Identification automatique d'entités pour l'enrichissement de contenus textuels

Rosa Stern

Université Paris 7 Denis Diderot

École doctorale de sciences du langage

Linguistique théorique, descriptive et automatique

THÈSE DE DOCTORAT

présentée par
Rosa STERN

Identification automatique d'entités pour l'enrichissement de contenus textuels

---

réalisée dans le cadre d'une convention CIFRE avec l'Agence France-Presse (Medialab)

Composition du jury

M. Frédéric BÉCHET (rapporteur)	Université de la Méditerranée
M. Éric CHARTON (examinateur)	École Polytechnique de Montréal
Mme Laurence DANLOS (directrice)	Université Paris 7 & INRIA
Mme Adeline NAZARENKO (rapporteur)	Université Paris Nord & CNRS
M. Benoît SAGOT (co-directeur)	INRIA
M. Denis TEYSSOU (responsable en entreprise)	Agence France-Presse

Identification automatique d'entités pour l'enrichissement de contenus textuels

Résumé

Cette thèse propose une méthode et un système d'identification d'entités (personnes, lieux, organisations) mentionnées au sein des contenus textuels produits par l'Agence France Presse dans la perspective de l'enrichissement automatique de ces contenus. Les différents domaines concernés par cette tâche ainsi que par l'objectif poursuivi par les acteurs de la publication numérique de contenus textuels sont abordés et mis en relation : Web Sémantique, Extraction d'Information et en particulier Reconnaissance d'Entités Nommées (REN), Annotation Sémantique, Liage d'Entités. À l'issue de cette étude, le besoin industriel formulé par l'Agence France Presse fait l'objet des spécifications utiles au développement d'une réponse reposant sur des outils de Traitement Automatique du Langage. L'approche adoptée pour l'identification des entités visées est ensuite décrite : nous proposons la conception d'un système prenant en charge l'étape de REN à l'aide de n'importe quel module existant, dont les résultats, éventuellement combinés à ceux d'autres modules, sont évalués par un module de Liage capable à la fois (i) d'aligner une mention donnée sur l'entité qu'elle dénote parmi un inventaire constitué au préalable, (ii) de repérer une dénotation ne présentant pas d'alignement dans cet inventaire et (iii) de remettre en cause la lecture dénotationnelle d'une mention (repérage des faux positifs). Le système Nomos est développé à cette fin pour le traitement de données en français. Sa conception donne également lieu à la construction et à l'utilisation de ressources ancrées dans le réseau des Linked Data ainsi que d'une base de connaissances riche sur les entités concernées.

Mots-clefs

extraction d'information, web sémantique, annotation sémantique, linked data, entités, reconnaissance d'entités nommées, liage

---

Automatic Entity Identification for Textual Content Enrichment

Abstract

This dissertation proposes a method and a system for the identification of entities (persons, locations, organizations) mentionned in the textual production of the news agency Agence France

Presse, in the prospect of the automatic content enrichment. The various fields concerned by this task are viewed through their relationship: Semantic Web, Information Extraction and in particular Named Entity Recognition (NER), Semantic Annotation, Entity Linking. Following this study, the industrial need expressed by the Agence France Presse is the subject of specifications, useful for the development of a solution relying on Natural Language Processing tools. The approach adopted for the identification of the target entities is then described: we propose a system taking charge of the NER step using any existing module, whose results, possibly combined with those of other modules, are evaluated by a linking module able to (i) align a given mention with the entity it denotes among an inventory, built prior to the task, (ii) to spot denotations without alignment in the inventory and (iii) to reconsider denotational readings of mentions (false positive detection). The Nomos system is developed to this end for the processing of French data. Its conception also gives rise to the building and use of resources integrated into the Linked Data network, as well as a rich knowledge base about the target entities.

Keywords

information extraction, semantic web, semantic annotation, linked data, entities, named entitiy recognition, linking

Je remercie Laurence Danlos, Benoît Sagot et Denis Teyssou de m'avoir accueillie pour réaliser cette thèse dans l'équipe Alpage et au Medialab de l'Agence France Presse. Je remercie également les rapporteurs, Adeline Nazarenko et Frédéric Béchet, d'avoir accepté d'évaluer ce travail et d'y avoir porté un regard ouvert et enrichissant. Je suis reconnaissante à Éric Charton pour son soutien dans l'élaboration de ma réflexion et pour son accueil au CRIM de Montréal.

Merci à mes proches, en particulier Dahlia et Joseph, pour leur soutien intellectuel et moral mis à l'épreuve de péripéties spirituelles en tous genres mais pourtant toujours tangible. Merci à Julie et Elena pour leur écoute d'une patience infaillible et qui ont réussi à ne pas oublier mes bons côtés.

She generally gave herself very good advice, (though she very seldom followed it).

Begin at the beginning and go on till you come to the end ; then stop.

"But I don't want to go among mad people," said Alice. "Oh, you can't help that," said the cat. "We're all mad here."

Lewis Carroll

Jamais, répondit le duc d'Auge. Je lui ai déjà expliqué que je ne voulais plus remettre les pieds dans ces bleds impossibles. Une croisade c'est beaucoup ; deux c'est trop.

D'abord, dit paisiblement Gabriel, c'est pas vrai et, deuzio, i comprendront pas.

Raymond Queneau

Table des matières

Introduction

L'enrichissement de contenus textuels : domaine de définition

Le sujet porté à notre étude s'articule autour d'un besoin applicatif de l'Agence France-Presse (AFP), abordé dans le cadre du Medialab, son département chargé de la prospection dans le domaine de l'innovation technologique liée aux métiers des medias. Ce besoin y est formulé par le terme d'*enrichissement de contenus textuels*, souvent étendu à l'expression *enrichissement sémantique de contenus textuels à l'aide de métadonnées* dans les travaux de spécification à l'initiative du Medialab. Si l'on peut identifier les contenus textuels à traiter ici comme la production de l'AFP, qui diffuse quotidiennement, sous forme numérique, plusieurs milliers de dépêches et autres supports informatifs accompagnés de texte — photographies légendées, retranscription de videos ou infographies —, l'enrichissement sémantique à l'aide de métadonnées doit quant à lui faire l'objet d'une définition. Ce terme désigne en effet un processus de traitement des contenus dans un but applicatif, dont il s'agit de comprendre la nature afin de proposer une réponse, notamment à l'aide de techniques de traitement automatique du langage (TAL), au besoin exprimé.

La figure 1 illustre de façon liminaire le type de résultat visé par l'enrichissement sur une dépêche diffusée par l'AFP. Au sein du contenu textuel relatant l'information donnant lieu à la dépêche elle-même, des indications sous forme de balises marquent un certain nombre d'éléments textuels ainsi mis en valeur par rapport au reste du document. Ces indications, distinctes du niveau textuel lui-même et donc des *données* diffusées, constituent ainsi des *métadonnées* de la dépêche. À la différence des métadonnées de documents entendues au sens usuel, telles que les informations de date, d'auteur ou de propriété associées au document mais distinctes du contenu informatif, ces métadonnées sont *ancrées* dans le contenu textuel et relient les éléments marqués à des ressources extérieures au document, par le mécanisme des URI (Uniform Resource Identifier). Dans cet exemple, il s'agit de mentions d'entités (personnes, organisations et lieux) qui, marquées par des balises, sont ainsi reliées à des ressources informatives les décrivant (biographie, description, localisation géographique).

On constate avec cette première illustration que l'enrichissement à l'aide de métadonnées relève d'un objectif de mise en relation de données, à partir de documents — par exemple les flux de dépêches de l'AFP — vers des ressources à même de compléter l'information source par de nouveaux éléments. Ceux-ci viennent ainsi *enrichir* l'espace informatif délimitée par le contenu d'un document par son ouverture sur des ressources externes.

Le Web Sémantique L'enrichissement de contenus se présente comme un enjeu d'intérêt pour l'AFP en tant qu'il reflète les développements technologiques initiés depuis environ une décennie dans le cadre du *Web Sémantique* : celui-ci se pose en effet comme un paradigme de publication documentaire proposant d'en renouveler les pratiques. Comme acteur de diffusion de l'information, l'AFP est concernée par ces pratiques, qui sont historiquement caractérisées par un effort important d'organisation, de représentation et de mise à disposition des données. Il s'agit avec le Web Sémantique, défini par l'un de ses principaux initiateurs, Tim Berners-Lee, comme une

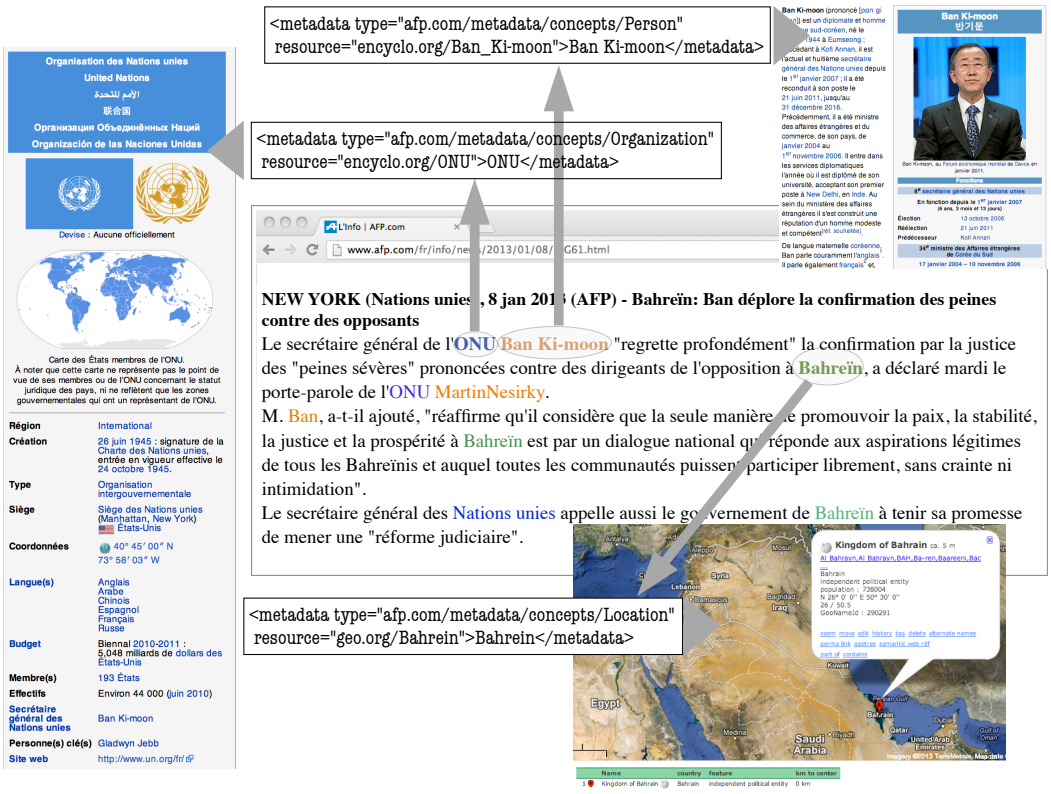


FIGURE 1 : Enrichissement de contenus textuels à l'aide de métadonnées.

« nouvelle forme de contenus sur le Web, muni de sens pour les machines » [BLHL01], d'intégrer pleinement la production de données sous toutes ses formes à l'organisation en réseau du Web, selon des standards de représentation permettant leur partage et leur exploitation par toute personne ou application automatisée, et ainsi une *« révolution faite de nouvelles possibilités »* d'usages. En ce qui concerne les données textuelles, la mise en œuvre d'un tel projet repose en grande partie sur une publication des contenus où la représentation des éléments informatifs véhiculés par le texte est explicitement spécifiée : dans notre exemple, cette représentation se traduit par les indications selon lesquelles la mention *Ban Ki-moon* réfère à une personne et est décrite par une biographie localisée par une URI (`encyclo.org/Ban_Ki-moon`). Cette représentation explicite est principalement envisagée par l'adjonction de métadonnées aux documents publiés, que l'enrichissement de contenus visé par l'AFP formule en soulignant la valorisation de la production ainsi réalisée. Comme l'illustre en effet la figure 1, les métadonnées sont le support de l'intégration des contenus dans un espace en réseau tel que le Web, et ainsi d'une diffusion reposant sur des modalités sophistiquées, dépassant la simple livraison et limitée aux documents eux-mêmes. Il s'agit là d'une orientation stratégique adoptée par nombre d'acteurs présents sur le Web, à titre commercial ou non, et formant une lignée dans laquelle l'AFP peut s'insérer de façon fructueuse. Les développements récents menés notamment par la BBC [Kob+09] ou le NYT [Wool0] en sont des exemples particulièrement pertinents.

Ces métadonnées concernent en premier lieu, dans cet exemple et dans notre contexte de travail, les entités mentionnées dans les documents publiés par l'AFP (personnes, organisations, lieux) : il s'agit en effet d'éléments informatifs centraux, autour desquels s'organisent principalement les événements rapportés dans l'actualité. Elles forment ainsi un ensemble sujet à l'acquisition de connaissances, que celles-ci soient encyclopédiques ou relatives à des domaines

particuliers. Les ressources constituées dans cette perspective sont dès lors visées par le processus d'enrichissement qui en intègre la valeur informative et valorise ainsi les contenus traités. On observe à ce titre que la mise en œuvre du Web Sémantique dans la décennie écoulée s'est traduite de façon prégnante par la constitution et la mise à disposition dans un vaste réseau informatif d'ensembles de connaissances, désigné par le terme de *Linked Data*[1] et concernant notamment les entités visées ici, par opposition à un véritable développement de pratiques de publication systématiquement associées à l'introduction de métadonnées. La taille et la diversité des Linked Data est communément illustrée par sa représentation sous forme de graphe, donnée à la figure 2.

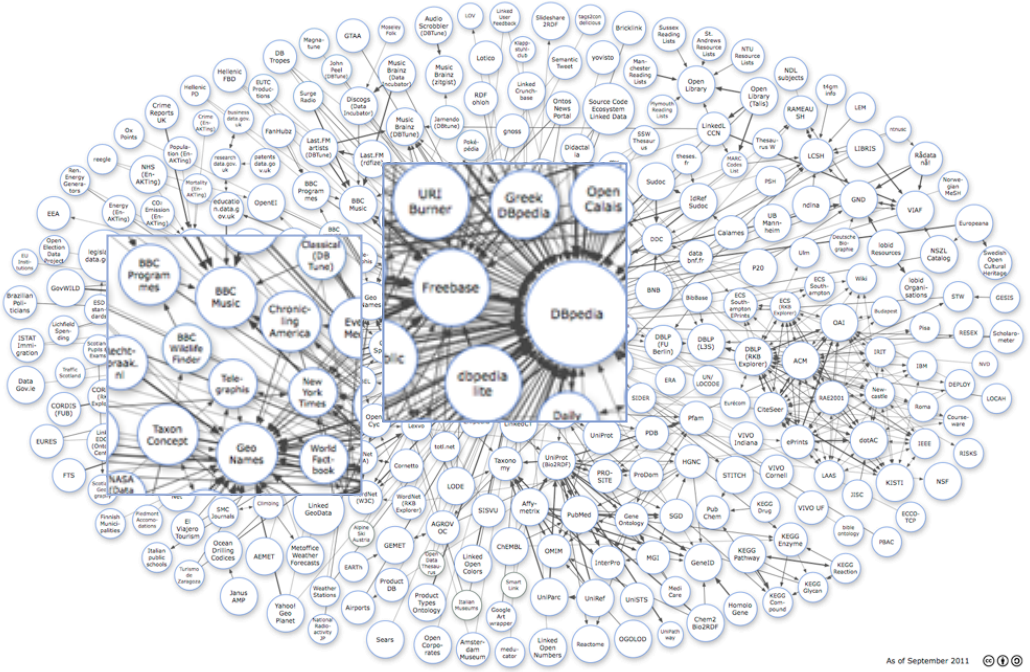


FIGURE 2 : Graphe des Linked Data.

Sémantique et enrichissement Si l'on peut ainsi comprendre de quel domaine est issue la formulation par l'AFP d'un besoin d'*enrichissement* des contenus, ainsi que le rôle central joué par les métadonnées telles qu'elles sont définies dans ce contexte, il reste à déterminer de quel *sens* relève cet enrichissement, caractérisé de *sémantique*. Son domaine de définition, le Web Sémantique, étant associé au même adjectif, il est utile de comprendre quelles idées et fonctionnalités ses initiateurs entendent par l'usage de ce terme. La définition de Berners-Lee, donnée plus haut, rappelle à ce titre que le Web Sémantique se présente avant tout comme un programme suivant lequel le Web dans sa version historique, sous le nom de World Wide Web (WWW), doit être refondé autour des principes de partage et d'intégration des connaissances. Cette ambition a notamment pour but de dépasser les limites inhérentes au WWW en termes d'accès à ces connaissances, disponibles dans leur grande majorité sous la forme de données textuelles distribuées au

1. `linkeddata.org`

travers de documents. Ces données atteignent en effet des quantités et une diversité de domaines telles qu'il est en pratique impossible, sous cette forme non structurée, d'en exploiter de façon systématique, exhaustive et efficace le contenu informatif. Le programme du Web Sémantique consiste à déterminer les modalités selon lesquelles la publication de nouveaux contenus, mais également la prise en charge des contenus existants, peuvent donner lieu à une mise à disposition des données sous une forme adaptée au partage et à l'intégration. La condition nécessaire identifiée à cet égard est celle d'une spécification explicite du *sens* des données, qui demeure implicite et inaccessible aux machines dans les données textuelles. Plus généralement, les données sous toutes leurs formes, non structurée dans le cas du texte ou structurée dans le cas de bases de données, par exemple, se présentent comme ininterprétables, autrement dit démunies de sens, pour les machines, dès lors qu'aucune sémantique ne leur est associée dans la perspective d'un traitement automatique. Dans l'exemple donné à la figure 1, l'association explicite de la mention *Ban Ki-moon* au terme PERSON par un attribut de balise nommé `type` n'a de pertinence, au niveau automatique, que si des programmes consommateurs de ces données disposent d'instructions liées à cette information de type et à l'entité mentionnée (rechercher la fiche biographique de la personne mentionnée dans les ressources adéquates, par exemple). À partir de la table de base de données suivante (figure 3), un traitement automatisé consistant par exemple à identifier les personnes employées par l'entreprise ABC Ltd. après 1999 implique, parmi d'autres conditions, que les champs `Arrivée` et `Départ` soient spécifiés comme relevant du type `date`, d'une part, et qu'un mode de calcul adapté à ce type de données soit défini, d'autre part, dans le programme utilisé :

Identifiant	**Nom**	**Fonction**	**Arrivée**	**Départ**
1	John Doe	PDG	1987-09-01	1996-03-15
2	Karl Hieronymus	DF	1992-10-10	2002-15-03
3	Tom Flinstone	DRH	2001-02-01	2009-12-06

FIGURE 3 : Structuration en base de données : table regroupant des informations sur les employés de l'entreprise ABC Ltd.

La sémantique ainsi recherchée s'entend donc principalement dans une acception relevant de l'informatique et des algèbres associant types de données et procédures. Le Web Sémantique fait le constat d'une hétérogénéité rédhibitoire dans la définition des modèles de données sur le Web d'une part, et de la simple absence de structuration au niveau des données textuelles, majoritaires sur le Web, d'autre part. Cette situation détermine l'une des orientations principales du Web Sémantique, qui consiste à initier et à encadrer la création et le développement de standards, essentiellement sous la forme de langages et de pratiques de publication de données. Ces standards visent à permettre à la large communauté concernée par la publication et la gestion documentaire de former un réseau cohérent et global, dans lequel la sémantique associée aux données est formalisée de façon explicite et interprétable automatiquement. Cet effort de standardisation constitue l'autre pan majeur du développement du Web Sémantique, en parallèle à la constitution massive d'ensembles de données dans les Linked Data.

Web Sémantique et organisations L'enrichissement de contenus s'inscrit dans le paradigme du Web Sémantique tant dans la perspective d'une valorisation de la production de l'AFP par leur mise en relation avec l'espace de connaissances ainsi construit sur le Web, que dans celle de l'adoption de pratiques, méthodes et technologies dérivant du Web Sémantique et applicables à toute forme d'organisation documentaire. Les principes d'échange et d'intégration de données à travers différents usages et applications, ainsi que de standardisation pour la cohérence et la facilité des traitements automatisés concernent en effet toute organisation, institutionnelle ou

commerciale, à visée publique ou privée, dont l'activité est liée à la production de données et en particulier sous leur forme documentaire et textuelle. C'est ce que montre notamment Wood [Woo10], qui souligne à travers plusieurs cas applicatifs l'efficacité de l'adoption du modèle architectural du Web, de la standardisation des modèles ainsi que de la pratique des Linked Data au niveau des entreprises et institutions. Ces stratégies permettent en effet de remédier à l'hétérogénéité caractérisant souvent les données des organisations, divisées en départements, services et silos de production employant des modes de représentation et de manipulation des données spécifiques et difficilement réutilisables, y compris au sein d'une même organisation. Ce constat est valable pour l'AFP, qui dispose de ressources dites *référentielles*, utilisées dans les différentes rédactions (texte, photographie. . .) principalement dans des objectifs d'indexation de la production. La mise en œuvre de l'enrichissement des contenus de l'AFP à l'aide de métadonnées pourra interroger les usages de ces ressources et la possibilité de leur mise en conformité avec les standards du Web Sémantique.

Web Sémantique, TAL et Extraction d'Information

Comme évoqué précédemment, la méthode principale d'explicitation de la représentation des données au sein des contenus textuels envisagée dans le cadre du Web Sémantique consiste à munir ces contenus de métadonnées ancrées sur les éléments sélectionnés pour leur valeur informative. Le terme d'*Annotation Sémantique* (AS) est employé pour désigner cette méthode, soulignant le caractère indissociable du texte et des métadonnées. Afin de fournir la sémantique attendue par les traitements élaborés dans le cadre du Web Sémantique, l'AS relie les éléments annotés à des ressources définies selon les standards du Web Sémantique, telles que les ensembles des Linked Data. Les métadonnées ainsi obtenues sont donc à la fois ancrées dans les contenus et dans des modèles dont la sémantique est spécifiée de façon standardisée et accessible.

