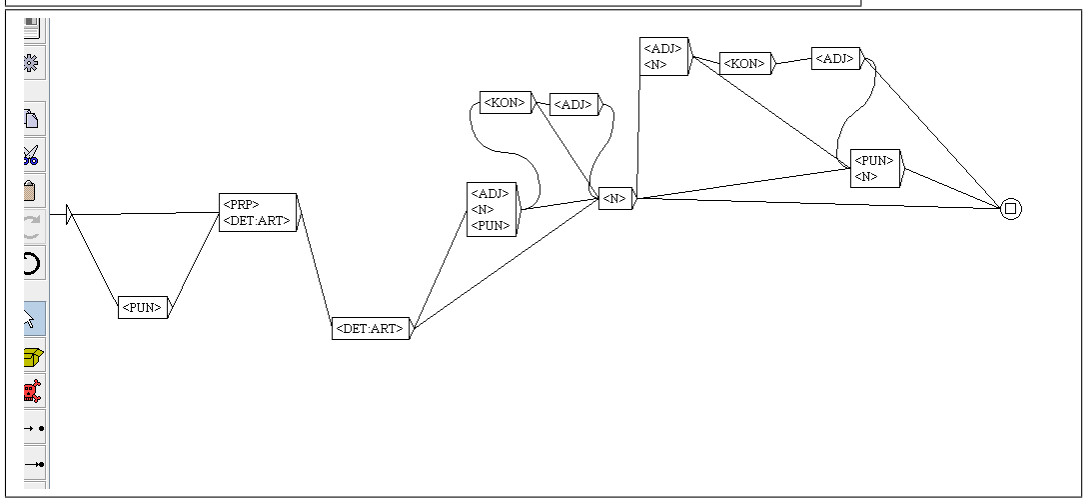


FIGURE B.2 – Exemple de grammaire construite par le système.


aCG
ARG1
aVa-A-CG
GV
aVa-A-CD
ARG2
aCD


aVa-A-CG.grf (F:\These\unitex\French\Graphs)
CONTEXT-ProRel
CONTEXT-ProRel-LONG
CONTEXT-Verbe
CONTEXT-PrepCircG
CONTEXT-ADV
CONTEXT-LocConj


<PUN>
<PRP>
<DET:ART>
<DET:ART>
<KON>
<ADJ>
<ADJ>
<N>
<PUN>
<ADJ>
<N>
<KON>
<ADJ>
<PUN>
<N>

Mani EZZAT

Acquisition de relations entre entités nommées à partir de corpus

Résumé

Les entités nommées ont été l'objet de nombreuses études durant les années 1990. Leur reconnaissance dans les textes a atteint un niveau de maturité suffisante, du moins pour les principaux types (personne, organisation et lieu), pour aller plus loin dans l'analyse, vers la reconnaissance de relations entre entités. Il est par exemple intéressant de savoir qu'un texte contient des occurrences des mots « Google » et « Youtube » ; mais l'analyse devient plus intéressante si le système est capable de détecter une relation entre ces deux éléments, voire de les typer comme étant une relation d'achat (Google ayant racheté Youtube en 2006).

Notre contribution s'articule autour de deux grands axes : tracer un contour plus précis autour de la définition de la relation entre entités nommées, notamment au regard de la linguistique, et explorer des techniques pour l'élaboration de systèmes d'extraction automatique qui sollicitent des linguistes.

MOTS CLES : Extraction d'information, relation, entités nommée, grammaire locale, corpus, linguiste, ontologie

Résumé en anglais

Named entities have been the topic of many researches during the 90's. Their detection in texts has reached a high level of performance, at least for the main categories (person, organization and location). It becomes now possible to go further, toward relation between entities recognition. For instance, knowing that a text contains the words "Google" and "Youtube" can be relevant but being able to link them and detect an acquisition relation can be more interesting (Google has bought Youtube in 2006).

Our work is focusing on two different aspects: to define a finer perimeter around the relation between named entities definition, with linguistic aspect in mind, and to explore new techniques that make use of linguists in order to build a relation between named entities recognition system.

KEYWORDS: Information extraction, relation, named entities, local grammar, corpora, linguist, ontology