Il s'agit, dans l'objectif d'un Web renouvelé tel que le présente Berners-Lee, de faire émerger les connaissances humaines à l'œuvre dans les productions de données afin de les rendre exploitables par des agents automatiques. Cette ambition vaste se heurte à la lourdeur de la tâche d'Annotation Sémantique, et par extension à l'impossibilité pratique d'annoter l'ensemble des contenus existants. Comme le souligne Wilks [Wil08], l'automatisation de l'Annotation Sémantique constitue un pré-requis fondamental à la réalisation du Web Sémantique, à moins d'ignorer les contenus existants et de n'envisager la production de données que sous une forme structurée, ce qui paraît difficilement acceptable. Dans le contexte de l'enrichissement de contenus de l'AFP, cette nécessité apparaît de façon prégnante devant la taille de la production textuelle, organisée sous forme de flux suivant le temps réel de l'actualité.

En suivant notamment l'argumentation de Wilks et Brewster [WB09], on observe que l'enjeu de l'Annotation Sémantique en tant que composant du Web Sémantique réside dans la possibilité d'automatisation de la compréhension des données langagières : faire émerger la sémantique des informations véhiculées au sein des contenus textuels revient à savoir de quoi ou de qui l'on parle, ce que les connaissances déjà acquises nous apprennent sur les éléments ainsi identifiés et quelles inférences peuvent être faites de leurs nouveaux usages. Le Web Sémantique se présente en ce sens comme une reformulation des objectifs de l'intelligence artificielle et du TAL, visant historiquement à atteindre une compréhension automatique du langage naturel. C'est notamment à ce titre que Wilks ([Wil08 ; WB09]) justifie le rôle crucial dévolu au TAL dans la réalisation du Web Sémantique : l'acquisition et la structuration des connaissances qui y est visée ne peuvent, selon lui, être menées qu'à partir des contenus eux-mêmes, donnant ainsi la base empirique nécessaire à la pertinence et la cohérence des résultats obtenus.

Au-delà de la parenté ainsi observable entre les objectifs aux fondements du Web Sémantique et du TAL comme dimension de l'intelligence artificielle, on constate une relation pratique et

méthodologique liant de façon étroite ces deux domaines : la recherche d'une structuration des données textuelles constitue en effet le cœur de l'Extraction d'Information (EI), développée dans le TAL depuis plusieurs décennies. Abandonnant progressivement l'objectif de compréhension globale des années 1960, elle suit au cours des années 1980 et 1990 une simplification progressive devant la nécessité de performances opérationnelles, aboutissant à une définition stabilisée de ses différents aspects et des résultats de bonne qualité. L'EI consiste principalement à opérer, à partir de données textuelles, la mise en correspondance d'éléments informatifs avec un modèle, la nature des éléments à extraire et ce modèle étant préalablement déterminés pour la tâche. Cette mise en correspondance vise à fournir une représentation structurée des informations ainsi collectées à d'autres traitements, tels que la fouille de données ou le résumé automatique.

Le parallèle entre AS et EI est dès lors manifeste : dans les deux cas, il s'agit d'aboutir à une représentation structurée des données textuelles à traiter, par l'adoption d'un modèle sur lequel s'appuie la tâche. La distinction entre AS et EI relève donc principalement de la nature des modèles associés à ces deux tâches : spécifiés formellement et explicitement selon les standards du Web Sémantique pour la première, ou définis relativement à une tâche et un domaine d'application pour le second, sans contrainte de formalisation externe. Cette distinction porte également sur la relation entretenue par ces deux tâches avec les données traitées. L'EI consiste fondamentalement à extraire des éléments des données textuelles, afin de les présenter sous une forme structurée ; elle intervient après la production de ces données, dans une tâche relevant de l'analyse. L'AS se place quant à elle dans la perspective d'une simultanéité par rapport à la production des données, qu'il s'agit de publier, diffuser ou exploiter à l'aide des métadonnées résultant de l'annotation. Elle implique cependant que les éléments à annoter fassent l'objet d'une identification préalable au sein des contenus traités : on peut ainsi considérer l'AS comme une forme d'EI ou alternativement, l'EI comme un composant nécessaire à l'automatisation de l'AS.

Ces différents constats et observations sur la nature des traitements envisagés dans le cadre du Web Sémantique et leur parenté avec le TAL, en particulier la structuration des données textuelles par l'EI, nous permettent de proposer une réponse pratique au besoin d'enrichissement de contenus formulé au départ de notre travail. L'AS se présente en effet comme un moyen adéquat pour la mise en œuvre de cet enrichissement, les annotations qu'elle produit correspondant aux métadonnées recherchées. L'EI s'intègre dans la méthodologie proposée en tant que composant permettant à l'AS de s'appliquer aux éléments informatifs identifiés au sein des contenus.

L'identification d'entités : proposition de définition méthodologique

Une fois identifiée l'Annotation Sémantique (AS) comme moyen de réaliser l'enrichissement des contenus textuels de l'AFP et l'EI comme composant méthodologique de cette approche, la question du traitement spécifique à accorder aux entités dans ce cadre se pose. Les éléments ciblés par l'enrichissement sont en effet les entités telles que les personnes, lieux et organisations mentionnées dans l'actualité, autour desquelles s'organise une grande partie de l'information véhiculée. Il s'agit de comprendre comment faire passer ces mentions au statut de métadonnées par l'AS, autrement dit quel modèle permet leur ancrage dans une sémantique adéquate.

Entités et entités nommées Les entités concentrent une grande partie des travaux réalisés en TAL et en EI, qui se penchent particulièrement sur leurs réalisations linguistiques désignées par le terme *entités nommées* (EN). Spécifiquement, la tâche de Reconnaissance d'Entités Nommées (REN), apparue et définie lors des campagnes d'EI telles que MUC [GS96], procède à la détection des EN au sein des données textuelles ainsi qu'à leur classification dans une typologie propre à chaque tâche de REN. Les types communs regroupent les personnes, organisations et lieux,

ainsi que les dates, montants ou événements, notamment. Il apparaît donc d'autant plus pertinent d'intégrer l'EI et plus particulièrement la REN à l'AS dans notre contexte dans l'objectif de traiter spécifiquement les entités. Cette intégration montre cependant une certaine limitation des modèles de structuration propres à la REN : la sémantique attendue par les traitements d'enrichissement est celle des entités en tant qu'individus extra-linguistiques, identifiables de façon univoque et descriptibles par un certain nombre d'attributs et liens avec d'autres entités. Cette sémantique se place hors du langage lui-même : les individus sont les objets *dénotés* par les EN, auxquelles la REN attribue une sémantique d'ordre typologique (PERSONNE, par exemple). Il s'agit donc de proposer un modèle d'ancrage des EN ou *mentions* d'entités qui détermine de façon précise et explicite quelle est l'entité dénotée par une mention donnée.

Dans l'exemple donné à la figure 1, la métadonnée ancrée sur la mention *Ban Ki-moon* spécifie que la mention est associée au type PERSON défini dans une ressource indiquée par l'URI qui préfixe le nom de ce type (`afp.com/metadata/concepts/`). Une telle spécification relève de l'AS, puisque la mention est ainsi ancrée dans un modèle explicite de représentation. Une procédure de REN permet d'arriver à ce résultat, si les types définis dans ce modèle sont accessibles à la tâche de reconnaissance. Un tel ancrage est en revanche insuffisant si l'on souhaite, comme c'est le cas dans notre contexte de travail, savoir de *qui* l'on parle, et non seulement de quelle sorte d'entité le segment annoté est une mention. Le second attribut caractérisant la mention *Ban Ki-moon* dans notre exemple (`resource="encyclo.org/Ban_Ki-moon"`) spécifie quant à lui la localisation d'une *description* associée de façon univoque et explicite à l'entité mentionnée. Cette seconde spécification relève elle de la référence : elle établit un lien référentiel entre une mention et une représentation de l'entité, autrement dit de l'individu dénoté par la mention.

On peut ainsi qualifier de *sémantique référentielle* la sémantique recherchée dans le cadre de l'enrichissement de contenus autour des entités. La mise en œuvre de cet enrichissement consiste alors à associer la tâche d'AS à des ressources appropriées à cet égard : conformes aux standards du Web Sémantique et intégrées aux Linked Data, ces ressources doivent présenter des descriptions d'entités identifiées de façon explicite, en un nombre et une diversité à même de couvrir la production de l'AFP.

Dénotation et identification d'entités Si les ressources développées selon les modèles proposés par le Web Sémantique permettent à la tâche d'AS de disposer des descriptions d'entités pouvant être mentionnées dans la production de l'AFP, il reste à établir de quelle façon les mentions repérées peuvent être mises en relation avec les entités qu'elles dénotent, autrement dit comment peut être établi automatiquement le lien de référence entre ces mentions et les individus recensées dans les ressources à disposition. La problématique rencontrée ici est celle de la dénotation, où les mentions, relevant du niveau linguistique et textuel, sont employées pour référer à des individus extra-linguistiques. Il s'agit donc de déterminer les facteurs permettant de lier une mention donnée à l'entité qu'elle dénote, ces facteurs relevant de la relation entre le *sens* de la mention dans son contexte d'occurrence et les descriptions d'entités accessibles.

Les développements technologiques liés à l'AS n'abordant que de façon relativement superficielle cette question, il paraît utile de nous appuyer dans la réalisation de notre tâche sur les travaux de recherche scientifique et académique intégrant pleinement la problématique de la dénotation. Il s'agit en particulier de recherches initiées par la campagne d'évaluation de TAC nommée KBP (*Knowledge Base Population*, Population de Bases de Connaissances ou PBC), tenant une édition annuelle depuis 2009 et dont les spécifications et résultats sont notamment rapportés par McNamee et Dang [MD09], Ji et al. [Ji+10] et Ji et al. [JGD11]. La tâche de PBC y est définie comme un renouvellement de l'Extraction d'Information, où une base de connaissances peuplée d'entités décrites formellement et identifiées de façon unique remplace les modèles traditionnels de structuration en Extraction d'Information. Deux sous-tâches sont spécifiées dans le cadre de

la PBC : *Entity Linking* et *Slot Filling*. La première, que nous proposons de nommer *Liage d'Entités*, consiste à identifier à quelle entrée de la base de connaissances réfère une mention textuelle d'entité indiquée dans un document. La seconde demande de collecter diverses informations concernant les entités de la base à partir des documents mis à disposition et de les intégrer à cette base, sous la forme de valeurs d'attributs pré-définis (date de naissance, fonction occupée dans une entreprise...). La représentation des entités dans une base de connaissances sur le modèle de la PBC est similaire à celle des ressources développées dans le cadre du Web Sémantique. La tâche de Liage d'Entités peut dès lors faire l'objet d'une intégration dans la méthodologie d'AS envisagée ici. Ses spécifications ainsi que les méthodes proposées dans le cadre de la PBC permettent en effet de développer une approche systématique de la résolution du phénomène dénotationnel à l'œuvre dans le processus d'ancrage des métadonnées au niveau des mentions d'entités.

Les différentes composantes de traitement à l'œuvre dans la méthode d'enrichissement introduite ici forment une chaîne opérationnelle dont nous proposons de désigner le résultat par le terme d'*identification d'entités*. Les métadonnées obtenues à l'issue de cette chaîne correspondent en effet à des mentions mises en relation avec des représentations d'individus, de nature à élargir l'espace informatif véhiculé par les documents ainsi annotés. Cette élargissement se fait bien à partir d'entités identifiées, autrement dit explicitement désignées et décrites au sein de ressources également spécifiées.

Problématiques liées au cas d'usage Le processus d'enrichissement de contenus de l'AFP implique le traitement de corpus bruts, où il s'agit à la fois de repérer les éléments devant donner lieu à des métadonnées (les mentions d'entités) et d'établir les liens référentiels existant entre ces mentions et les entités définies dans les ressources d'AS adoptées pour la réalisation de cette tâche. L'intégration du Liage dans cette méthodologie, spécifiquement au niveau de son second composant (établissement des liens), se heurte au problème de l'application séquentielle de plusieurs modules d'automatisation : la REN permettant dans un premier temps de repérer les mentions d'entités peut en effet produire des erreurs, notamment la détection de faux positifs, dont la propagation au module suivant nuit à la qualité générale de l'enrichissement obtenu. Nous proposons à cet égard une prise en charge explicite de la propagation d'erreurs afin d'assurer une précision satisfaisante des performances de la tâche d'enrichissement. Il s'agit notamment de modifier la méthodologie générale du Liage d'Entités adoptée dans le cadre de la PBC, puisque celle-ci porte sur la mise en relation d'entités et de mentions préalablement isolées dans les données de référence, où la détection automatique n'est donc pas un enjeu. Cette modification donne lieu à une approche jointe des deux composants de l'AS : REN et Liage, où le second intègre des critères d'évaluation des résultats du premier.

Parallèlement à l'élaboration de cette approche jointe, différentes configurations séquentielles de la méthodologie proposée peuvent être envisagées, notamment afin de vérifier l'hypothèse selon laquelle plusieurs modules d'analyse en cascade peuvent bénéficier de la conservation d'ambiguïté à certains niveaux et ainsi améliorer la qualité des résultats finalement obtenus.

Contributions

Afin de mettre en œuvre une réponse pratique et fonctionnelle au besoin d'enrichissement de contenus formulé par l'AFP, nous proposons une méthodologie d'Annotation Sémantique adaptée au traitement de données textuelles brutes. Cette méthodologie intègre un niveau de sélection des éléments à même de constituer des métadonnées, ainsi qu'une approche systématique du phénomène dénotationnel et de l'établissement des liens référentiels entre éléments annotés et

entités représentées dans les ressources adoptées. La relation entre ces deux niveaux, se traduisant par une possible propagation d'erreurs, y est explicitement prise en charge.

La méthodologie proposée est concrètement réalisée par le système Nomos, développé dans le cadre de la tâche d'enrichissement présentée ici. La conception de ce système inclut la constitution des ressources nécessaires à son développement et à son évaluation (données de référence), ainsi que de la base de connaissances concernant les entités cibles des liens référentiels.

La tâche d'AS accomplie par Nomos nécessite par ailleurs la mise à disposition de ressources de recensement et de description des entités ciblées par l'enrichissement. La constitution de telles ressources a été menée parallèlement à la mise en place de la méthodologie adoptée ici et au développement du système Nomos.

À un niveau plus général, la tâche d'enrichissement se traduit par l'intégration d'outils de TAL à la chaîne de production de l'AFP, permettant à leur suite l'application de Nomos pour l'identification des entités. Cette intégration implique notamment l'adaptation d'outils issus de la recherche académique à la configuration industrielle du traitement de données de l'AFP, qui vise une qualité et une efficacité conformes aux attendus d'un contexte de production.

L'intégration de la production de l'AFP au paradigme et à l'espace de publication du Web Sémantique, visée par le Medialab avec l'enrichissement de contenus textuels, se traduit également par la conception d'une ressource référentielle propre à l'Agence : il s'agit de collecter, représenter et maintenir les métadonnées produites à partir de l'enrichissement de contenus dans une structure conforme aux standards du Web Sémantique et des Linked Data, autrement dit un modèle de nature ontologique dont les métadonnées jugées pertinentes pour le métier de l'AFP constituent la population. La conception, la création ainsi que la population initiale de ce modèle référentiel forment ainsi une partie des travaux réalisés dans le cadre de la thèse présentée ici.

Enfin, l'enrichissement de contenus à l'aide de métadonnées devant permettre des formes renouvelées d'exploitation de la production, notre travail a abordé deux cas applicatifs à cet égard : la Recherche d'Information reposant notamment sur l'indexation par entités, et la détection automatique de citations avec attribution d'auteurs, où l'identification d'entités joue un rôle crucial.

Organisation du mémoire

Chapitre 1 : L'enrichissement de contenus et le paradigme du Web Sémantique
Nous proposons tout d'abord une exploration du contexte de définition de l'enrichissement de contenus, sous la forme d'une description du Web Sémantique, de son état actuel de développement ainsi que des principaux standards de représentation dont il est à l'origine. La relation entre le Web Sémantique et l'effort de compréhension automatique du langage naturel y est notamment abordée. Cette approche du Web Sémantique permet de constater sa filiation avec l'évolution historique des principes d'organisation documentaire, dont les organisations telles que l'AFP sont des lieux de mise en œuvre. Ces principes permettent notamment d'éclairer le rôle assigné aux métadonnées de documents dans notre contexte de travail.

Chapitre 2 : L'Extraction d'Information : jalon méthodologique pour l'enrichissement de contenus textuels
La parenté entre Web Sémantique et traitement du langage naturel est ensuite étudiée du point de vue de l'Extraction d'Information, dont le Web Sémantique constitue une forme de prolongement dans la nature des objectifs poursuivis. À la suite d'une présentation générale de l'Extraction d'Information et de la nature des modèles qui la caractérisent, nous étudions le rôle crucial de l'Extraction d'Information dans le traitement des entités à travers la tâche de Reconnaissance d'Entités Nommées. Notre analyse souligne notamment les limitations des modèles employés

dans ce cadre quant à la sémantique référentielle qu'il s'agit d'associer aux entités dans la tâche d'enrichissement à l'aide de métadonnées.

Chapitre 3 : Annotation sémantique et identification d'entités
Nous menons une étude générale de l'Annotation Sémantique et des ressources sur lesquelles elle s'appuie et soulignons la nécessité d'une intégration de l'Extraction d'Information à cette tâche dans la perspective de son automatisation. L'Annotation Sémantique doit être complétée par une approche systématique du problème de l'identification des entités, qui fait l'objet de la tâche de Liage d'Entités proposée dans le cadre de la conférence TAC (*Text Analysis Conference*). Nous en présentons les problématiques et les spécifications, ainsi que les méthodologies proposées pour sa mise en œuvre.

Chapitre 4 : Expression de besoins : enrichissement de contenus textuels pour l'AFP
Le processus d'enrichissement formulé par l'AFP s'inscrit dans un cadre de développement stratégique mené par divers acteurs des medias autour du Web Sémantique, que nous illustrons par deux cas d'usage existants, ceux de la BBC et du *New York Times*. Nous présentons ensuite l'organisation de la production de l'AFP, dans laquelle doit s'intégrer le processus d'enrichissement traité ici. Nous décrivons les spécifications, contraintes et objectifs de l'enrichissement envisagé dans le cadre spécifique de cette organisation.

Chapitre 5 : Approche de l'identification d'entités dans les contenus textuels de l'AFP
L'approche générale que nous proposons d'adopter pour la réalisation de l'enrichissement des contenus de l'AFP est décrite dans le chapitre 5, d'abord sous ses aspects méthodologiques puis au travers des ressources, connaissances et outils qu'elle implique. Nous soulignons notamment les problématiques liées au traitement de données textuelles brutes, nécessitant une prise en charge de la propagation d'erreurs dans une configuration de traitements en cascade. Nous introduisons également la possibilité de conservation d'ambiguïtés au sein de différentes étapes d'analyse. La constitution d'une base de connaissances associée aux entités cibles de notre tâche est décrite de façon étendue. Nous présentons également les outils de Reconnaissance d'Entités Nommées intégrés à notre approche, ainsi qu'un système d'identification développé durant notre travail selon une approche différente ; il s'agira notamment de le comparer au système Nomos, présenté au chapitre suivant.

Chapitre 6 : Un système d'identification d'entités : Nomos
Nous présentons le système d'identification automatique d'entités Nomos, qui repose sur l'approche, les ressources, connaissances et outils décrits au chapitre précédent. Les différentes configurations possibles du système sont spécifiées puis évaluées à la suite d'expériences dont nous rapportons les modalités et les résultats.

Chapitre 7 : Applications : acquisition de métadonnées et enrichissement de dépêches AFP
Les travaux applicatifs réalisés durant ce travail de thèse sont présentés dans le chapitre 7. Il s'agit d'abord de la création et de la population d'une base référentielle destinée à rassembler et représenter de façon structurée, selon les standards du Web Sémantique adoptés dans le processus d'enrichissement, les métadonnées considérées comme pertinentes et utiles pour le métier de l'AFP et le traitement de ses contenus. Nous avons également exploré l'orientation d'une tâche de Recherche d'Information vers les entités, qui constitue un volet applicatif important de l'identification et de l'enrichissement. Nous décrivons enfin un système de détection automatique de citations avec attribution d'auteurs, pour lequel l'identification d'entités joue un rôle crucial, ainsi que les modalités de son intégration concrète dans une application mise en production par l'AFP et utilisée par le journal *Libération*.

Conclusion et perspectives
Nous conclurons la présentation de notre travail par l'exploration de différents points restant à traiter relativement à la tâche d'enrichissement et à l'identification automatique d'entités. Il s'agit notamment de la résolution des dénotations métonymiques, ainsi que de la prise en charge du phénomène de la nouveauté touchant les entités. Ce second point concerne en particulier le problème de la synchronisation et de l'enrichissement dynamique des ressources référentielles adoptées.

Collaborations

Les outils d'analyse linguistique automatique utilisés dans ce travail, en particulier la chaîne de traitement SxPipe et le système de Reconnaissance d'Entités Nommées LIANE, sont développés dans le cadre de recherches académique et distribués sous licence libres, en permettant l'utilisation commerciale. SxPipe est principalement développé et maintenu au sein de l'équipe-projet Alpage (INRIA et Université Paris 7) par Benoît Sagot et distribué sous licence CeCILL-C. LIANE est issu de travaux menés dans les laboratoires LIA et LIF par Frédéric Béchet et disponible sous licence GNU-GPL.

Nous avons participé à l'élaboration de la ressource d'Annotation Sémantique Aleda employée dans les systèmes NPNORMALIZER et Nomos, principalement développées et maintenues par Benoît Sagot (Alpage). Le module NPNORMALIZER, intégré à la chaîne SxPipe, est également développé et maintenu par Benoît Sagot. Nous avons réalisé son intégration dans les traitements de contenus de l'AFP ainsi que rapports d'erreurs et de corrections nécessaires à son amélioration.

Les annotations manuelles de données utilisées dans les différents travaux présentées ici ont été réalisées par les journalistes du Medialab de l'AFP (Denis Teyssou, Dominique Ferrandini et Bernard Apfeldorfer) et par nous-même. La validation manuelle des résultats de traitements automatiques présentés au chapitre 7 a été réalisée par les journalistes du Medialab.

Nous avons réalisé le système de détection automatique de citations à partir de la chaîne SxPipe, après de travaux de recherche portant notamment sur l'étude des verbes de citations menés en collaboration avec Laurence Danlos et Benoît Sagot (Alpage). Le module de résolution d'anaphores intégré à ce système a été conçu et développé par Myriam Majdoub dans le cadre d'un stage de six mois pour l'obtention d'un Master 2 professionnel de l'Université Paris 7 (cursus Linguistique Informatique) effectué au Medialab sous notre encadrement.

L'intégration sous forme de prototype des travaux réalisés à partir des outils d'analyse linguistique dans la chaîne de production de l'AFP a été réalisée par Bertrand Goupil, ingénieur au Medialab.

Chapitre 1

L'enrichissement de contenus et le paradigme du Web Sémantique

La production de données textuelles en grande quantité associée à la diffusion de l'information sous forme numérique constitue un enjeu important du développement des pratiques de publication, notamment pour une organisation telle que l'AFP. L'enrichissement de contenus à l'aide de métadonnées se présente comme une évolution majeure et indispensable au renouvellement de ces pratiques : ces métadonnées rendent possibles l'interprétation et l'exploitation automatique des contenus en langage naturel par des applications diverses, ouvrant ainsi des perspectives de traitement sophistiqué de l'information. À ce titre, le Web Sémantique constitue un cadre de présentation du processus d'enrichissement à deux niveaux. D'une part, il apparaît comme un paradigme contemporain de publication dont l'influence sur les acteurs concernés est en constante augmentation, en particulier au sujet de la structuration des données à large échelle ; dans ce paradigme, l'enrichissement à l'aide de métadonnées joue un rôle central. D'autre part, le Web Sémantique représente par et pour lui-même un espace de publication qu'une organisation comme l'AFP se doit d'investir au même titre que d'autres diffuseurs d'information, publics et privés. Ce rattachement à l'espace du Web Sémantique se fait principalement par le truchement des métadonnées, support de l'intégration et du partage d'information proposé par le Web Sémantique.

C'est donc en premier lieu une description de ce que recouvre et représente à ce jour le Web Sémantique qui sera faite, afin de donner un ancrage pratique à la problématique de l'enrichissement (1). Il sera ensuite possible d'examiner comment la production documentaire, qui constitue l'activité centrale de l'AFP, peut relever des schémas de structuration proposés par le Web Sémantique, notamment en termes de classification (2). La notion de *sémantique* attachée au Web sera alors discutée dans le cadre de ce type de production documentaire et des applications recherchées par l'inscription d'organisations et d'entreprises dans un tel modèle (3).

1 Le Web Sémantique

La dénomination de *Web Sémantique* ne présente pas toujours de contours clairs ni de définition précise quand à l'objet qu'elle désigne et suscite peu de propositions de définitions dans le cadre académique. On peut en effet s'interroger sur sa nature : base de données, réseau, programme informatique ou énonciation de pratiques logicielles peuvent tour à tour le qualifier, à l'image des divers points de vue, partiels et hétérogènes, des communautés s'y intéressant. Il est en tout cas concerné par le Web dans sa forme actuelle, celle du World Wide Web, dont la maturité et l'universalité sont à la fois facteurs de limitations quant au traitement de l'information qui y est déposée et sources de potentialités pour de nouveaux développements. Le Web Sémantique

peut dès lors être envisagé comme une extension et une redéfinition du Web en lui-même, ainsi que des pratiques et technologies présidant à sa création et à son évolution, où la notion de *sémantique* constitue un noyau fondateur tout en réduisant la couverture conceptuelle de ce terme.

1.1 Évolution du World Wide Web vers une pratique de l'interprétabilité

1.1.1 Le World Wide Web entre potentialités et limitations

Depuis sa mise à disposition du public au début des années 1990, le World Wide Web (WWW), communément identifié comme le *Web*, constitue un lieu de dépôt et de partage d'information, en grande majorité sous forme de documents textuels, en constante augmentation. À la manière d'une gigantesque bibliothèque universelle, le Web rassemble une quantité inédite de connaissances selon une architecture distribuée et minimalement contrainte. La production, le dépôt et le partage de documents sur le Web ne sont en effet réglementés ni pris en charge par aucune autorité centrale, mais laissés aux communautés et individus utilisateurs, les technologies nécessaires se fondant sur les principes de consensus et de standardisation.

La masse de contenus, notamment textuels, hébergée par le Web est le reflet tant d'une forte activité humaine de publication de connaissances que d'une grande variété de domaines, points de vue et degrés de spécialisation caractérisant cette publication. Son intérêt réside autant sinon plus dans son organisation sous forme interconnectée que dans la quantité d'informations disponibles : les documents publiés sur le Web sont en effet reliables entre eux par le méchanisme des liens hypertexte [RFC866] et localisables par celui des URI (Uniform Resource Identifier) [RFC1738], tous deux proposés par Tim Berners-Lee dans l'initiative ayant mené à la mise en place du WWW.

La richesse et la disponibilité de ces contenus, en constante augmentation, place le Web dans le paradigme de la Recherche d'Information (RI), puisque l'espace documentaire à explorer afin d'obtenir une information particulière dépasse les capacités d'un utilisateur humain. Concernant les contenus textuels, c'est donc sur le principe de l'indexation des documents sur des unités ou *mots* qui permet la mise en œuvre d'une RI fondée sur la pertinence de description des *mots-clés* pour accéder à l'information. Le mécanisme des liens hypertexte est loin d'être absent de ce paradigme, puisque l'algorithme de RI le plus efficace et populaire aujourd'hui intègre une interprétation de ces liens pour le calcul de l'ordonnancement des pages du Web [Pag+99].

Le regroupement sans précédent de connaissances que constitue le Web est au centre des réflexions sur une nouvelle forme de réseau d'information : le modèle de développement du Web, distribué et en réseau, augmentant continuellement les quantités de données et les domaines concernés, mène en effet à une limitation inhérente. Il devient souvent en pratique malaisé voire impossible de trouver l'information recherchée avec une précision satisfaisante, en raison de l'explosion en termes de volume de données, mais également de diversité des sources d'information et des représentations associées. Bien que le Web continue de se présenter comme un espace de recherche d'information incontournable, son utilisation comme outil de réponse à des questions — largement répandue comme emploi étendu de la RI, la réponse cherchée se trouvant généralement dans les premiers documents retournés par les moteurs de recherche les plus populaires — se heurte aux problèmes d'ambiguïté et de dissolution des informations dans de multiples silos non connectés.

Comme le souligne Horrocks dans plusieurs exemples de cette limitation du Web [Hor08], chercher des renseignements sur une personne devient une opération complexifiée par le nombre possiblement élevé d'homonymes, que l'indexation par mots-clés ignore en tant que phénomène pertinent pour l'information retournée ; obtenir une liste des noms de chefs d'État des pays de l'Union Européenne peut se révéler impossible en un nombre minimal de requêtes, ce genre d'informations existant sur le Web mais en de multiples lieux et sans schéma de représentation

unifié.

Il reste cependant que le Web constitue une source d'informations qui ne peut être ignorée dans la formulation d'un *nouveau* Web, le Web Sémantique : c'est à partir du contenu du Web, principalement sous forme de documents en langage naturel, que le Web Sémantique peut se concevoir en tant qu'ensemble de techniques et de pratiques permettant d'utiliser les connaissances véhiculées par ces contenus. Cette opération d'extension et d'exploitation ne peut s'envisager que par la mise en place de protocoles et le développement de technologies faisant passer les contenus documentaires de leur forme textuelle, inaccessible aux machines, à un ensemble de données interprétables automatiquement. L'idée définitoire du Web Sémantique, formulée par Horrocks [Hor08] est celle d'une réponse à un problème, celui de la localisation et de l'intégration de l'information parmi l'ensemble documentaire réuni sur le Web ; cette réponse réside dans la production *« d'annotations à la sémantique accessible aux machines »* , cette sémantique *« donnant un sens formellement défini aux termes utilisées dans les annotations. »*

1.1.2 Une information compréhensible par les machines

La formulation d'un renouvellement du Web sur la base de contenus munis d'une *sémantique* accessible aux machines remonte à la fin des années 1990 et se trouve énoncée dans l'article « The Semantic Web » publié en 2001 par Tim Berners-Lee publié dans la revue *Scientific American* [BLHL01]. L'auteur y rassemble ses réflexions et arguments pour un Web renouvelé et réalisant toutes ses potentialités en tant que vecteur d'information. On y trouve la description anticipatoire d'*« une nouvelle forme de contenus Web dotés de sens pour les ordinateurs »* , qui *« lancera une révolution de nouvelles possibilités. »* Ces dernières sont présentées sous les traits d'agents automatiques capables d'exploiter les connaissances accessibles sur l'espace du Web afin de réaliser un certain nombre de tâches, complexes et interreliées, jouant ainsi le rôle d'assistants personnels sur lesquels il serait possible de s'appuyer dans le futur. Au cœur de cet objectif se trouve la nécessité de rendre accessibles les informations et connaissances placées dans les contenus Web aux machines, et ainsi de leur donner une sémantique.

Dans cette perspective, le Web Sémantique est présenté comme une extension du Web existant, dans lequel les documents publiés le sont non plus seulement à destination des humains mais aussi d'agents automatiques. Afin que le Web Sémantique puisse fonctionner ainsi, il apparaît nécessaire de mettre à disposition des sources d'information structurées, ainsi que des règles d'inférence permettant aux ordinateurs de mener des raisonnements sur ces informations. Il s'agit donc d'ajouter une forme logique aux contenus du Web, en s'appuyant principalement sur le déploiement d'ontologies pour la définition de concepts et de règles d'inférence et sur le mécanisme référentiel des URI pointant vers ces ontologies.

Par cet exposé d'objectifs, la vision du Web Sémantique initiée par Berners-Lee, que l'on retrouve dans la formulation de Horrocks [Hor08] donnée plus haut, souligne l'enjeu majeur auquel les technologies idoines doivent répondre : il s'agit d'opérer une migration à partir de données non structurées, principalement le texte contenu dans les documents du Web, vers une représentation de l'information structurée et interprétable grâce à une modélisation des connaissances accessible aux machines. Cette migration peut s'entendre comme le transfert de contenus existants vers le statut de données munies d'une sémantique explicite, ou bien comme la production de nouveaux contenus suivant cet objectif. C'est en fonction du type de réponse apportée à cet impératif que s'organisent les différentes approches et réflexions autour du Web Sémantique, selon que l'accent est mis sur la création et la manipulation des données constituant l'information, sur les représentations conceptuelles sous-jacentes ou sur le passage entre langage naturel et données structurées.

1.2 Le Web Sémantique en réalisations

En 2006, un second article dont Tim Berners-Lee est co-auteur reprend la description définitoire du Web Sémantique sous le titre « The Semantic Web Revisited » [SHBL06] et nuance déjà fortement les évidences formulées en 2001 au sujet des services devant naturellement et aisément découler d'une nouvelle vision du Web. Il y est reconnu que la transformation d'un Web limité à l'accès par moteurs de recherche en une globalité munie de sens et capable de réaliser des inférences est loin d'être réalisée. Les scénarios esquissés avec optimisme en 2001 n'émergeant pas comme escompté, les auteurs refondent la vision originale du Web Sémantique en s'appuyant sur les progrès tout de même réalisés en la matière, arguant que le besoin d'intégration des données et d'une sémantique de l'information se fait toujours plus crucial et manifeste. Ils soulignent notamment que le développement des standards et conceptualisations nécessaires à la réalisation du Web Sémantique est largement entamé, avec l'existence d'un grand nombre d'initiatives d'intégration de connaissances à l'aide d'ontologies prouvant le bien-fondé du projet originel.

Les avancées les plus probantes se situent donc au niveau du développement de standards d'échange et principalement du langage RDF[1], qui permet une représentation des connaissances minimaliste mais efficiente pour un usage dans le cadre du Web, ainsi que de l'utilisation des URI. Le langage de définition d'ontologies OWL[2], plus expressif que RDF, tient également bonne place dans les avancées mises en vedette, même si le coût de développement des ontologies devant sous-tendre le Web Sémantique est présenté comme un frein. Les ontologies dites profondes (SUMO[3], DOLCE[4]), qui concentrent d'importants et coûteux efforts humains, font l'objet d'une remise en question en raison de la difficulté à les faire fonctionner à l'échelle du Web et à atteindre une universalité de conceptualisation pour toutes les utilisations. Afin d'ancrer leur vision dans une perspective concrète et solide, les auteurs évoquent le rôle important des *communautés*, notamment scientifiques, dans le développement du Web Sémantique, comme cela a été le cas avec le WWW : le coût de développement des ontologies pourra baisser à mesure que le nombre d'utilisateurs concernés augmentera. La définition d'ontologies légères (*shallow ontologies*, par opposition aux ontologies profondes) pour tout type de domaine et de besoin est présentée comme un moteur important des développements à venir, du fait de leur simplicité (peu de concepts et de relations définis) s'équilibrant avec une capacité à fournir de grandes quantités de données. C'est finalement surtout du côté de la définition des règles d'inférence et de leur support logiciel que le Web Sémantique peine encore en 2006 à se concrétiser.

L'autre pan des avancées connues par le Web Sémantique est celui de la publication de données, parallèlement aux efforts de conceptualisation et de partage pour l'intégration. Suivant une série de bonnes pratiques énoncées dans le cadre du Web Sémantique[5], les organisations et individus producteurs de contenus se proposent d'étendre aux données les principes de publication de documents, c'est-à-dire de les mettre à disposition des utilisateurs du Web sous une forme interconnectée, créant ainsi un vaste réseau de ressources informatives, toujours selon l'architecture typiquement distribuée du Web. Sous le nom de Linked Data[6], ces ressources atteignent une taille et une présence parmi des producteurs et utilisateurs suffisamment importantes pour que le Web soit fréquemment rebaptisé *Web de données*, par opposition à sa version historique, reposant sur l'unité documentaire.

La réalisation pratique du Web Sémantique telle qu'elle émerge aujourd'hui s'organise ainsi autour de trois objets centraux :

1. Resource Description Framework `http://www.w3.org/RDF/`
2. Ontology Web Language `http://www.w3.org/2004/OWL/`
3. `http://www.ontologyportal.org/`
4. `http://www.loa.istc.cnr.it/DOLCE.html`
5. `http://www.w3.org/DesignIssues/LinkedData.html`
6. `http://linkeddata.org/`

1. les *documents*, principale source d'information et de connaissances (exemple : un article en ligne sur le site Web du NYT),

2. les *ressources Web*, pouvant représenter tout concept, objet ou entité, référencées par le mécanisme des URI et préférablement liées à une *description* d'ordre ontologique (la page Wikipedia de l'article consacré à Barack Obama, président des États-Unis, par exemple)

3. les conceptualisations partagées sous forme d'*ontologies*, jouant le rôle d'ancrage sémantique pour les ressources référencées dans le réseau des Linked Data.

Documents, Linked Data et ontologies constituent le squelette du Web Sémantique par un processus d'annotation sémantique au niveau des documents, les annotations pointant vers des éléments des Linked Data, eux-mêmes ancrés en tant qu'instances dans les ontologies correspondantes, qui en fournissent les descriptions conceptuelles et relationnelles utiles à des traitements automatiques. Ce processus d'annotation revient à munir les documents concernés de *métadonnées* sémantiques et concentre une grande partie des travaux de recherche consacrés au développement du Web Sémantique.

Avant d'examiner plus en profondeur la mise en œuvre de l'Annotation Sémantique de contenus textuels, il apparaît utile de donner un aperçu des standards d'échange et de représentation créés pour le Web Sémantique et utilisés pour la production de métadonnées, en particulier RDF et OWL. Afin de situer le niveau d'intervention de ces langages, le schéma usuellement proposé, notamment par le W3C, pour la description structurelle du Web Sémantique est reproduit à la figure 1.1. Parmi les nombreux niveaux concernés par les efforts de développement de la communauté du Web Sémantique, allant du texte aux raisonnements logiques abstraits et à la notion de confiance associée aux traitements effectués, on retrouve le niveau documentaire, soit celui des données, dans les deux niveaux inférieurs (URI et Unicode pour le texte, XML et espaces de nom pour leur formattage). Les trois briques centrales (RDF dans ses différentes syntaxes et *ontologie*) constituent le niveau de représentation et de description de l'information visé par l'Annotation Sémantique. La position de la brique « URI » correspond à l'annotation : celle-ci se fait au niveau des données, à l'aide d'URI renvoyant aux Linked Data, elles-mêmes ancrées dans le niveau de représentation ontologique, qui fait l'objet du développement des langages RDF et OWL. Les niveaux supérieurs correspondent aux processus d'inférence envisagés à partir des traitements effectués en amont et demeurent encore peu spécifiés mais figurent d'ores et déjà sur ce diagramme programmatique.

1.2.1 Resource Description Framework (RDF)

Ainsi que le rappelle Horrocks dans [Hor08], le processus d'Annotation Sémantique de contenus Web cherche à répondre au problème d'intégration et de partage de l'information. L'annotation est réalisée au niveau local, celui d'un document, quand l'objectif d'intégration demanderait à pointer vers des ressources partagées, autrement dit non locales. L'intérêt d'annotations locales, uniquement valides pour le cadre de production des contenus considérés, serait en effet limité et un ancrage dans une représentation de l'information définie de façon externe et partagée devient dès lors nécessaire.

Le langage RDF, correspondant à une suite de recommandations du W3C publiées en 2004[7], permet la description structurée de ressources, au sens du Web et évoquées en 1.2, ainsi que des relations entre ces ressources. Son composant principal, les URI, permet d'établir une référence directe vers toute ressource Web non locale à la description considérée, remplissant ainsi la condition d'intégration de l'information. Pour faire référence à la femme politique américaine Hillary

7. `http://www.w3.org/RDF/`

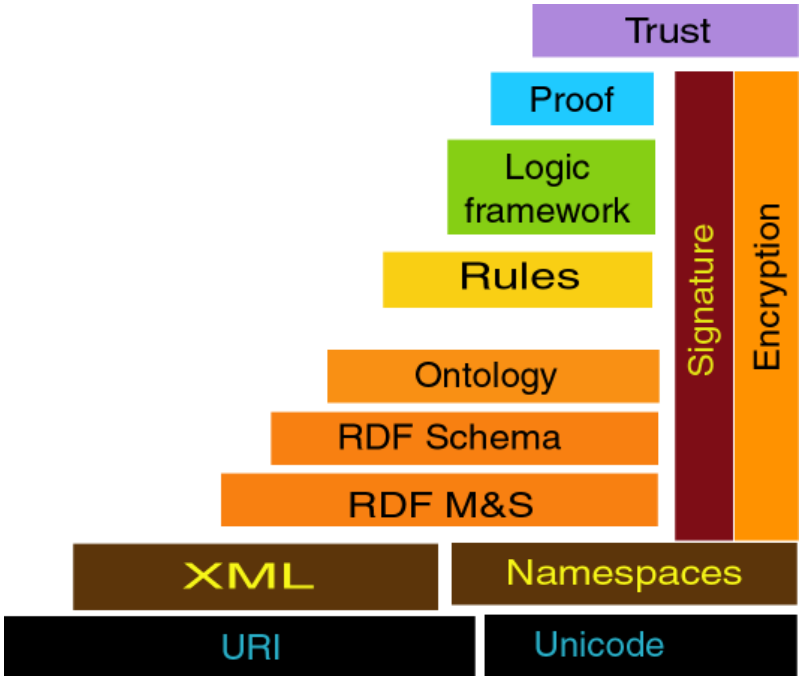


FIGURE 1.1 : Schéma usuel de description structurelle du Web Sémantique.

Clinton, on pourra utiliser l'URL[8] `http://politike.org/person/HillaryRodhamClinton`, abrégée dans la suite en `HillaryClinton`.

RDF est donc un mécanisme de représentation distinct du langage naturel, permettant de représenter des graphes orientés et étiquetés dont les arcs sont décrits sous forme de triplets : sommet d'origine, étiquette de l'arc et sommet final modélisant des relations prédicatives entre sujets et objets. Ainsi, on peut exprimer la fonction actuelle de Hillary Clinton et sa relation avec Bill Clinton, identifié par l'URI `http://politike.org/person/WilliamJeffersonClinton`, abrégée dans la suite en `BillClinton`, par les triplets suivants, correspondant au graphe de la figure 1.2 :

```
HillaryClinton isUSsecretaryOf StateDepartment

HillaryClinton isMarriedTo BillClinton
```

où `HillaryClinton` est le sujet, `isUSsecretaryOf` et `isMarriedTo` sont les prédicats, et `StateDepartment` et `BillClinton` sont les objets des relations ainsi définies. Sujet et objet peuvent avoir la forme d'une URI ou d'un nœud vide, c'est-à-dire un nœud sans étiquette ou une ressource non identifiée ; l'objet peut également consister en une valeur littérale (chaîne, entier...). Quant au prédicat, qualifié ici de *propriété*, il s'agit toujours d'une URI, telle que `http://politike.org/relation/isUSsecretaryOf`.
Sous sa forme sérialisée en XML, un triplet aura la forme suivante :

```
<rdf:Description rdf:about="#HillaryRodhamClinton">
<isUSsecreataryOf>StateDepartment</isUSsecreataryOf>
<isMarriedTo rdf:resource="#WilliamJeffersonClinton"/>
</rdf:Description>
```

8. Les URL (Uniform Resource Locator, `http://www.w3.org/TR/url/`) sont un type particulier d'URI, spécifiant le domaine ainsi que le chemin de localisation d'une ressource, sur le Web ou tout autre espace de données.

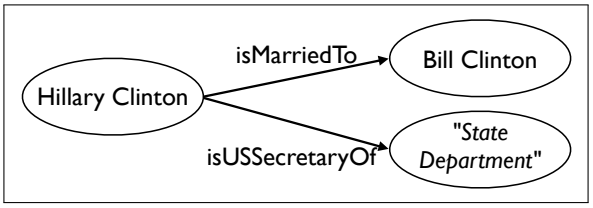


Figure 1.2 : Exemple de graphe RDF.

RDF fournit ainsi un mécanisme d'annotation structurée, auquel il reste encore à donner une sémantique, c'est-à-dire un moyen d'interpréter les annotations considérées. Dans les précédents exemples, le sens des propriétés spécifiées doit en effet se trouver défini parmi un ensemble de termes accessible aux annotations *via* le mécanisme des URI. RDF présente un certain nombre de ressources spéciales permettant ce type de définition, telles que les propriétés `rdf:class`, `rdf:subClassOf`, `rdf:domain` ou `rdf:range`. Cette extension du langage, nommée RDF Vocabulary Description Language ou RDF Schema, correspond au besoin, pour le Web Sémantique, d'un formalisme dans lequel exprimer le sens de termes nouveaux à partir de ceux déjà définis, afin de fournir les vocabulaires munis de sens nécessaires à l'interprétation des annotations. On pourra alors déclarer, relativement aux exemples précédents, une classe conceptuelle `Person`, ainsi que `Politician` comme sous-classe de `Person`, et l'appartenance de l'individu `HillaryClinton` à cette sous-classe par les triplets :

```
Politician rdf:subClassOf Person
HillaryClinton rdf:type Politician
```

qui permettront d'inclure Hillary Clinton dans des résultats de recherche portant sur des politiciens mais également des personnes en général.

1.2.2 Ontologies pour le Web Sémantique

Afin de répondre à ce besoin d'expression de sens, le Web Sémantique a recours aux ontologies comme dépôt et véhicule de vocabulaires de termes extensibles et dotés de sens bien défini. Sujet de nombreuses recherches et études scientifiques, les ontologies constituent avant tout, dans le cadre du Web Sémantique, un outil formel et fonctionnel. Bien que ne formant pas l'objet central du présent travail consacré à l'enrichissement de contenus, elles apparaissent néanmoins dans tout compte-rendu de travaux relatifs au Web Sémantique, que ce soit en tant que sujet principal ou comme accessoire aux réflexions et développements présentés. Il faut en effet nécessairement parler des ontologies en tant que choix décisif de représentation de l'information et des connaissances sur le Web Sémantique, même si nous nous limitons en la matière en proposant une description synthétique de leurs aspects les plus pertinents pour notre cadre de travail.

Définitions : philosophie versus informatique Employé au singulier et dans le domaine de la philosophie, le terme *ontologie* désigne la discipline s'attachant à déterminer quelles entités et quels types d'entités existent et ainsi d'étudier la structure du monde, réel ou non. Cette étude est indépendante de toute perspective particulière ou de considérations ultérieures et se distingue ainsi des sciences expérimentales. Remontant à Platon et Aristote, l'ontologie se fonde sur le développement de catégorisations hiérarchisées des différents types d'entités et des attributs qui les distinguent ; sa représentation prototypique est celle de l'arbre de Porphyre, reproduit en figure 1.3, du nom du philosophe néoplatonicien du III^{e} siècle après J.C. En informatique, une ontologie constitue un artefact d'ingénierie invoquant une représentation de l'existence dans

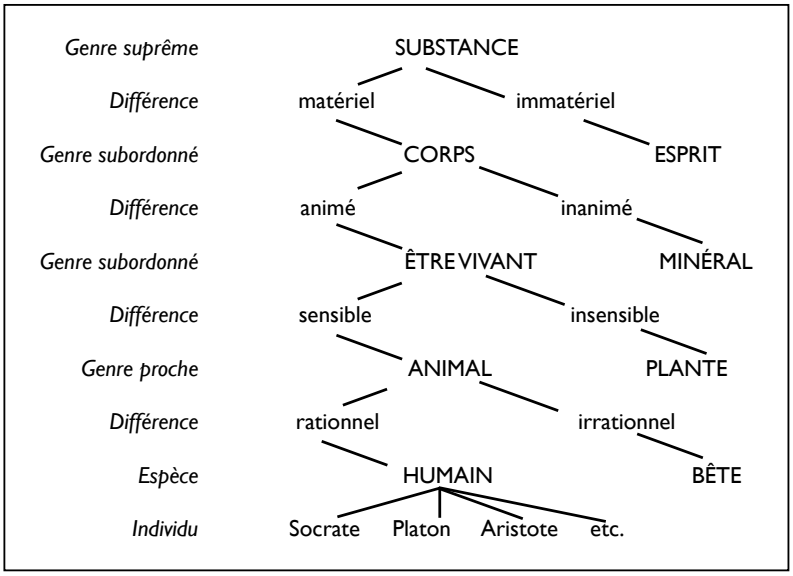


FIGURE 1.3 : Arbre de Porphyre.

une perspective pragmatique : il s'agit d'un moyen de modélisation d'un aspect du monde, usuellement désigné comme *domaine*, dont l'observation fait émerger un certain nombre d'entités et de relations pertinentes à sa représentation. Comme le souligne Horrocks [Hor08], ce processus de spécification n'est pas étranger à l'ontologie philosophique puisqu'elle implique souvent une classification fondée sur un type d'information similaire aux attributs de différence générique de l'arbre de Porphyre. Une classe Politicien peut ainsi être décrit comme une sous-classe de Personne sur la base du trait distinctif de la pratique professionnelle de la politique.

À la suite de Guarino et al. dans [GOS09], nous employons la définition donnée par Studer et al. [SBF98], formulée ainsi :

> [1.1] Une ontologie est une spécification formelle et explicite d'une conceptualisation partagée.

Notions définitoires Cette définition fait intervenir trois notions principales mises en exergue :

Conceptualisation Selon la formulation de Genesereth et Nilsson [GN87], il s'agit d'une vue abstraite et simplifiée du monde que l'on souhaite représenter dans un but particulier. Tout système reposant sur des connaissances implique un choix de conceptualisation, qu'il soit explicite ou implicite.

Spécification formelle et explicite La spécification d'une conceptualisation constitue l'expression des faits qui la constituent. Elle est nécessairement explicite dans la mesure où elle doit garantir que seuls les modèles attendus par une conceptualisation sont retournés par l'interprétation des symboles utilisés pour la spécification. Une ontologie spécifie explicitement une conceptualisation par un ensemble d'axiomes la décrivant et contraignant ses interprétations de façon intensionnelle. Ces axiomes doivent être donnés dans un langage formel afin de permettre leur accessibilité par les machines, ce qui exclut le langage naturel. Le degré de formalisation sous-jacent au langage adopté détermine le penchant du compromis entre puissance d'expressivité et efficacité sur le plan de la décidabilité et du raisonnement.

Partage Bien qu'informelle, la notion de partage dans la définition des ontologies est déterminante : le bénéfice de leur usage reste concrètement limité si la conceptualisation proposée ne correspond pas à une compréhension minimalement généralisée. Les ontologies sont en effet, dans leur usage, un instrument destiné la communication dans le champ des conceptualisations considérées et impliquent à ce titre un processus référentiel portant sur les entités et relations en jeu. Le partage est intégré aux ontologies par la simplification et la clarification de ce processus, où le sens est déterminé de façon logique et précise plutôt que laissé à l'arbitraire et l'ambiguïté.

Typologie fonctionnelle et usage du Web Sémantique À partir de ces notions définitoires, une typologie fonctionnelle des ontologies peut être dressée, suivant la réinterprétation par Brewster et O'Hara dans [BO07] de Davis et al. [DSS93] : cinq fonctions d'une représentation des connaissances, ici sous la forme d'ontologies, sont considérées :

- *Substitution* aux objets du monde représenté. Le degré de fidélité de la représentation dépend de ce que l'ontologie capture et omet dans ce processus.
- Ensemble d'*engagements ontologiques* : critère nécessaire à toute théorie et formulé par Quine [Qui80], il reflète les éléments du monde choisi comme pertinents pour sa représentation. Ce choix nécessaire est en même temps le moyen de remédier à la complexité du monde dans le processus de représentation.
- *Théorie fragmentaire du raisonnement*. En tant que mode représentation de connaissances, une ontologie traduit une approche particulière du processus de raisonnement humain. Ainsi que l'énonce Minsky [Min74], « Chaque type de problème requiert un type de pensée et de raisonnement approprié, ainsi qu'un type de raisonnement approprié. » Toute représentation ontologique rend compte du choix relatif au type de raisonnement pour un domaine donné en adoptant un formalisme approprié.
- *Outil de traitement informatique*. Le but fonctionnel d'une ontologie est le traitement informatisé de problèmes relatifs à des domaines particuliers, ce qui pose le problème de l'efficacité computationnelle. Le formalisme adopté pour une ontologie doit garantir cette efficacité, ce qui conduit nécessairement à une limitation de l'expressivité permise.
- *Moyen d'expression pour l'humain*. Comme toute forme de représentation, une ontologie constitue un moyen d'expression pour l'humain, ainsi que de communication entre l'humain et la machine.

Ces différents aspects de la représentation ontologique de domaines participent de l'usage des ontologies dans le Web Sémantique. Sur chacun d'eux, cet usage se traduit par un certain nombre de limitations : il s'agit en effet principalement de permettre le partage de données, par le biais d'un accord sur les portions de représentation les concernant. La prétention n'est pas, dans ce cadre, celle d'un consensus plus large ni d'une cohérence et d'une complétude globale des modèles utilisés. Les ontologies rattachées au Web Sémantique sont donc fréquemment essentiellement destinées à structurer formellement, avec une profondeur minimale pour l'expression de concepts, les données déposées sur le Web.

OWL, un langage d'ontologies pour le Web Afin de doter les moyens de réalisation du Web Sémantique d'un langage ontologique plus expressif que RDF, le W3C crée en 2001 un groupe de travail pour la mise au point d'un tel langage, suivant l'objectif usuel de standardisation du Web, qui aboutit en 2004 à la recommandation du standard OWL. Une syntaxe RDF est spécifiée

pour ce langage, ce qui permet de rendre les ontologies ainsi définies accessibles sur le Web par le mécanisme des URI. Les conditions à remplir pour un langage ontologique, rappelées par Antoniou et Harmelen [AH09] auxquelles répond OWL, sont les suivantes :

- Définition d'une syntaxe, de la même façon que pour les langages de programmation, afin de permettre des traitements automatiques. La syntaxe de OWL est celle de RDF et RDFS.
- Définition d'une sémantique, non subjective et non ambigüe, permettant des traitements automatiques et notamment le raisonnement.
- Support efficace de raisonnement, portant sur :
  - l'appartenance à une classe : si x est une instance de la classe C, et si C est une sous-classe de la classe D, alors x est une instance de D ;
  - l'équivalence de classes : si la classe A est équivalente à la classe B et si la classe B est équivalente à la classe C, alors A est équivalente à C ;
  - la cohérence de l'ontologie : la déclaration commune de la classe A comme sous-classe de la classe B et des classes A et B comme disjointes mène à une incohérence ;
  - la classification : si des paires attribut-valeur sont déclarées comme conditions suffisantes d'appartenance à la classe A, si l'individu x satisfait ces conditions, alors x est une instance de A.
- Puissance d'expression suffisante.
- Facilité de l'expression.

La sémantique d'un langage ontologique permet le support nécessaire au raisonnement, où les dérivations ci-dessus peuvent être réalisées mécaniquement, et non manuellement, afin de vérifier l'intégrité et la cohérence d'une ontologie, ainsi que de classifier automatiquement de nouvelles instances. Cette caractéristique des ontologies est usuellement permise par l'ancrage du langage adopté dans un formalisme logique ; OWL correspond partiellement aux logiques de description et peut ainsi faire usage des raisonneurs existants pour ce formalisme (FaCT [9], RACER [10]).

Comme RDF, OWL présente les primitives de modélisation ontologique permettant d'organiser le vocabulaire correspondant en hiérarchie typées : classes et relation de sous-classe, propriétés et relation de sous-propriété, restrictions de domaine et de portée des propriétés, instanciation de classes. Les limitations de RDF sont abrogées par la mise à disposition d'un plus grand nombre de constructeurs, dont la description exhaustive est mise à disposition dans plusieurs documents de référence du W3C [11]. Le compromis entre cette plus grande richesse d'expressivité et le maintien d'une efficacité de raisonnement suffisante est notamment pris en charge par la définition de trois sous-langages de OWL : OWL Full permet l'utilisation de toutes les primitives du langage et est entièrement compatible avec RDFS, mais ne garantit pas la décidabilité et ne permet donc pas l'utilisation de raisonneurs ; OWL DL, en spécifiant des restrictions sur les constructeurs utilisables, permet l'expressivité maximale tout en maintenant la complétude et la décidabilité, et peut donc faire usage des raisonneurs adaptés ; OWL Lite réduit encore les constructeurs pour la spécification de hiérarchies minimales et peu contraintes. L'adoption du langage OWL dans sa

9. www.cs.man.ac.uk/ horrocks/FaCT/
10. www.sts.tu-harburg.de/r.f.moeller/racer/
11. `http://www.w3.org/2004/OWL/` : Page générale d'informations sur OWL
`http://www.w3.org/TR/owl2-overview/` : Présentation générale d'OWL (version 2)
`http://www.w3.org/TR/owl2-syntax/` : Spécifications structurelles et syntaxiques complètes d'OWL (version 2)
`http://www.w3.org/TR/2012/REC-owl2-new-features-20121211/` : Présentation de l'évolution entre OWL 1 et OWL 2, notamment en ce qui concerne les différents profils OWL (OWL DL, OWL Lite, OWL Full)

version OWL DL pour la spécification d'une ontologie sera exemplifié au chapitre 7, où l'usage des constructeurs les plus courants ainsi que certains traits caractéristiques de l'ontologie ainsi obtenue pourront être discutés. L'exemple de la figure 1.2 peut être exprimé en OWL des façons suivantes, suivant la syntaxe adoptée pour la sérialisation XML (RDF/OWL ou OWL/XML) :

```
<owl:NamedIndividual rdf:about="&politike;HillaryRodhamClinton">
<rdf:type rdf:resource="&politike;Person"/>
<isUSsecretaryOf rdf:resource="&politike;StateDepartment"/>
<isMarriedTo rdf:resource="&politike;WilliamJeffersonClinton"/>
</owl:NamedIndividual>
```

```
<ObjectPropertyAssertion>
<ObjectProperty IRI="#isMarriedTo"/>
<NamedIndividual IRI="#HillaryRodhamClinton"/>
<NamedIndividual IRI="#WilliamJeffersonClinton"/>
</ObjectPropertyAssertion>
<ObjectPropertyAssertion>
<ObjectProperty IRI="#isUSsecretaryOf"/>
<NamedIndividual IRI="#HillaryRodhamClinton"/>
<NamedIndividual IRI="#StateDepartment"/>
</ObjectPropertyAssertion>
```

Construction et acquisition d'ontologies Deux types d'approche distinguent les méthodes de construction d'ontologies, selon que la conceptualisation dont il est question dérive d'un processus d'introspection et de spécification manuelle ou de données à partir desquelles des traitements automatiques infèrent une structure possible pour leur représentation. Il s'agit alors respectivement

- d'ontologies construites par des spécialistes du domaine considéré ou ontologies *expertes*, obtenues par une dérivation « de haut en bas » (« top-down ») ; les personnes en charge de la spécification constituent en effet la hiérarchisation conceptuelle et relationnelle destinée à être peuplée d'instances — entités et relations — vérifiant les axiomes énoncées par cette spécification ; il peut exister autant d'ontologies expertes que de projets industriels ou communautaires liés à un domaine particulier : logiciel (*Core Software Ontology* [OGS09]), anatomie (*Foundational Model of Anatomy* [RM08]) ou droit (*LKIF-core* [Hoe+07]) ;
- d'ontologies acquises à partir de corpus rassemblant les données représentatives du domaine (dérivation « de bas en haut »— « bottom-up ») ; des méthodes automatiques ou semi-automatiques permettent d'extraire les concepts pertinents ainsi que les relations qui les caractérisent ; leur organisation au niveau intensionnel, c'est-à-dire les axiomes de l'ontologie à constituer, peut être inférée grâce à des algorithmes de partitionnement ou l'application de patrons linguistiques ; la structure résultante est le plus souvent bruitée et nécessite une supervision humaine et des processus de révisions manuelles. Buitelaar et Cimiano [BC08] présente dans une étude récente et approfondie les problématiques et méthodes de l'acquisition automatique d'ontologies.

1.2.3 Linked Data

À l'image du Web Sémantique lui-même, les Linked Data (LD) constituent à la fois un regroupement d'objets identifiables dans l'espace du Web et un ensemble de pratiques de publication de

données mises en œuvre de façon distribuée. Il s'agit en effet, comme évoqué en 1.2, de données publiées par des organisations et individus producteurs de contenus afin d'en faire des ressources Web, de fait organisées en un réseau global. Les bonnes pratiques correspondantes, énoncées par Tim Berners-Lee [12] mais également discutées notamment par Heath et Bizer [HBl1] reposent sur les principes suivants :

- Accessibilité des données sur le Web sous licence libre, dans un format structuré et lui-même non-propriétaire
- Usage des standards libres du W3C pour l'identification des ressources (RDF, URI)
- Connexion des données à celles d'autres producteurs de données

En mettant ces principes en relation avec les modes de représentation des connaissances évoqués précédemment, c'est-à-dire les triplets RDF et les ontologies exprimées en OWL, on obtient un ensemble de données dont la localisation et l'identification par URI revient à l'expression d'une instanciation ontologique : chaque élément publié dans le cadre des LD correspond en effet à une instance définie dans une ontologie, que celle-ci émane de l'agent de publication ou que ce dernier emploie une modélisation tierce pour accomplir cette opération d'ancrage sémantique. Par le mécanisme des liens, cette instance peut être localisée au sein de plusieurs ontologies en donnant des définitions complémentaires. L'association des annotations en RDF, des ontologies et de leurs instances organisées dans les LD constituent donc la mise en œuvre principale du Web Sémantique à ce jour, sur laquelle peuvent s'appuyer les services et applications destinés à exploiter les connaissances du Web en vue de tâches automatisées. Une telle organisation permet en effet non seulement une accessibilité des données sur le Web, mais également, par le mécanisme des liens, une ouverture systématique de l'espace de connaissances disponibles. Lors de l'accès à une instance correspondant à une ontologie donnée et ainsi aux propriétés qui y sont définies pour elle, la spécification de liens vers d'autres instances équivalentes dans d'autres ontologies augmente le nombre et la sophistication de l'information à la portée de l'agent, humain ou automatique.

Les premiers acteurs de publication sur le réseau des Linked Data sont les organisations s'intéressant en tant que telles aux données et à leur maintenance sous forme de bases, telles que les encyclopédies en ligne (Wikipedia [13]), l'inventaire Freebase [14], ou DBpedia [15], version entièrement convertie en RDF de Wikipedia. De nombreuses organisations dans divers secteurs (presse, édition, politique...) publient également depuis plusieurs années tout ou partie des données pertinentes à leur activité sur le mode des LD, par exemple le journal *New York Times* [16]. De récentes avancées ont par ailleurs été menées au niveau des données administratives et gouvernementales, rendues de plus en plus accessibles par les États prenant part à la politique dite des « données ouvertes » (*Open Data*). ; associées à une publication sur le réseau des Linked Data, elles contribuent à former les Linked Open Data (LOD).

La figure 1.4 donne une représentation visuelle du graphe ainsi créé, dont la taille continue d'augmenter tant par le nombre de nœuds — chaque nœud représentant un agent de publication, et le plus souvent une modélisation ontologique du domaine concerné — que par celui des arcs reliant ces nœuds, correspondant à la mise en relation d'ensemble de données entre plusieurs agents et ontologies. Les données ainsi publiées sont recensées et documentées par la commu-

12. http://www.w3.org/DesignIssues/LinkedData.html
13. http://www.wikipedia.org
14. http://www.freebase.com
15. http://www.dbpedia.org
16. http://data.nytimes.com/

nauté des Linked Data et rendues accessibles *via* la plateforme de distribution Web CKAN[17] et un site dédié[18], qui compte à ce jour plus de 5 000 ensembles de données.

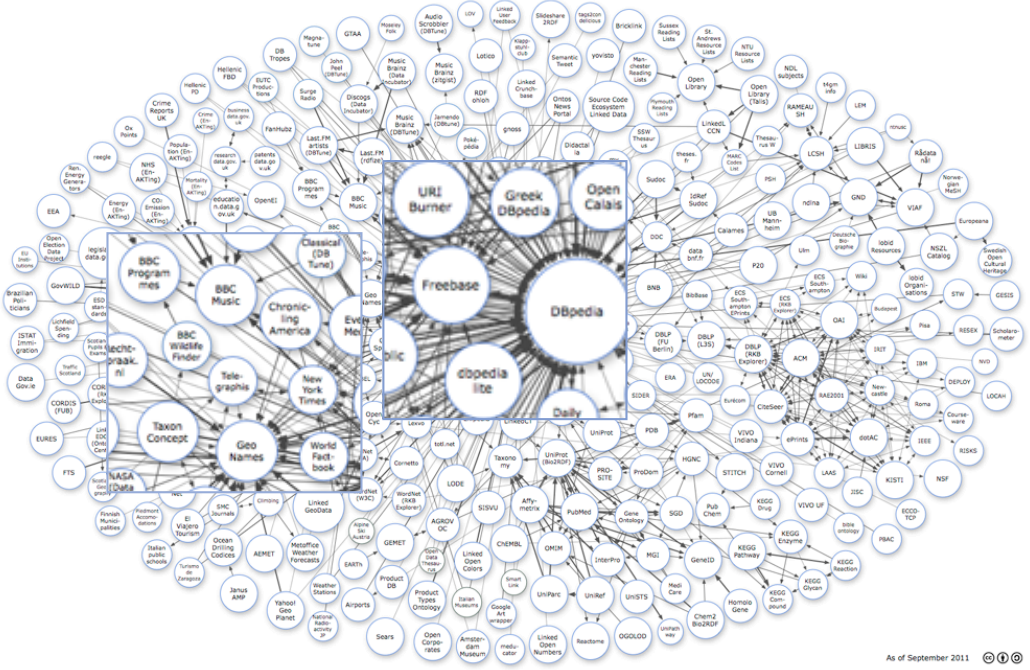


Figure 1.4 : Graphe des Linked Data : ressources publiées (nœuds) et liens existant entre ressources (arcs) (adapté de `http://linkeddata.org/`).

On peut finalement observer que l'importance et les dimensions de ce *Web de données* (cf. 1.2) relativisent la prégnance des ontologies dans le Web Sémantique selon sa définition de départ : présentées comme le cœur du processus de représentation sur laquelle la sémantique du Web devrait se fonder, leur fonction de formalisation semble reléguée au second plan, derrière la puissance d'inventaire quasi-universel des LD. Les modèles ontologiques sur lesquels reposent les ensembles de données des LD sont en effet moins exposés et exploités qu'attendu. Il reste cependant que la structuration de ces ensembles par le mode ontologique constitue la clé technologique sur laquelle peuvent se bâtir l'expansion et la fonctionnalité d'intégration des LD. On peut ainsi reformuler cette situation de prédominance des données sur les modèles en termes ontologiques, où le composant terminologique (aussi nommé « T-box ») est réduit un à faible nombre d'axiomes ou à une faible complexité conceptuelle et relationnelle, tandis que le composant assertif (aussi nommé « A-box »), autrement dit la population de l'ontologie ou ses instances, peut être de très grande taille. Cette configuration se retrouve par ailleurs dans les travaux centrés sur les *bases de connaissances*, non nécessairement définies relativement au Web et aux LD, où le focus des recherches porte souvent plus nettement sur les problèmes de population que de structuration en tant que telle ; ce thème sera notamment abordé au chapitre 3.

17. `http://ckan.org/`
18. `http://datahub.io/`

1.3 Annotation Sémantique, Intelligence Artificielle et TAL

1.3.1 Nécessité du lien entre contenus et représentations pour le Web Sémantique

L'intelligence artificielle rebaptisée Les accomplissements technologiques associés au renouvellement de la pratique de publication documentaire initié par le Web Sémantique justifient leur existence et leur mode de mise en œuvre par l'objectif d'une information accessible et compréhensible par les machines. Le but énoncé, en particulier dans [BLHL01], est celui d'agents automatiques capables d'accomplir des tâches diverses en exploitant les connaissances jusqu'alors uniquement réservées à la compréhension humaine. À la suite de Wilks et Brewster [WB09 ; Wil08], on peut observer qu'il s'agit là d'une reformulation, dans les termes contemporains du Web et des pratiques logicielles associées, des objectifs originels poursuivis par les recherches en intelligence artificielle (IA) au vingtième siècle. Si une continuité entre IA et Web Sémantique peut être constatée dans ces objectifs, Wilks observe cependant qu'elle n'existe que peu dans les lignes de recherches suscitées par le second. Le Web Sémantique encourage sur ce plan l'adoption de schémas de représentation simples, favorisant des traitements algorithmiques à la complexité limitée. Ce manque de sophistication peut contribuer à poser la question de la puissance de représentation permise par les outils du Web Sémantique. On peut en tout cas faire état d'une transition entre IA et Web Sémantique où la représentation des connaissances formelle et traditionnelle laisse place à l'adoption des ontologies comme outil central de représentation et de traitement.

Un manque d'annotations Si le Web Sémantique devait constituer une forme contemporaine d'IA, on ne peut que constater que son objectif de départ ne correspond pas à une réalité flagrante : de nombreux services ayant pour base le Web existent, mais ne témoignent pas encore d'un accomplissement généralisé de tâches par les machines comme cela était imaginé au début des années 2000 par Tim Berners-Lee. Le Web Sémantique actuel existe davantage en tant qu'ensemble d'avancées technologiques, et plus encore comme centre de gravité d'un effort de standardisation et de définition de pratiques modernisées. La carence la plus manifeste en ce qui concerne une réalisation concrète du Web Sémantique se situe au niveau des annotations : les données annotées selon les directives énoncées dans le cadre du Web Sémantique sont loin de constituer une masse visible dans l'espace de publication documentaire. Or, c'est bien l'Annotation Sémantique qui doit permettre une exploitation automatique par l'extraction de connaissances à partir de contenus, afin de constituer un ensemble d'informations sous une forme distincte du langage naturel. L'Annotation Sémantique de contenus, même si elle dispose désormais d'un certain nombre de conditions nécessaires pour sa mise en œuvre, n'est pas encore une norme généralisée dans la publication de données.

L'annotation : partie prenante du processus rédactionnel Ce constat nous ramène au problème de l'enrichissement de contenus textuels pour lequel le Web Sémantique constitue un cadre formel et pratique : il s'agit d'ancrer les contenus d'intérêt dans une sémantique permettant l'interprétabilité, afin de proposer une publication documentaire augmentée au niveau informatif et disponible pour des traitements ultérieurs sophistiqués. Cet ancrage dispose de la structure nécessaire à l'endroit des outils fournis par le Web Sémantique décrits précédemment (1.2) : le langage de description RDF, le mécanisme de référencement et de localisation des URI, ainsi que les ontologies pouvant être conçues pour tout domaine à l'aide du langage OWL. Il reste néanmoins à déterminer les modalités concrètes de mise en œuvre de l'Annotation Sémantique attendue pour un fonctionnement effectif du Web Sémantique : l'Annotation Sémantique se conçoit comme partie prenante du processus rédactionnel, et c'est en ce sens que le Web Sémantique en propose un renouvellement. Il apparaît dès lors que ce versant productif du Web Sémantique ne se situe pas au niveau du développement informatique et logiciel mais à celui de ses utilisateurs,

c'est-à-dire des auteurs des documents eux-mêmes. On observe en cela un trait typique du Web depuis sa conception originale, celui de la distribution, où les contenus et leur interconnexion se construisent par une intrication d'initiatives et de communautés plutôt que sous l'autorité d'un organe central.

Coûts et freins de l'annotation Le faible taux de données annotées révèle cependant que ce mode de pratique rédactionnelle est encore peu répandu, et ce pour des raisons tenant manifestement au coût qu'elle induirait. L'Annotation Sémantique de contenus demande en effet du temps aux rédacteurs qui en sont chargés, ce qui peut être en soi rédhibitoire. Mais elle réclame également la mise en place d'une structuration de l'information du côté des données publiées, qu'il s'agisse d'une organisation, publique ou privée, ou d'un rédacteur isolé : les annotations doivent établir des liens référentiels vers des ensembles de données existants, constitués en interne dans une ontologie de domaine par exemple, ou externes, comme DBpedia ou tout autre nœud des Linked Data pertinent pour les contenus considérés. Cette nécessité de structuration se traduit par un effort de développement technique non négligeable, mais également de conceptualisation sans lequel le processus d'annotation ne peut aboutir aux résultats escomptés. Construire et adopter un modèle de connaissances, puis le lier à des contenus en langage naturel en maintenant les impératifs de sens et de formalisation sont des opérations non triviales sur lesquelles il convient de mener une réflexion.

Périmètre de l'annotation Dans la perspective d'une Annotation Sémantique accomplie par les rédacteurs au moment de la production des contenus, elle peut être envisagée comme plus ou moins aisée, dès lors que les structures et outils adéquats sont mis à leur disposition — au moins une ontologie peuplée de cibles pour l'annotation ainsi qu'une interface de sélection et d'ajout de métadonnées au contenu liée à cette ontologie. Mais une telle réduction de l'Annotation Sémantique aux contenus produits à compter d'aujourd'hui semble impossible, puisqu'elle reviendrait à laisser de côté l'ensemble des contenus d'ores et déjà existants, sur le Web ou au sein d'organisations, c'est-à-dire à ignorer la majeure partie des connaissances déjà publiées. Si l'on considère les contenus du Web comme devant être traités de façon généralisée, le processus d'Annotation Sémantique nécessite la mise en œuvre de techniques et méthodologies adaptées, ce qui constitue un pan crucial des recherches à mener au sujet du Web Sémantique.

1.3.2 Les deux filiations du Web et de l'annotation sémantiques

L'Annotation Sémantique constitue donc un problème de nature double quant à sa mise en œuvre : vecteur fondamental de l'acquisition de connaissances pour un Web Sémantique fonctionnel, elle pose la question de l'expression du sens à partir du langage naturel vers une autre modalité de représentation ; destinée à l'acquisition du sens sur un très grand ensemble de contenus touchant à tous les domaines, elle demande à s'interroger sur les moyens de sa réalisation concrète à une si large échelle ainsi que sur les possibilités de validation de la sémantique ainsi constituée. Ces deux axes problématiques permettent d'envisager l'Annotation Sémantique dans une filiation avec, d'un côté, les réflexions menées en intelligence artificielle (IA) sur la place du langage et de la représentation des connaissances, et, de l'autre, la tradition d'annotation et d'analyse textuelle en traitement automatique du langage (TAL). Ces deux lignées entretiennent des relations historiques et fondamentales dans lesquelles se place l'Annotation Sémantique nécessaire à la mise en œuvre du Web Sémantique.

L'Annotation Sémantique invoque tout d'abord une dichotomie classique en IA, celui de la représentation des connaissances et de sa relation au langage naturel : il s'agit de savoir si les connaissances véhiculées par la communication linguistique peuvent trouver une représentation

dans un autre système symbolique, c'est-à-dire si le langage naturel ne tient que par et pour lui-même et si toute autre représentation est parasitaire ou insuffisante quant au sens exprimé. À l'inverse, cette relation interroge la nature parasitaire du langage lui-même et la nécessité de ramener le sens à des formalismes non linguistiques afin d'en avoir une connaissance exacte. Autrement dit, l'Annotation Sémantique se situe dans la problématique du recodage de contenus, dans laquelle la place et l'expression du sens dépend du degré de formalisation et d'autonomie accordé au langage naturel.

Par ailleurs, l'Annotation Sémantique promeut le document et son contenu textuel comme objet central dans le Web Sémantique, et se place ainsi dans la lignée du TAL, qui s'intéresse de façon primordiale à ces objets. Le principe de l'annotation de contenus textuels dans le but de formaliser de façon explicite les divers niveaux et types de connaissances qui y sont exprimés sous forme linguistique est en effet un élément constitutif des méthodes et objectifs du TAL. Ce principe est notamment réalisé par une association entre certains éléments textuels dans un document donné et des éléments de codage indiquant une information particulière au sujet des premiers. Les éléments de codage, autrement dit les annotations elles-mêmes, sont au moins vues comme une traduction plus formelle des éléments textuels en question.

Dans le Web Sémantique et à la suite des paradigmes de représentation des connaissances développés notamment dans le cadre de l'IA, cette formalisation repose sur un ancrage des éléments de codage dans un schéma de conceptualisation partagé et bien défini, c'est-à-dire dans une ontologie. D'autres modalités d'annotation sont toutefois envisageables, la plus prégnante à ce jour étant celle des *étiquettes* (en anglais « tags ») associées aux contenus sous la forme d'une indexation et largement répandue dans l'espace du Web et des services associés identifiés comme « Web 2.0 ». Dans cette pratique, l'annotation relève également des utilisateurs mais suivant un principe de libre choix et de collaboration, où l'ancrage sémantique est formé par une synthèse sociale plutôt que par une définition préalable d'un schéma partagé. On parle alors de *folksonomies*, qui de distinguent des conceptualisations formelles en ce qu'elles ne définissent pas de vocabulaire pour les représentations, qui conservent les caractéristiques du langage naturel — variations, ambiguïté... —, même sous la forme d'étiquettes. Il faut toutefois rappeler, comme le fait Wilks [Wil08] dans ses observations sur la validité de l'ancrage sémantique choisi pour une représentation formelle des connaissances à partir de contenus, que l'annotation doit être fondée empiriquement afin que le sens ainsi dérivé corresponde à une réalité justifiée par les usages humains.

Ainsi, pour Wilks et Brewster [Wil08 ; WB09], l'Annotation Sémantique doit s'effectuer sur la base d'ontologies construites de façon empirique, c'est-à-dire acquises à partir des contenus textuels eux-mêmes. Cette vision empiriste de l'ancrage sémantique pour le Web se ramène à la tâche d'acquisition automatique d'ontologies évoquée en 1.2.2, pour laquelle le TAL présente depuis plusieurs décennies des méthodes et techniques adaptées, le rendant ainsi indispensable à la mise en œuvre d'un Web Sémantique fonctionnel.

Indépendamment de cette vision et de la question du choix de représentation pour l'Annotation Sémantique, celle-ci dépend également du TAL en raison de l'espace et du volume de données visées. Il ne serait en effet envisageable ni de procéder manuellement à l'annotation des millions de documents potentiellement utiles, ni de ne considérer que les publications à venir en laissant de côté le Web déjà existant dans le processus de migration vers le Web Sémantique, ce qui constitue également un point de l'argumentation de Wilks en faveur d'un Web Sémantique cohérent [Wil08]. L'Annotation Sémantique doit de fait s'envisager comme une tâche automatisable afin d'atteindre ses objectifs. Le TAL se présente ici aussi comme un recours indispensable, puisqu'il permet, notamment à travers le large sous-domaine de l'Extraction d'Information, de dériver automatiquement les éléments visés par l'Annotation Sémantique à partir de contenus textuels.

Les méthodes développées par les recherches en TAL et en Extraction d'Information depuis

plusieurs décennies permettent donc de définir un cadre de réalisation pour l'Annotation Sémantique. Il lui fournit d'une part les moyens d'un traitement de données à grande échelle, incontournable en ce qui concerne le Web Sémantique. Mais l'Annotation Sémantique repose également sur le versant du TAL lié à des considérations centrales de l'IA, c'est-à-dire le franchissement de la frontière entre données linguistiques et représentation logique et formelle.

Le Web Sémantique consiste en une réalité faite de développements de standards et de pratiques de publication encouragés par des visées applicatives à la fois larges et donc encore peu spécifiées, mais motivées par une volonté de progrès dans le champ des connaissances et de leur représentation. L'enrichissement de contenus textuels constitue un élément de ce vaste projet, en tant qu'il est le résultat premier du processus d'Annotation Sémantique. En amont de spécification d'applications, de tâches ou de services permis à terme par l'Annotation Sémantique, l'enrichissement de contenus se présente en effet comme un objectif à part entière : il s'agit de donner une forme concrète et active à la mise à disposition de connaissances sous une forme interprétable, quelque soit la nature et la réalité des traitements envisageables sur cette base.

Le renouvellement des pratiques de publication documentaire proposé dans le paradigme du Web Sémantique se traduit donc principalement par l'enrichissement de contenus, où les documents considérés se voient augmentés d'une couche sémantique utile, celle des métadonnées fournies par l'Annotation Sémantique. Avant d'examiner plus en avant à quelles lignes de recherches et méthodologies se rattache cette annotation (chapitres 2 et 3), il semble utile de poursuivre la présentation de l'enrichissement de contenus en nous intéressant aux objets fondamentaux qu'il manipule — documents et métadonnées —, ainsi qu'aux raisons de l'intérêt porté au Web Sémantique par une organisation telle que l'AFP.

2 Documents et métadonnées : formalisation pour le traitement de l'information

Le Web Sémantique et ses objectifs d'exploitation des connaissances rappellent que le Web est avant tout une forme moderne de publication, de recueil et de consultation de l'information. Il se place à ce titre dans la lignée des traditions de regroupements documentaires destinées à conserver, inventorier et rendre accessible la connaissance humaine, notamment les bibliothèques. Celles-ci constituent la réalisation la plus manifeste de cette activité de gestion de l'information et c'est autour de leur organisation que se forment les paradigmes de formalisation d'accès aux connaissances, transposables à toute organisation concernée par les pratiques documentaires.

Le tournant numérique de l'information poursuit cet effort par de nouveaux moyens et contribue à renouveler un certain nombre de concepts fondamentaux dans le champ du traitement de l'information, aboutissant notamment aux développements dans le cadre du Web Sémantique et des ontologies évoquées précédemment. L'exploration de ces concepts ainsi que de leur évolution historique contribue à tracer les contours de l'enrichissement de contenus et des métadonnées, notamment en tant que moyen d'accès à l'information et aux connaissances dégagé de contraintes physiques et fondé sur la notion de réseau.

2.1 Modalités d'organisation et de description des contenus documentaires

2.1.1 De la catégorisation à l'indexation

La structure organisationnelle de l'information trouve un ancrage historique dans les méthodes successives employées par les dispositifs d'arrangement de ressources, qu'il s'agisse d'encyclopédies ou de bibliothèques notamment. Il s'agit avant tout d'organiser ces ressources elles-mêmes,

autrement dit des objets physiques : l'Encyclopédie, publiée en France entre 1751 et 1778, présente un état des connaissances sous forme d'articles classés selon l'alphabet, vu comme un ordre conforme à la raison ; le système de classification documentaire de Dewey[19], destiné aux bibliothèques, repose sur le système décimal et divise les connaissances en 10 grandes catégories, elles-mêmes sous-divisées par multiples de 10 sur deux niveaux supplémentaires (cf. figure 1.5). Dans ces deux exemples, la rationalité se pose comme pré-requis pour le bien-fondé du mode organisationnel adopté ; dans le second, c'est le principe de *catégorisation* qui préside à l'organisation.

0 (General).	**300 Sociology.**	**500 Natural Science.**
10 BIBLIOGRAPHY.	310 STATISTICS.	510 MATHEMATICS.
20 BOOK RARITIES.	320 POLITICAL SCIENCE.	520 ASTRONOMY.
30 GENERAL CYCLOPEDIAS.	330 POLITICAL ECONOMY.	530 PHYSICS.
40 POLYGRAPHY.	340 LAW.	540 CHEMISTRY.
50 GENERAL PERIODICALS.	350 ADMINISTRATION.	550 GEOLOGY.
60 GENERAL SOCIETIES.	360 ASSOCIATIONS AND INSTITUTIONS.	560 PALEONTOLOGY.
70	370 EDUCATION.	570 BIOLOGY.
80	380 COMMERCE AND COMMUNICATION.	580 BOTANY.
90	390 CUSTOMS AND COSTUMES.	590 ZOOLOGY.

20 Book Rarities.	**370 Education.**	**510 Mathematics.**
21 Manuscripts.	371 Teachers, methods, and discipline.	511 Arithmetic.
22 Block Books.	372 Elementary.	512 Algebra.
23 Early Printed.	373 Higher.	513 Geometry.
24 Celebrated Printers.	374 Self-education.	514 Trigonometry.
25 Celebrated Binders.	375 Classical and real.	515 Conic sections.
26 Materials.	376 Female.	516 Analytical geometry.
27 Ownership.	377 Religious and secular.	517 Calculus.
28 Prohibited.	378 Schools and Colleges.	518 Quaternions.
29 Other.	379 Reports.	519 Probabilities.

FIGURE 1.5 : Extrait de la table de classification décimale de Dewey.

Remontant à l'approche classique d'Aristote, la catégorisation se présente comme le procédé prototypique d'organisation, où la catégorie se définit par un ensemble de caractéristiques communes à tous ses membres. Augmentée d'une dimension hiérarchique, la catégorisation définit des classes et sous-classes, dont la précision du sens et la portée dépendent de leur degré de spécificité. Comme dans le système de Dewey, une catégorisation consiste en l'énonciation d'un ensemble de classes censées représenter un domaine donné — ici, tous les domaines considérés comme possibles par Dewey. Cette énonciation est indissociable d'un choix de représentation et fige la catégorisation selon cette restriction. Une catégorisation, ainsi que son composant hiérarchique, est de fait limitée dans son rôle organisateur puisqu'elle constitue le reflet d'une certaine vision, que celle-ci dépende d'une époque, d'une opinion ou d'un statut social particulier et ne peut prétendre à l'exhaustivité, de nouveaux domaines et objets de connaissances émergeant continuellement. Ainsi, Dewey place la philosophie comme catégorie première de son système, suivie par la religion, dans laquelle la Bible et le christianisme occupent sept des dix sous-catégories prévues. La catégorie 376, dédiée au thème de l'éducation des femmes, existe encore de

19. Reproduit en intégralité à l'adresse `http://www.gutenberg.org/files/12513/12513-h/12513-h.htm` dans le cadre du projet Gutenberg

nos jours. Lors de l'apparition de l'informatique, les sujets associés ont été placés par les mainteneurs du système dans la catégorie 000, dévolue aux « généralités », la catégorie « Technologies et sciences appliquées » ne présentant plus d'entrée décimale disponible. Dans le 23e édition du système de Dewey datant de 2003, le thème informatique accède au statut de catégorie plénière avec le renommage de la catégorie 000 en « Informatique, information et généralités ».

Un tel mode de catégorisation hiérarchisée est non seulement limité quant à la description de la réalité qu'il propose, mais également en tant que reflet de l'état du patrimoine considéré : les livres, en tant objets physiques, ne peuvent par nature être disposés à plus d'un endroit, ce qui empêche une catégorisation multiple d'ouvrages dont le sujet peut être caractérisé par plusieurs classes du système. Cette limitation correspond à ce que l'on peut appeler un « premier ordre » de l'organisation de l'information, ou celle-ci est contrainte par les objets eux-mêmes et fondée sur une vue unique des connaissances.

Bien que le système de Dewey continue d'être utilisé dans de nombreuses bibliothèques du monde entier, et que le principe de catégorisation demeure un mode fondamental d'organisation de l'information, un « second ordre » existe qui s'attache à la distinction entre les objets contenant les connaissances et les éléments informatifs eux-mêmes. Le système documentaire des bibliothèques présente là aussi une instanciation prototypique de ce second ordre, en tant que l'accès aux ressources qu'il permet repose en grande partie sur le principe de l'*indexation*. Par le relevé, pour chaque ouvrage, d'une liste d'indications descriptives et en les associant avec une méthode d'identification, le fonds d'ouvrage se trouve doublé d'un *catalogue* permettant de retrouver, à partir des descripteurs disponibles, les ouvrages pertinents. En termes plus généraux, les éléments d'un tel catalogue constituent des *métadonnées*, c'est-à-dire des unités d'information distinctes du contenu, chargées d'un sens particulier le rattachant à ce contenu.

Les métadonnées usuelles concernant les ouvrages conservés en bibliothèque sont les titres, noms d'auteurs, année d'édition et autres informations relatives à la publication, ainsi que les sujets auxquels un ouvrage donné se rattache, choisis par exemple parmi les classes d'une catégorisation similaire à celle de Dewey. Plus généralement, les documents constituant l'unité informative générique dans les systèmes d'information sont dotés de telles métadonnées et sont ainsi rattachées à la catégorisation dont relève le système correspondant. Cette dernière peut être plus ou moins générale ou spécialisée, organisée en catégories de diverses granularités, selon une hiérarchisation profonde ou non.

Ce processus d'indexation rend possible un accès aux connaissances par plusieurs points d'entrée, mais ne permet cependant pas de s'affranchir de l'arrangement physique des objets considérés. De plus, dans sa forme historique, il maintient une distinction entre le rôle d'organisateur et de pourvoyeur de la connaissance, notamment celui du bibliothécaire, et celui du public souhaitant y accéder. Dans ces deux ordres d'organisation, le principe d'une autorité chargée de la définition, du maintien et de la diffusion de l'information préside à un mode bi-directionnel d'accès aux connaissances, où la recherche d'information est limitée à l'espace et à l'organisation définis par cette autorité.

2.1.2 Organisation en réseau et contenus numériques

Plusieurs propositions alternatives pour l'organisation de l'information sous sa forme documentaire émaillent le vingtième siècle, comme le rappellent Wood ou Weinberger [Woo10 ; Wei07], autour de la remise en question de la catégorisation comme principe d'organisation.

Le « Répertoire bibliographique universel » (RBU) [Otl34] de Paul Otlet, entamé à la fin du XIXe siècle, se présente en 1934 sous la forme d'un catalogue constitué de 15 millions de fiches d'indexation, dans le but de répertorier tous les ouvrages publiés dans le monde. Le RBU repose sur la « classification décimale universelle », qui comprend des catégories mais également des notations algébriques permettant de faire référence à des intersections de sujets et donc de

représenter un réseau de concepts. Cette catégorisation multiple traversant les ressources permet à Otlet de répondre par courrier à des requêtes pour lesquelles il fournit les fiches d'indexation pertinentes, à la manière d'un réseau documentaire préfigurant le Web et les moteurs de recherche associés. Otlet imagine d'ailleurs dans l'avenir un « *télescope électrique, permettant de lire de chez soi des livres exposés dans la salle teleg des grandes bibliothèques, aux pages demandées d'avance. Ce sera le livre téléphoté.* » [Otl34].

Le documentaliste Jesse Shera, dans une réflexion sur l'introduction de la technologie informatique dans la gestion des bibliothèques, propose en 1965 [She65] une catégorisation non hiérarchique et indépendante des ouvrages physiques qui, selon lui, ne mettent pas au jour les relations existant entre leurs contenus ; ceux-ci sont constitués d'« unités de pensée », vers lesquelles l'organisation et la recherche de l'information doit être redirigée et ainsi dédiée à la conservation de l'intégrité intellectuelle présentée par ces contenus plutôt que des livres eux-mêmes.

En 1945, Vannevar Bush [Bus45] décrit le système théorique du « Memex », répondant ainsi au besoin d'une nouvelle organisation de l'information pour l'accès et la recherche, considéré comme nécessaire devant l'augmentation de la production documentaire, à laquelle Bush considère qu'il devient impossible de faire face avec les moyens alors à disposition. Ce système, invoquant les notions de « mémoire » et d'« index », aurait consisté en une bibliothèque miniaturisée munie d'un accès intelligent et à la disposition de tout un chacun. Les « pistes associatives » en constituent le fondement : à la manière d'annotations produites par l'utilisateur, ces pistes viendraient associer les contenus reliés et permettre d'y accéder ultérieurement. Il influence notamment Ted Nelson qui propose en 1965 [Nel65] le terme « hypertexte » dans le cadre de son modèle de création et d'utilisation de contenus interreliés, finalement concrétisé par les travaux de Tim Berners-Lee et l'avènement du Web.

Le tournant numérique touchant les ressources documentaires et le développement d'Internet comme support d'expansion du Web fait passer l'information de lieux de dépôts uniques, structurés selon un ordre centralisé et figé, à un espace de publication dont l'architecture est distribuée, tant au niveau de la production des contenus que de leur mise à disposition, construite en un réseau fondé sur le principe des liens hypertexte. Il s'agit là d'un changement de paradigme majeur, où les contraintes organisationnelles liées aux objets physiques laissent place à la possibilité de l'ubiquité. Le principe de catégorisation, loin de disparaître, s'inscrit dans cette structure en réseau en voyant sa forme fondamentale renouvelée : à partir d'une représentation sous forme d'arbre, la catégorisation spécifie par les nœuds et arcs les sujets et relations de subsomption existant entre sujets, les feuilles de l'arbre correspondant aux ressources du domaine représenté. Dans ce qui peut être qualifié de *troisième ordre* d'organisation, les feuilles peuvent être rattachées à autant de nœuds qu'il se trouve de sujet dans la catégorisation pour les caractériser. Les moyens informatiques permettent à ce processus de rattachement, c'est-à-dire de catégorisation multiple et indépendant d'une vue figée, d'avoir lieu lors de l'accès aux ressources, selon des critères définis par l'utilisateur, en fonction de besoins particuliers et éventuellement ponctuels, plutôt qu'au moment de la conception de l'ensemble documentaire considéré.

Ce nouvel ordre d'organisation de l'information est largement adopté par de nombreux secteurs d'activités présents sur le Web. L'entreprise Amazon[20], dont le site Web est majoritairement consacré au commerce en ligne de livres, présente une classification des ouvrages selon des schémas variables, dépendants à la fois de catégories assignées statiquement aux ouvrages mais multiples, de revues fournies par les utilisateurs ainsi que des historiques de recherche et d'achat liés aux ouvrages consultés. L'encyclopédie en ligne Wikipedia traite de tout sujet abordé par un de ses collaborateurs dans des articles dont l'organisation sous-jacente est très faiblement spécifiée : les catégories assignées aux articles sont assimilées à des étiquettes plutôt qu'aux nœuds d'une classification hiérarchisée et sont rassemblées dans une liste indexée peu utilisée dans les

20. `http://www.amazon.com/`

modalités d'accès aux contenus de Wikipedia. Ces modalités sont davantage conditionnées par l'enrichissement des articles en liens hypertexte renvoyant à d'autres articles de l'encyclopédie ou à des ressources externes, favorisant ainsi une exploration des ressources fondée sur des relations pertinentes et modulables, intégrant également une dimension de sérendipité pouvant bénéficier à ce processus d'exploration.

Bien que le principe de catégorisation continue de jouer un rôle déterminant dans l'organisation de l'information sur le Web, elle tend à se réaliser selon un axe allant des ressources vers la structuration, au gré de besoins variables au fil du temps et selon les individus, communautés et domaines. Le processus principal de caractérisation des données permettant d'obtenir les catégorisations en question est celui de l'*étiquetage* des ressources, à tout niveau — fragment textuel, page ou document, par lequel les catégories pertinentes peuvent être spécifiées explicitement ou inférées à partir d'ensembles relevant d'étiquettes similaires ou liées. Le paradigme du Web 2.0 exploite principalement ce processus en laissant les utilisateurs en charge de sa mise en œuvre. Les étiquettes ainsi créées, de même que les catégories qui en dérivent, constituent des *folksonomies* où la notion de taxonomie attachée aux catégorisations dérive de ce processus participatif et individualisé. L'effacement du principe d'une autorité à l'origine de l'organisation de l'information et contrôlant les modalités d'accès aux connaissances est particulièrement prégnant dans ce modèle.

Dans le troisième ordre d'organisation de l'information permis par la numérisation documentaires et le développement du Web, l'ensemble des objets pouvant constituer des descripteurs définis pour l'accès à une ressource s'élargit à tout élément lui appartenant, dépassant le cadre des métadonnées classiques : il peut s'agir d'un nom d'auteur ou d'un titre d'ouvrage, mais également de tout ou partie de son contenu textuel — une œuvre peut être retrouvée à partir de la citation d'un passage ou du nom de personnages qui y sont évoqués, grâce à l'indexation « plein-texte » sur laquelle repose le Web dans sa forme la plus usuelle. C'est en effet le paradigme de la *Recherche d'Information* (RI) qui, parallèlement à l'ubiquité caractérisant les ressources documentaires numériques en matière de classification, définit pour une large part les modalités d'usage et de développement du Web. La RI dans sa forme moderne et informatisée procède ainsi à l'indexation des ressources documentaires, en particulier textuelles, par la sélection de l'ensemble des mots ou termes de ces ressources — moyennant des processus de filtrage et de transformation — pour jouer le rôle de descripteurs exploitables par le modèle de RI en question. Il s'agit là aussi d'une émancipation au regard de la maîtrise du processus de description, réservé dans la RI manuelle à l'expertise du bibliothécaire : chargé de la production des descripteurs pertinents pour chaque ouvrage et relativement à l'ensemble du fonds, il est également l'intermédiaire obligé pour l'interrogation des ressources selon le schéma de description ainsi produit. Le principe de pertinence des descripteurs en RI contemporaine repose en revanche sur la notion de *mots-clés*, identifiés empiriquement parmi l'ensemble des termes d'indexation.

2.2 Vers des métadonnées sémantiques

L'enjeu pour le Web Sémantique en tant que cadre renouvelé de publication documentaire consiste notamment à dépasser le paradigme « plein-texte » de la RI telle qu'elle régit en grande partie les usages du Web actuel. Bien qu'indispensable au fonctionnement du Web et profondément ancrée dans les pratiques de la majorité des utilisateurs, la RI classique présente en effet un certain nombre de limitations freinant encore un accès aux ressources sophistiqué et directement exploitable par des services automatisés. L'indexation et la recherche par mots-clés ne caractérisent en effet les contenus qu'à un niveau lexical peu profond, en s'appuyant sur leur présence ou leur absence. Cette méthode, bien que prenant en compte des mesures de pertinence sophistiquées et ayant prouvé son efficacité pour l'accès à des ressources qui ne sauraient, à l'échelle du Web, être manipulées manuellement, exclut en revanche une approche sémantique de l'information. Le ni-

veau lexical considéré est en effet isolé des structures plus profondes à l'œuvre dans les contenues textuels : un mot-clé ne représente que lui-même, sa relation au reste du document comme à la requête est ignorée ainsi que le sont tous les phénomènes d'ambiguïté intervenant dans l'usage du langage naturel. La polysémie caractérisant un mot-clé pourra provoquer un résultat de recherche bruité voire non pertinent, tandis qu'une relation lexicale telle que la synonymie, non modélisée par les systèmes de RI classiques, empêchera la sélection de toutes les ressources pertinentes. Plus généralement, ce niveau lexical freine toute conceptualisation dans les processus d'indexation et de recherche et contraint à une manipulation de l'information dépourvue de sémantique.

L'étiquetage collaboratif évoqué précédemment, notamment développé sur les plateformes du Web 2.0, permet une distinction de nature entre contenus et mots-clés, les étiquettes étant choisies en regard du contenu et non extraites de celui-ci. La sémantique des catégorisations ainsi obtenues est néanmoins très faiblement définie : aucun schéma conceptuel précis et partagé n'étant prévu pour ces étiquettes, celles-ci relèvent, comme les contenus, du langage naturel, et les problèmes d'ambiguïté s'y appliquent donc de la même façon ; cette absence de schéma exclut également l'exploitation de l'étiquetage à des fins d'automatisation, du moins en ce qui concerne la définition d'une sémantique pour l'interprétation. Si cette faible conceptualisation n'empêche pas une utilisation satisfaisante des applications dérivées, elle ne s'inscrit pas dans les propositions de formalisation et de manipulation des contenus du Web Sémantique.

La refondation du modèle de représentation des connaissances proposée par le Web Sémantique peut être interprétée comme une RI *sémantique*, où les descripteurs sur lesquels reposent l'indexation et le processus de requêtes sont munis de sens et non uniquement considérés sous leur forme surfacique. Si les mots-clés en RI classique peuvent être vus comme des métadonnées, dans la mesure où ils constituent effectivement un mode de description des contenus, il s'agit dans le Web Sémantique d'user du principe des métadonnées afin d'ancrer les contenus dans une sémantique définie par les conceptualisations spécifiées pour un domaine donné. La RI reposant sur des métadonnées sémantiques franchit ainsi le seuil de la description surfacique pour intégrer ces conceptualisations : une requête visant à obtenir des informations sur les *États membres de l'Union européenne* ou des *acteurs français* se limite, en RI classique, aux documents mentionnant les mots de la requête eux-mêmes ; les documents mentionnant l'Allemagne ou la Grèce, Isabelle Huppert ou Lambert Wilson, mais pas les termes de la requête, seront alors ignorés dans les résultats de recherche. Avec une Annotation Sémantique des documents et une indexation sur les métadonnées ainsi produites, elles-mêmes associées à une ontologie dans lesquelles les individus référencés (Allemagne ou Isabelle Huppert) sont spécifiés comme membres de classes conceptuelles correspondant à des termes utilisables dans des requêtes, la RI peut alors accéder à un ensemble de documents plus pertinents. La RI sémantique donne lieu depuis quelques années à des travaux de recherche et de développement proposant différentes modalités de mise en œuvre ; celle-ci est notamment déterminée par le niveau d'intégration de la conceptualisation dans l'application, selon que les requêtes elles-mêmes ou seuls les résultats retournés sont de nature sémantique. Mangold [Man07] établit une classification de plusieurs de ces travaux autour de critères définis pour la caractérisation de la RI sémantique (*semantic search* en anglais).

Dans le paradigme de la RI sémantique et plus généralement de la caractérisation sémantiques des contenus, ceux-ci fournissent eux-mêmes les descripteurs nécessaires, comme c'est le cas avec les mots-clés, mais les métadonnées ainsi produites sont porteuses d'information dépassant la simple identification des mots ou termes sélectionnés : elles en spécifient la sémantique par association avec le schéma conceptuel correspondant. Les contenus ne sont donc pas étiquetés au niveau documentaire mais enrichis au niveau textuel lui-même par des annotations. Une dimension collaborative peut donc venir caractériser ce processus, puisque tout rédacteur puis lecteur et utilisateur peut augmenter cette couche d'annotation, usant de tout schéma considéré comme pertinent. Cette flexibilité associée à un ancrage dans des conceptualisations définies

et identifiées renvoie aux notions de partage et d'intégration aux fondements du Web Sémantique. Elle contribue également à effacer la distinction entre tenants de l'autorité concernant la structuration de l'information et utilisateurs confinés à la consultation.

La gestion documentaire envisagée par le Web Sémantique repose donc sur l'addition de métadonnées aux contenus, celles-ci étant dérivées d'annotations dont la production est intégrée au processus rédactionnel. La couche sémantique ainsi adjointe aux documents doit permettre à la fois une exploration plus riche et sophistiquée pour les utilisateurs humains et un mode d'identification et d'extraction à destination d'agents automatiques. En termes de RI, le Web Sémantique propose ainsi un renouvellement et une orientation vers une recherche à *facettes*, tenant compte de propriétés sémantiques des contenus vues comme autant de points d'accès aux connaissances.

2.3 Acquisition de métadonnées à partir des contenus

Si le renouvellement de la gestion et de la publication documentaire proposé par le Web Sémantique se distingue de la RI classique par la sélection de descripteurs munis de sémantique, il s'inscrit néanmoins dans une lignée similaire en envisageant l'acquisition de ces descripteurs à partir des contenus eux-mêmes, par opposition à des descripteurs inventoriés ou créés indépendamment d'eux, en partant du processus d'annotation des contenus.

Comme cela a été évoqué en 1.3, les métadonnées adaptées au Web Sémantique sont dérivées d'une annotation qui peut être manuelle ou automatique. Dans ces deux configurations, l'annotation est partie prenant du processus rédactionnel et produit des métadonnées associées à des fragments textuels — mentions de concepts, d'entités —, pointant vers le schéma conceptuel choisi pour l'adjonction de sémantique attendue. Ces métadonnées viennent caractériser les documents annotés à deux niveaux : elles sont étroitement liées à des mentions locales au sein du texte, mais peuvent en être physiquement détachées et ainsi constituer des descripteurs pour le document lui-même.

La dérivation directe des métadonnées à partir des contenus permet une description et une organisation documentaires empiriquement fondée, où la notion de pertinence s'émancipe de l'arbitraire éventuel de classifications figées, ainsi que de la faible définition sémantique des mots-clés dérivés de l'indexation plein-texte.

Tout en étant profondément ancrées dans les contenus, les métadonnées du Web Sémantique permettent l'ouverture du document vers un espace informatif plus vaste, puisque chacune d'elles explicite la sémantique liée à l'élément annoté par association avec un schéma conceptuel. Cette association identifie le contenu de l'annotation comme une instance de classe conceptuelle définie dans un schéma — typiquement une ontologie. Il s'agit donc d'un lien de référence qui conduit à l'obtention d'une sémantique que l'on peut qualifier de *référentielle*. Grâce à la mise en relation de ces instances dans un réseau tel que celui des Linked Data, principalement par des liens de synonymies, l'ouverture du document dépasse le seul schéma conceptuel sous-jacent à son annotation : les métadonnées de document permettent l'accès à toute autre conceptualisation définissant l'instance en question, ainsi qu'à tout autre document du Web munis des métadonnées similaires.

Les principes et pratiques de gestion et d'organisation des connaissances sous leur forme documentaire présentés ici sont indissociables des propositions et réalisations du Web Sémantique. Celui-ci se place en effet au niveau documentaire, suivant le modèle du Web, mais formule les conditions et moyens d'un dépassement de cet espace, principalement concrétisé par l'enrichissement des contenus à l'aide de métadonnées.

Si ces réflexions sont menées au sujet de l'espace de connaissances constitué par le Web, elles

concernent également les organisations, entreprises ou institutions qui fondent également une grande partie de leur activité sur des collections documentaires, que celles-ci soient des moyens ou l'objet même de cette activité. À ce titre, les objectifs et les schémas méthodologiques proposés par le Web Sémantique peuvent s'appliquer à ces organisations, ce qui rejoint la présentation initiale du Web Sémantique comme cadre de définition du besoin formulé par l'AFP et traité dans le présent travail. La section suivante expose les éléments principaux de ce rattachement des contenus documentaires privés au cadre du Web, du Web Sémantique et de l'enrichissement de contenus à l'aide de métadonnées.

3 Données d'entreprise et sémantique

Le terme *enterprise data* est fréquemment utilisé dans les travaux en langue anglaise sur les modalités de gestion des données, principalement documentaires, au sein d'entreprises mais également d'institutions et d'organisations en général en raison d'un fonctionnement commun. Le terme *données d'entreprise* ou DE est retenu ici. Le caractère privé constitue la distinction majeure entre les DE et les données du Web ; au-delà, c'est davantage une analogie entre elles que l'on peut décrire, conduisant à l'intégration des technologies du Web Sémantique dans l'organisation et la gestion documentaire autour des DE.

3.1 Analogie entre Web et données d'entreprise

L'analogie entre l'espace documentaire du Web et celui des DE peut d'abord être observée au niveau de leur forme : l'information produite et collectée par une organisation peut en effet constituer de très grandes quantités de documents, dont la masse pose des problématiques de gestion et d'organisation similaires à celles du Web. De façon comparable, des moyens de RI sont donc indispensables pour rendre possible un accès efficace aux DE et sont généralement déployés *via* les systèmes d'information ou intranets des organisations à destination de leurs personnels. La RI, combinée à d'autres méthodes de manipulation des contenus, peut répondre à des besoins d'exploitation de l'information pour la conduite de l'activité même d'une organisation, indirectement dans la conception de produits sur la base d'une présentation et d'une agrégation particulière de l'information, ou encore dans le processus de mise à disposition des DE hors du champ de l'organisation : diffusion publique, notamment sur le Web, ou à destination de clientèles, etc.

Les besoins suscités par la taille considérable souvent atteinte par les DE trouvent dans la formation et le mode de fonctionnement du Web une réponse pertinente quant aux moyens méthodologiques et technologiques à mettre en œuvre. Le Web présente en effet un ensemble de solutions efficaces au regard de l'échelle immense de l'information qui y est déposée, à la fois en ce qui concerne les pratiques de publication et les modes d'accès. Son architecture distribuée ainsi que la standardisation des moyens de communication et d'échange qui lui sont rattachés figurent en tête des éléments permettant d'expliquer et de décrire son succès et son efficacité. Cette architecture est donc communément adoptée par les organisations dont les données, notamment documentaires, présentent une échelle comparable.

Les développements portant sur la gestion et l'organisation des DE sont donc fréquemment intégrées aux recherches et propositions technologiques concernant le Web Sémantique. Une grande partie des innovations dans ce domaine trouvent d'ailleurs leur origine dans le secteur privé et industriel. Plusieurs exemples de cette intégration sont présentés dans des publications autour du Web Sémantique telles que *Semantic Technologies in Content Management Systems* [MK12].

La relation entre Web et DE s'illustre également par une tendance accrue des entreprises et institutions à développer leur présence sur le Web, procédant ainsi à une exportation de tout

ou partie de leurs contenus selon des méthodologies conformes à cet espace et à son mode de fonctionnement. C'est notamment par les fonctionnalités de recherche et plus généralement d'accès à l'information que cette présence peut adopter, à des degrés variables, les pratiques proposées dans le cadre du Web Sémantique. L'investissement du Web par les DE se reconnaît également dans le nombre grandissant d'organisations publiant leurs données sur le réseau des Linked Data, évoqué en 1.2.3 : Nations Unies (UN/LOCODE), Royaume-Uni (Crime Reports UK), Association for Computing Machinery (publication scientifique) ou index du *New York Times*.

3.2 Technologies du Web Sémantique dans l'entreprise : structuration et interconnexion des données

L'architecture générale du Web apporte à l'organisation des DE les notions connexes de distribution et de décentralisation. Il apparaît en effet que les larges quantités de données à produire et manipuler le sont plus efficacement hors d'une structure unique et centralisée [Woo10]. La distribution permet une répartition de la charge ainsi qu'une approche pragmatique et experte de l'information, traitée dans ce cadre par l'ensemble des acteurs concernés.

Indépendamment du Web Sémantique et de son émergence, les DE connaissent des techniques d'organisation adaptées au contexte privé et industriel. Il s'agit principalement de stocker, répertorier et rendre accessible les ensembles documentaires véhiculant l'information propre à une organisation. Cette information présente une structure correspondant à son activité, et les trois opérations en question doivent maintenir un cadre référentiel défini par cette structure. Les *vocabulaires contrôlés* sont largement répandus et employés dans cet objectif [Woo10] : les éléments informatifs y sont listés sous forme de lexique fermé, par ailleurs constitué et maintenu par une entité chargée de cette tâche au sein de l'organisation et faisant autorité en la matière. Des schémas peuvent s'ajouter à ces vocabulaires pour expliciter les types d'éléments informatifs manipulés ainsi que les relations existant entre eux. Ces schémas concernent en premier lieu les bases de données développées autour des DE et devant y faire référence afin de garantir leur interprétabilité dans le circuit informatif de l'organisation considérée.

Les organisations connaissent cependant des modalités de fonctionnement pouvant mettre à mal la cohérence et la validité des DE dans le processus de production et de gestion. L'information dans une entreprise est en effet souvent créée, collectée, mise sur la chaîne de production ou modifiée à plusieurs endroits et niveaux correspondant à la distribution des tâches, services et spécialisations caractéristiques de telles organisations. Cette situation conduit à la co-existence de multiples silos de données souvent spécialisées et structurées selon des modèles différents, rendant leur intégration difficile voire impossible et nécessitant fréquemment une nouvelle constitution des données, vue comme plus aisée que l'effort d'intégration alternatif [Woo10]. On peut observer qu'il s'agit là d'un problème similaire à celui de la production documentaire du Web, où l'absence de schéma commun et de moyens d'interprétation des données fondés sur une sémantique formalisée donne lieu à un vaste espace informatif sans relation et difficilement exploitable.

Les technologies proposées dans le cadre du Web Sémantique peuvent donc apparaître comme pertinentes pour la mise en œuvre de méthodes d'intégration des DE. La spécification de schémas conceptuels modélisant le domaine et les types d'éléments informatifs associés permettent de dépasser les limitations des vocabulaires contrôlés en termes de définition sémantique. Ces schémas sont à la fois applicables à la constitution des bases de données au cœur des DE, mais rendent également possible l'enrichissement des contenus documentaires, ainsi qu'une mise en relation effective des différents silos de données. La conception de schémas multiples peut être envisagée, notamment dans le but de maintenir et de traduire la variété correspondant aux différents services et spécialités d'une entreprise. Le principe des Linked Data peut alors être intégré au fonctionnement d'une organisation [Woo10] pour permettre la production et la distribution de

données de façon non centralisée, mais néanmoins associées à une sémantique définie ; les différents ensembles ainsi produits se trouvent mis en relation par le mécanisme référentiel propre aux Linked Data, conservant ainsi leurs particularités fonctionnelles tout en assurant une cohérence et une intégration globale du système de DE.

3.3 Contraintes fonctionnelles et pratiques pour des données d'entreprises liées

Les modalités de production et de gestion des DE sont fortement conditionnées par les CMS (abréviation du terme anglais *Content Management System*, soit « système de gestion des contenus »), outils principaux destinés au traitement de l'information dans les organisations. Les développements d'ingénierie chargés de mettre en œuvre les technologies dites sémantiques dérivées du paradigme du Web Sémantique doivent adapter les CMS afin d'y intégrer les fonctionnalités associées telles que la référence aux schémas définis pour une organisation donnée et l'annotation des contenus à l'aide d'éléments informatifs ancrés dans ces schémas.

Il s'agit autrement dit d'introduire le principe de l'enrichissement à l'aide de métadonnées dans le processus de production documentaire, parallèlement à la gestion générale de l'information qui peut également se présenter sous forme de données structurées, notamment dans des bases de données relationnelles. L'adaptation des CMS dans cette perspective s'accompagne nécessairement d'un changement dans les pratiques usuelles de gestion des DE : de façon comparable à la publication documentaire connaissant un renouvellement dans le cadre du Web Sémantique, notamment par l'ancrage sémantique des contenus dans les conceptualisations appropriées, la production de DE doit prendre en compte la nécessité de leur intégration et de leur caractère partageable pour une exploitation efficace de l'information à travers l'ensemble de l'organisation considérée.

À la différence du Web, les DE relèvent d'activités et d'acteurs privés et ne partagent donc pas systématiquement son caractère libre, ouvert et non contrôlé. La production de l'information sur un mode distribué s'accompagne usuellement du maintien de schémas centralisés sous la forme de vocabulaires contrôlés, même lorsque ceux-ci migrent vers des spécifications conceptuelles plus sophistiquées que des listes d'entrées lexicales. Le processus de production intègre souvent une contrainte de correction maximale des données, notamment lors des étapes d'enrichissement à l'aide de métadonnées. L'introduction d'outils d'automatisation de la manipulation des contenus s'accompagne alors de méthodes de contrôle et de validation des résultats plus prégnantes que lors de leur déploiement dans l'espace public du Web.

Les DE constituent un champ d'application des technologies du Web Sémantique en tant qu'elles présentent des similarités de forme avec l'espace documentaire du Web, ainsi que des besoins et contraintes de fonctionnement pour lesquelles l'introduction de traitements dits sémantiques peut constituer une réponse adéquate. Cette analogie se concrétise surtout dans le renouvellement de la RI classique vers sa version sémantique, permettant une utilisation des DE plus sophistiquée, des mises en relation de l'information plus aisées et moins coûteuses ainsi qu'une visibilité accrue et facilitée par l'adoption du schéma de publication des Linked Data.

Chapitre 2

L'Extraction d'Information : jalon méthodologique pour l'enrichissement de contenus textuels

L'enrichissement de contenus textuels, placé dans le contexte du Web Sémantique au chapitre 1 de ce mémoire, correspond à un objectif de traitement de l'information reposant sur les notions de partage et d'accessibilité. La présentation générale du Web Sémantique a souligné le rôle structurant de l'Annotation Sémantique à cet égard : il s'agit du processus par lequel l'enrichissement de contenus à l'aide de métadonnées peut être réalisé, et nécessite à ce titre une définition méthodologique et fonctionnelle. Dans cette perspective, l'Extraction d'Information se présente comme un jalon fondamental : développée dans le champ du traitement de l'information et du traitement automatique du langage depuis plusieurs décennies, son objectif majeur de structuration de l'information en vue de la facilitation d'une exploitation automatique peut être considéré comme une formulation historique du paradigme du Web Sémantique. Elle fournit ainsi un ensemble de méthodologies et techniques permettant d'envisager la mise en œuvre de l'Annotation Sémantique, notamment en termes d'automatisation.

Un examen de la parenté existant entre Web Sémantique et Extraction d'Information ainsi que du lien méthodologique que l'on peut en dériver permettra de justifier la nécessité de l'intégration de l'Extraction d'Information dans la mise en œuvre de l'Annotation Sémantique (section 1). À la suite d'une synthèse historique du développement de l'Extraction d'Information et de sa systématisation (sections 2.1 et 2.2), la structuration fondée sur la classification qui y occupe une place centrale (section 2.3) pourra être mise en regard du processus d'instanciation ontologique évoqué au chapitre 1. Les entités, qui se placent au centre de l'objectif d'enrichissement de contenus, font l'objet d'un traitement extensif en Extraction d'Information sous leur forme linguistique par le biais de la Reconnaissance d'Entités Nommées, abordée à la section 3.1. Dans la perspective d'une intégration de l'Extraction d'Information à l'Annotation Sémantique, la sémantique attribuée aux entités nommées par les méthodes de reconnaissance pourra être discutée (section 3.2) et mise en regard d'autres approches des entités tenant compte de leur aspect référentiel (section 3.3).

1 Web Sémantique et Extraction d'Information : parenté et relation méthodologique

1.1 Définitions et périmètre analogique

La proposition de définition du Web Sémantique développée dans le chapitre 1 place au premier plan le problème de la représentation des connaissances à partir du langage naturel dans un objectif de formalisation sémantique pour la facilitation de traitements automatisés. Il a en effet été souligné dans le chapitre précédent que le déploiement du Web Sémantique requiert l'existence d'un niveau informatif formalisé se superposant aux contenus du Web — principalement des documents textuels —, et les modalités d'un tel déploiement ont été discutées. La question du passage entre niveau textuel et représentation sémantique demeure cependant prégnante puisqu'il s'agit là de l'obstacle principal à une réalisation immédiate du Web Sémantique et des modèles de publication de données qui lui sont assimilés.

Cette question occupe depuis plusieurs décennies l'Extraction d'Information (EI), qui se définit et se développe en tant que sous-domaine du traitement automatique du langage (TAL). Celui-ci, notamment dans sa composante de linguistique formelle, pose dès ses débuts le problème de l'obtention de structures non-linguistiques à partir d'énoncés en langage naturel et permettant une représentation du sens. L'EI aborde cette problématique à un niveau applicatif en proposant des moyens de structuration du langage naturel fournissant une représentation de l'information adaptée à des traitements automatiques ultérieurs, tels que la fouille de données, le résumé ou les systèmes de question-réponse, typiquement sous une forme similaire aux bases de données.

Avant de s'intéresser plus en avant à l'EI, on peut d'ores et déjà observer dans cette définition liminaire une communauté d'objectifs avec le Web Sémantique concernant le traitement du langage naturel, qui vise dans les deux paradigmes à rendre manipulable automatiquement l'information qui s'y trouve exprimée. Au-delà de la possible analogie caractérisant le Web Sémantique et l'EI, cette dernière se présente comme le recours méthodologique venant répondre à la question du passage entre contenus textuels et représentation sémantique telle qu'elle se pose pour la mise en œuvre du Web Sémantique, c'est-à-dire pour l'Annotation Sémantique.

On peut rappeler ici l'organisation du diagramme illustrant le déploiement du Web Sémantique (figure 1.1, p. 28) : la spécification à son niveau inférieur d'une couche données textuelles (brique « Unicode »), associées aux URI et à un balisage au format XML, correspond à l'Annotation Sémantique sur laquelle reposent les niveaux supérieurs. À partir de ces annotations, il est envisagé que soient constitués des inventaires de connaissances, typiquement sous forme de triplets RDF, mettant en relation des entités de diverses natures par le biais de prédicats eux aussi inventoriés. Wilks et Brewster souligne dans [WB09] que ces inventaires sont similaires au résultat attendu d'un système d'EI, chargé de repérer les entités pertinentes au niveau informatif dans des contenus textuels, de les typer et d'indiquer les relations existant entre elles. L'exemple suivant, repris à la présentation de Grishman [Gril2], illustre la démarche de l'EI ; il concerne l'extraction d'informations au sujet de personnels d'organisations, prenant et quittant leur fonction. À partir du passage suivant :

> *Frédéric Pierrafeu a été nommé directeur des systèmes d'information de Time Bank Inc. en 2031. Il s'est marié l'année suivante et est devenu PDG de Dinosaur Savings & Loan.*

le processus d'EI cherche à produire une structure informative, illustrée par la table 2.1, qui pourra être utilisée dans des traitements ultérieurs, tels que la constitution d'un historique de ressources humaines à partir d'un grand nombre de documents comportant ce type d'informations. Un graphe RDF dérivant du même passage, obtenu par une analyse prenant place en 2032 et au

Personne	Entreprise	Fonction	année	arrivée/départ
Frédéric Pierrafeu	Time Bank Inc.	DSI	2031	arrivée
Frédéric Pierrafeu	Time Bank Inc.	DSI	2032	départ
Frédéric Pierrafeu	Dinosaur Savings & Loan	PDG	2032	arrivée

Figure 2.1 : Exemple de résultat d'Extraction d'Information.

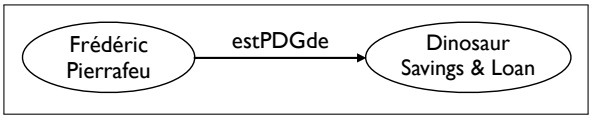


Figure 2.2 : Graphe RDF (sujet, prédicat, objet).

périmètre informatif réduit, peut être représenté par la figure 2.2. La relation entre l'EI et un tel graphe, dont la forme et la structure sont propres aux attentes formulées dans le cadre du Web Sémantique, est d'ordre méthodologique : c'est par une opération relevant de l'EI qu'il est possible de repérer dans les contenus considérés les mentions d'entités — ici, le nom de personne *Frédéric Pierrafeu*, les noms d'organisations *Time Bank Inc.* et *Dinosaur Savings & Loan* — ainsi que la réalisation surfacique et linguistique des prédicats par lesquels elles sont mises en relation.

Le recours à l'EI implique que l'on considère les éléments informatifs recherchés comme présents dans les contenus — à la différence de méthodes s'attachant à l'inférence de connaissances implicites — et que leur organisation surfacique permet, par le biais de l'analyse adéquate, de les reconnaître au sein de ces contenus. Si l'on accepte l'idée que cette organisation surfacique résulte d'une superposition complexe de niveaux linguistiques et que le sens qu'elle véhicule est fonction de ses différents composants, l'analyse en question consiste alors en une projection inverse, par laquelle il est possible de retrouver les éléments informatifs à partir de la forme linguistique de surface.

Si cette opération ne constitue qu'une étape des traitements nécessaires à une Annotation Sémantique complète, où les entités et prédicats relevés sont identifiés en fonction de ressources sémantiques adaptées — ce qui fera l'objet d'un examen idoine dans le chapitre suivant de notre travail —, elle apparaît cependant nécessaire à sa mise en œuvre. La dimension indispensable de l'EI pour l'Annotation Sémantique requise par le Web Sémantique et plus concrètement l'enrichissement de contenus textuels se justifie d'autant plus que de tels traitements ne peuvent s'envisager sans automatisation, si l'on suit l'argumentaire proposé par Wilks et Brewster [WB09], discuté au chapitre 1 (section 1.3.2) et selon lequel il est difficilement imaginable de réaliser une Annotation Sémantique complète de façon exclusivement manuelle.

1.2 Modalités de structuration en Extraction d'Information

L'obtention d'une représentation de l'information à partir de contenus textuels se réalise en EI sur le principe de la *structuration* : les données sous leur forme linguistique se présentent en effet au niveau informatique de façon *opaque*, selon la formulation de Moens et De Busser dans [Moe06], et c'est par l'assignation d'une structure aux informations véhiculées que l'EI propose de les rendre transparentes, c'est-à-dire manipulables et exploitables automatiquement. La structuration réalisée en EI procède d'une spécification des éléments informatifs à repérer et distinguer du reste du texte, vu comme non pertinent. L'EI produit ainsi des résultats sous forme de segments textuels, à la différence de la recherche d'information par exemple, qui retourne des documents. Les éléments informatifs sont non seulement spécifiés en amont de la réalisation de la tâche d'EI mais également caractérisés en termes de nature et d'étendue : il peut s'agir de mots, de termes,

de groupes tels que des groupes nominaux, de verbes, etc. Cet ensemble de spécifications forme la structure vers laquelle les éléments informatifs extraits doivent être mis en correspondance, et correspond ainsi à un *modèle* du domaine et de la tâche concernés.

L'EI est en effet employée, dans sa forme traditionnelle, dans le cadre de domaines et de tâches définis au préalable et selon lesquels sont opérés une sélection, une caractérisation ainsi qu'une classification des éléments informatifs recherchés. Le modèle ainsi obtenu circonscrit un ensemble de classes informatives dans lesquelles il s'agit de placer les extractions textuelles afin de leur conférer une sémantique induite par l'appartenance à la classe, en vue des traitements pour lesquels l'EI prépare ainsi les données. La spécification du modèle que l'on souhaite instancier par application de l'EI consiste donc en une conceptualisation locale, puisque restreinte à un domaine particulier, et définie *a priori*. Dans l'exemple de Grishman (table 2.1), les mouvements de personnels entre différentes fonctions dans différentes organisations sont ainsi représentés dans une structure où les personnes et organisations, identifiées par un nom, ainsi que la nature des mouvements, leur date et les fonctions occupées constituent des classes informatives assignées aux fragments textuels extraits. Ceux-ci sont ainsi typés sémantiquement et mis en relation selon la spécification de la conceptualisation donnée pour ce domaine particulier.

La définition d'un modèle de domaine selon une structuration en classes ou types sémantiques rejoint la composante de représentation des connaissances du Web Sémantique, évoquée au chapitre 1 (section 1.2) et principalement réalisée par le biais d'ontologies. L'idée partagée ici entre Web Sémantique et EI est celle de tâches à accomplir dans le cadre d'un domaine particulier et de la nécessité d'une conceptualisation donnée *a priori* afin de munir les informations pertinentes d'une sémantique pour leur exploitation. La complexité et la profondeur des modèles utilisés en EI sont variables et largement déterminées par la tâche considérée, selon le degré de sophistication et de finesse de grain requis pour une description adéquate de ses éléments pertinents. Les représentations en EI consistent généralement en un ensemble de paires associant un attribut à une valeur : chaque paire caractérise un aspect du domaine et de la description qui en est faite — par exemple, le fait que des personnes, des entreprises et des fonctions soient les composants principaux du domaine des ressources humaines — à l'aide d'un attribut typant l'élément extrait, sa valeur étant le fragment textuel lui-même. Un tel ensemble peut être spécifié à l'aide d'une simple liste, mais peut également faire l'objet d'une description plus formelle : les attributs peuvent en effet se voir ancrés dans une taxonomie, une classification conceptuelle hiérarchisée ou, dans la perspective d'une interprétabilité accrue, une ontologie.

À partir de ses composants principaux, du mode de structuration de l'information qui la caractérise et de sa parenté avec certains des requis fondamentaux du Web Sémantique, la question se pose de savoir comment l'EI peut constituer une voie méthodologique pour l'accomplissement du processus d'Annotation Sémantique.

1.3 Intégration de l'Extraction d'Information dans l'Annotation Sémantique

L'enrichissement de contenus textuels repose sur le processus d'Annotation Sémantique, définie dans le cadre du Web Sémantique, puisque c'est par ce biais que peut émerger, à partir des contenus textuels, un ensemble de métadonnées porteuses d'informations sémantiques, destinées à une utilisation ultérieure par des services automatisés. L'Annotation Sémantique consiste en effet en l'adjonction d'un niveau informatif superposé au texte, localisant les fragments textuels porteurs de l'information recherchée et indiquant l'ancrage sémantique qui peut être associé à chacun d'eux. Ces deux opérations de localisation et d'ancrage sémantique peuvent être mises en regard du repérage et du typage sémantique accomplis en EI et évoqués précédemment.

L'EI intervient en effet dans l'Annotation Sémantique comme composant méthodologique essentiel en tant que moyen de localisation des éléments informatifs pertinents dans un cadre donné : ce n'est pas l'ensemble du texte d'un document qui est visé, mais un certain nombre de fragments textuels pouvant être ramenés aux éléments d'un modèle spécifié au préalable — même si ces unités informatives peuvent constituer une partie plus ou moins importante du texte, jusqu'à éventuellement recouvrir l'ensemble d'un document. La capacité de repérage de l'EI, traditionnellement déployée sur des contenus constitués indépendamment de la tâche d'extraction, c'est-à-dire en mode analytique et *a posteriori*, est employée en Annotation Sémantique au niveau de la production des données : l'Annotation Sémantique est en effet partie prenante du processus rédactionnel (cf. 1.3.1) et le caractère pertinent des unités informatives à retenir pour l'annotation est évalué dans ce cadre. On peut à cet égard retenir la formulation proposée par Wilks et Brewster [WB09] selon lequel l.e Web Sémantique et l'Annotation Sémantique se présentent comme producteurs, quand l'EI se situe elle du côté de la consommation. L'intégration de l'EI en Annotation Sémantique se fait donc dans un sens productif, d'une part, et affaiblit la distinction entre contenus et résultats d'extraction d'autre part, puisque ceux-ci ne sont plus attendus en tant que structure informative indépendante du texte. Les annotations peuvent se présenter physiquement séparément du texte — dans un en-tête de document ou dans une base de données attenante par exemple, plutôt qu'à l'endroit même du texte annoté —, mais lui demeurent associées. Il ne s'agit pas, en effet, d'extraire définitivement un ensemble de connaissances à partir de contenus, mais de munir ces derniers des indications nécessaires à des traitements automatiques fondés sur l'interprétabilité de l'information, ce qui consiste, en premier lieu, à explorer et associer des ensembles documentaires par le truchement des annotations. Le lien entre annotation et texte reste donc essentiel et son caractère plus ou moins lâche — selon que les annotations accompagnent le texte ou se présentent séparément — dépend de chaque contexte d'utilisation.

Le typage sémantique des unités informatives repérées par l'EI fait également apparaître une parenté avec l'Annotation Sémantique pouvant donner lieu à une intégration méthodologique : il s'agit d'assigner une classe définie par le modèle sous-jacent à la tâche d'EI à chaque unité extraite. En Annotation Sémantique, le modèle consiste en une ontologie dont les classes conceptuelles sont les cibles des liens à établir entre les segments textuels localisés et cette ontologie. Si l'Annotation Sémantique se présente donc selon un schéma fonctionnel parallèle à celui de l'EI, autour des deux opérations principales de repérage et de typage, l'ancrage sémantique qu'elle cherche à réaliser doit cependant relever d'une formalisation avancée, notamment formulée par le biais d'ontologies fondées sur les logiques de description. Il semble nécessaire, à ce titre, d'évaluer l'adéquation de l'EI quant à cette condition de formalisation. Un examen plus précis de l'EI et du traitement de l'information qu'elle propose permettra ainsi de déterminer le périmètre méthodologique dans lequel cette discipline constitue un jalon pour la réalisation de l'Annotation Sémantique, ainsi que les limitations qui peuvent la caractériser au niveau de l'ancrage sémantique recherché.

2 La tâche d'Extraction d'Information

La synthèse historique et méthodologique de l'EI présentée ici propose un examen de cette discipline dans sa dimension de traitement de l'information textuelle, partant d'ambitions de compréhension du langage naturel propres au TAL dans lequel elle s'inscrit, et aboutissant à la définition stable et systématique d'une tâche applicative, efficace et éprouvée. Les éléments de définition de l'EI donnés précédemment trouveront ainsi une illustration concrète dans les stades de sa mise en œuvre au cours des dernières décennies. Au sortir de cette description générale, l'attention sera portée plus précisément sur la composante de l'EI qualifiée de sémantique, bâtie

autour du principe de classification, afin de déterminer son adéquation à la problématique de l'Annotation Sémantique.

2.1 Origines : un objectif du TAL et de l'intelligence artificielle

L'EI constitue un sous-domaine du TAL et trouve à ce titre ses origines dans les théories et développements orientés vers l'objectif de compréhension automatique du langage naturel formulé dans la seconde moitié du XX^{e} siècle. L'idée d'une mise en correspondance entre données textuelles et structures sémantiques identifiées pour un domaine remonte à Harris [Har58] et peut être considérée comme une formulation originelle de l'EI telle qu'elle se présente aujourd'hui. Les travaux les plus notoires en la matière dans les années suivantes se concentrent sur la mise en œuvre d'une analyse textuelle ambitieuse reposant sur la représentation des connaissances, d'abord envisagée de façon générale.

Cet objectif du TAL renvoie à son intégration dans le champ plus large de l'intelligence artificielle (IA) et place au premier plan la problématique de dérivation sémantique à partir de données linguistiques. La compréhension visée est ainsi ramenée à la possibilité de représentation du sens véhiculé par les énoncés en langage naturel par le biais d'une formalisation permettant l'automatisation de son interprétation. C'est dans cette perspective que sont proposées, à partir des années 1960, des théories et modalités de représentation partant d'un effort de formalisation des structures de connaissances vues comme sous-jacentes au langage.

La théorie de la Dépendance Conceptuelle À la fin des années 1960, Roger Schank propose une contribution majeure à l'IA sous la forme de la Théorie de la Dépendance Conceptuelle (*Conceptual Dependency Theory*, CDT) [Sch72]. Le modèle dérivant de la CDT suppose l'existence d'une base conceptuelle indépendante de la langue, rendant compte de scénarios ou de schémas d'exécution d'actions humaines. Organisée autour de primitives (objets, actions, attributs, lieu, temps), cette base conceptuelle présente des interconnexions régies par un ensemble de règles établissant les dépendances possibles entre concepts. Les structures linguistiques sont mises en correspondance avec cette base lors de la compréhension, et créées à partir d'elle lors de la génération, formant ainsi des conceptualisations manipulables selon les règles de dépendance établies. La définition des unités linguistiques (noms, verbes, etc.) est ainsi formulée en termes de primitives ou de prédicats conceptuels. La CDT vise à l'extraction d'informations sémantiques à propos d'événements atomiques à partir de phrases *via* leur conversion au niveau conceptuel, ce qui peut être illustré par le diagramme reproduit à la figure 2.3. Plusieurs systèmes adoptant ce modèle ont

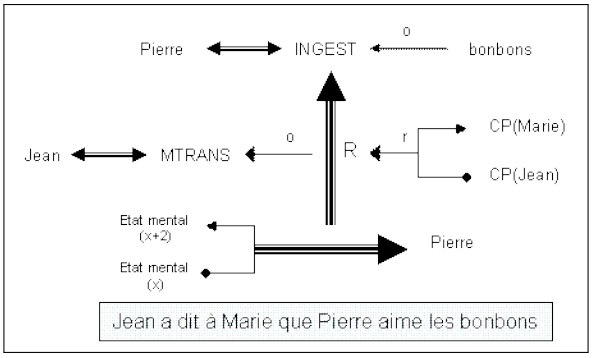


FIGURE 2.3 : Exemple de diagramme conceptuel de Schank (extrait de [Sab90]).

été développés, notamment à l'Université de Yale, jusqu'aux années 1980 : SAM [SA77] produit un réseau de dépendances conceptuelles entièrement instancié à partir d'un texte, usant entre autres

de scripts complets pour l'explicitation des conceptualisations, c'est-à-dire contenant l'intégralité des informations associées à un scénario. La majorité des systèmes de CDT utilisent cependant des scripts incomplets, munis des conceptualisations les plus importantes ou pertinentes. L'analyse partielle ainsi menée permet l'extraction de certains éléments informatifs seulement, ce en quoi elle prédéfinit l'EI, comme le font observer Moens et De Busser [Moe06]. Le système FRUMP, également développé à Yale [DeJ77 ; DeJ82] constitue une réalisation typique de ces scripts partiels. Dans la lignée de la CDT et de la compréhension de scénarios, Lehnert propose la construction de graphes connexes d'unités (*plot units*) représentant la structure narrative d'un texte [Leh82].

Parallèlement aux recherches en CDT, la compréhension de textes donne lieu au système de Rumelhart ; Rumelhart [Rum77 ; Rum75] fondé sur des grammaires narratives, rendant compte du texte sous forme de structures hiérarchiques. Le « Linguistic String Project » (LSP) commencé en 1965 à l'Université de New York se tourne vers le développement de méthodes de structuration et d'accès à l'information dans la littérature scientifique et technique. L'analyse de document y est fondée sur des principes linguistiques dans une optique de démonstration de l'analyse grammaticale automatique, menant aux méthodes d'analyse de sous-langages [Sag81] et ainsi à la spécialisation en domaines.

Mais les travaux de Schank et la CDT, quoique peu implémentés dans leur totalité théorique, demeurent une influence majeure et durable dans les recherches portant sur les modalités de représentation du langage naturel pour un traitement automatique.

La théorie des cadres Les méthodes de représentation des connaissances fondées sur les cadres (*frames*) constituent un courant important dans la lignée de la CDT. Formulée explicitement par Minsky en 1975, la notion de cadre s'inscrit dans le champ des structures de représentation des connaissances pour l'IA. Elle se traduit par des structures de données — les cadres — correspondant à des situations stéréotypées, pour lesquelles chaque cadre enregistre les propriétés des entités, actions ou événements pertinents. Un cadre possède ainsi un certain nombre de champs destinés à être remplis par une valeur, qui peut consister en une référence à un autre cadre. Les cadres disposent de valeurs par défaut pour le remplissage des champs, du principe de l'héritage des valeurs de champs entre cadres, ainsi que de la possibilité d'obtenir une valeur par application dynamique d'une procédure. Un ensemble de cadres avec relations mutuelles définit un réseau sémantique de cadres.

De nombreux systèmes d'EI, développées au cours des années suivantes dans le cadre d'une identification plus précise de cette tâche, reposent sur la structure des cadres. Ceux-ci se présentent en effet comme un mode de représentation largement utilisé et ne se limitant pas à l'EI, comme en témoigne le projet de ressource lexicale FrameNet [BFL98] qui emploie la structure conceptuelle des cadres, ainsi que la conception du langage ontologique OWL et les logiques de description de façon générale, pour lesquelles les cadres constituent une base paradigmatique [Gra+08].

Ce premier mouvement vers l'EI, que l'on peut circonscrire à ces différentes théories et mises en œuvre, donne lieu, de la fin des années 1970 aux années 1980, à la conception de systèmes génériques. Ceux-ci se caractérisent par une attention portée de façon privilégiée sur un fondement théorique et linguistique, devant garantir par des représentations adéquates une forme de compréhension automatique du sens. L'EI en tant que discipline y trouve de premières réalisations, mais l'idée de collecter des informations selon des structures définies par domaine reste un moyen de démonstration des capacités des systèmes mis au point plutôt qu'une tâche circonscrite et envisagée en tant que telle. Comme cela est rappelé par Poibeau et Nazarenko [PN99], de tels

systèmes, reposant sur des structures de nature logico-conceptuelles destinées à formaliser l'information contenue dans l'ensemble d'un texte et souvent conçus indépendamment du domaine, donnent lieu à des représentations d'une complexité peu opérationnelle. Le fondement linguistique, logique ou cognitif dominant les recherches de cette période tendent donc à réaffirmer le problème de la relation entre langage naturel et sens, sans permettre de dégager de méthodes manifestement adéquates pour sa compréhension et son implémentation informatique.

2.2 Systématisation de l'Extraction d'Information

Au cours des années 1980, l'intérêt porté aux possibilités de traitements automatisés de l'information est grandissant et est notamment manifesté par des communautés et institutions pour lesquelles elle constitue un enjeu crucial. Dans ce contexte, l'EI émerge en tant que tâche identifiée et utile, autour de processus d'évaluation par comparaison avec les performances humaines sur les problèmes visés. L'empirisme devient concomitamment l'approche dominante, tant au niveau de la définition de l'EI que de sa mise en œuvre fonctionnelle.

Comme le souligne Grishman [Gril2], une caractéristique notable de l'EI réside dans l'importance de l'évaluation, encouragée par des institutions gouvernementales américaines, dans les recherches menées au cours des années 1980. En particulier, les campagnes MUC (*Message Understanding Conference*) initiées par la Marine américaine en 1987 et financées par La DARPA (*Defense Advanced Research Projects Agency*) jusqu'en 1998 réunissent des équipes de recherche autour du problème de la compréhension de messages et contribuent principalement à une définition systématique de la tâche d'EI. L'histoire et la méthodologie des campagnes MUC sont désormais bien connues et étudiées, notamment par Grishman et Sundheim [GS96], Hirschman [Hir98] ou Cowie et Wilks [CW00]. En suivant Ehrmann [Ehr08], on peut dégager trois cycles dans la description de MUC, autour de la définition et de la complexité de la tâche d'EI, des données fournies aux participants et des modalités d'évaluation mises en œuvre.

Les deux premières éditions de MUC sont principalement exploratoires et ne concernent que de courts messages de la Marine américaine. Le principe des formulaires de structuration des données (nommés *templates*), à remplir par les participants à partir du contenu des messages, ainsi que les premières mesures d'évaluation — précision et rappel, empruntés à la recherche d'information — sont adoptés lors de MUC-2 (1989). À partir de MUC-3 (1991), les données intègrent des corpus journalistiques sur le thème du terrorisme en Amérique Latine et les formulaires présentent des champs à remplir en plus grand nombre. MUC-4 (1992) ajoute la F-mesure aux métriques d'évaluation, combinant précision et rappel pour une meilleure comparaison entre les systèmes. Des corpus de domaines différents — microélectronique, vente d'entreprises — sont incorporés à MUC-5 (1993), qui présente des tâches à réaliser pour le japonais en plus de l'anglais. La diversification et la complexité accrue de MUC-5 révèlent un besoin de généricité et donnent lieu à un effort de développement vers la portabilité des systèmes, sans que leur performance ne progresse notablement. Leur adaptation reste longue et laborieuse, mais un certain nombre de sous-tâches sont identifiées comme distinctes et indépendantes de la tâche globale et orientent les développements vers la conception de modules liés à des fonctionnalités précises. MUC-6 et MUC-7 (1995 et 1998) présentent ainsi des systèmes modulaires et encouragent leur conception autour de la portabilité. Les tâches de résolution de coréférence, de désambiguïsation lexicale et de détection des structures prédicatives sont envisagées, la première seulement étant réalisée. Les formulaires sont simplifiés et normalisés ; ils modélisent les entités (*Template Element*) ainsi que les relations entre entités (*Scenario Template*). Le principe de sous-tâches distinctes et de modules indépendants s'illustre dans la création de la tâche de détection d'entités nommées, sur laquelle nous reviendrons plus loin dans ce chapitre (section 3.1). Ces dernières éditions voient également

la généralisation des méthodes probabilistes et d'apprentissage automatique concurrencer les méthodes symboliques. Les performances encourageantes obtenues par les participants légitiment l'approche favorisant une décomposition des tâches.

À l'issue de cette décennie de campagnes, la tâche d'EI apparaît comme bien définie, rassemble une communauté de recherche et de développement active et fournit des résultats d'une qualité proche des performances humaines. Cowie et Wilks [CW00] soulignent cependant un travers lié à une évaluation placée au centre des préoccupations, qui consiste en une tendance à concevoir des solutions de court terme, où l'innovation ne joue pas un rôle central, afin de répondre aux attentes précises et restreintes des applications dans le cadre de MUC. L'utilité des technologies émergeant de MUC est cependant reconnue, mais il est également mis en avant que l'impératif d'adaptabilité et de vitesse de développement tend à éviter le déploiement de l'ensemble des méthodes de TAL disponibles, ce qui pose la question d'une relation lâche entre EI et TAL. On peut à cet égard observer que la tâche de résolution de coréférence introduite lors de la dernière édition de MUC ouvre des perspectives de recherches plus fondamentales.

Parallèlement et en relation avec les campagnes MUC, la majorité des systèmes d'EI proposés au cours des années 1990 prennent la suite des ambitions de compréhension du langage naturel, formulées lors des décennies précédentes et évoquées précédemment, autour du principe de localité et selon un mode applicatif. L'EI se présente alors, selon la formule de Poibeau et Nazarenko [PN99], comme un outil de *compréhension locale et guidée par le but*. Il s'agit en effet d'instancier des schémas informationnels définis en fonction de tâches précises ainsi que de types de texte et de domaines circonscrits. Les méthodes symboliques s'attachant à la reconnaissance des éléments de ces schémas sont à ce titre longtemps dominantes ; elles opèrent à partir du repérage d'amorces, du déploiement d'automates et d'heuristiques locales. Chaque nouvelle tâche d'EI donne donc lieu à la mise au point d'un nouveau système, dont les fonctionnalités sont étroitement liées à l'identification de ses éléments informatifs caractéristiques ; ceux-ci pouvant être différents pour chaque domaine et type de texte, l'adaptabilité des systèmes est donc peu opérationnelle.

Les dernières éditions des campagnes MUC orientent la conception des systèmes vers un effort de généralisation et de modularité qui constitue un tournant dans l'approche de l'EI. Cette discipline demeure cependant ancrée dans l'objectif de l'applicabilité et le principe de définition préalable des structures informatives relatives aux domaines traités reste prégnant dans la mise en œuvre de l'EI. L'adaptabilité s'y trouve alors concentrée sur les modes d'acquisition des ressources nécessaires — règles de reconnaissance, lexiques ou bases de connaissances — pour le traitement adéquat de domaines variés.

Les développements informatiques de la fin du XX^{e} siècle se traduisant notamment par une capacité accrue des ordinateurs en termes de mémoire, la manipulation de grandes quantités de données constitue un problème de moins en moins prégnant. La période s'accompagne donc d'un mouvement vers le recours aux données dans la constitution des modèles employés en TAL, qui se traduit notamment par l'émergence de la linguistique dite de corpus, mais également dans l'idée que les structures sous-jacentes au langage, permettant une représentation de l'information, peuvent émerger des corpus de textes eux-mêmes : Cowie et Wilks [CW00] soulignent ainsi une attention moindre portée aux théories linguistiques dans un effort de dérivation des structures à partir des données, désormais en quantité suffisante pour rendre possible les généralisations sur le langage.

Les méthodes numériques et notamment probabilistes à partir d'apprentissage supervisé, c'est-à-dire de données annotées selon les schémas informationnels visés, s'inscrivent dans ce mouvement empirique et apparaissent comme l'une des réponses apportées au besoin d'adap-

tabilité accrue. Il s'agit d'ancrer les capacités des systèmes d'EI dans une représentation dérivée des données elles-mêmes, mais également de contourner le problème du coût engendré par la conception manuelle de systèmes : celle-ci nécessite la mise au point de règles et motifs de reconnaissance par des experts, ainsi qu'un processus d'acquisition de ressources, lexicales ou relevant des connaissances du monde. L'annotation de corpus se présente comme plus aisée et favorise une distinction des compétences entre expertise du domaine et implémentation informatique. Le coût de l'annotation n'étant lui-même pas négligeable, les méthodes d'apprentissage semi-supervisé ou non supervisé connaissent également un intérêt grandissant, surtout à partir des années 1990. Moens et De Busser [Moe06] soulignent cependant que la majorité des travaux à base d'apprentissage portent sur l'acquisition des motifs pour la reconnaissance et le typage d'entités, le repérage des relations entre entités, la classification de rôles sémantiques et la résolution d'expressions temporelles, c'est-à-dire sur les éléments composant les scénarios sous-jacents aux tâches d'EI. L'apprentissage des scénarios eux-mêmes ou des scripts qui leur sont associés, à la manière de ce qui est proposé par la CDT (cf. *supra* 2.1), reste rare. Une présentation détaillée des différentes méthodes, symboliques et probabilistes, employées depuis les débuts de l'EI jusqu'aux systèmes récents, est proposée par Moens [Moe06].

La systématisation permise notamment par des campagnes d'évaluation dédiées se traduit par une architecture générale et commune, dans un espace de variation autour de caractéristiques typiques de tout système d'EI. Nous reprenons ici une description synthétique de cette architecture à Moens et De Busser [Moe06], en regard du schéma reproduit à la figure 2.4. Ce dernier illustre

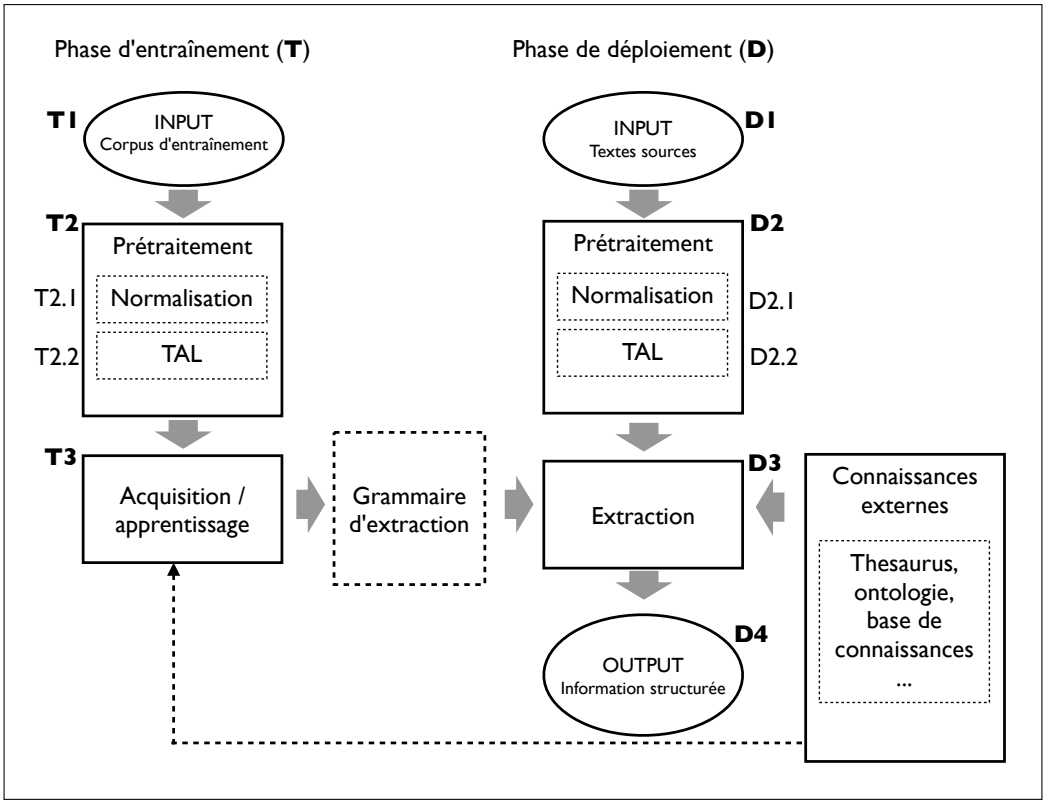


FIGURE 2.4 : Architecture générale d'un système d'EI (adapté de [Moe06]).

les deux phases distinctes d'un système d'EI : entraînement et déploiement. La première concerne l'acquisition des motifs d'extraction, réalisée par l'intervention d'un spécialiste humain ou par le biais d'un apprentissage automatique. La première étape de sélection d'un corpus textuel

représentatif de la tâche (fig. 2.4, T1) est suivie d'un prétraitement (fig. 2.4, T2) au cours duquel le texte est normalisé grâce à des outils usuels de TAL, notamment la segmentation en phrases et en mots (fig. 2.4, T1.1), et enrichi d'indications linguistiques (fig. 2.4, T1.2), par exemple un étiquetage en parties du discours, qui pourront être utilisées lors de l'acquisition. Dans cette dernière étape (fig. 2.4, T3), le corpus prétraité est utilisé comme base pour la conception de règles d'extraction dans le cas d'une approche manuelle ; dans le cas d'un apprentissage automatique, il est d'abord annoté selon les éléments textuels pertinents pour la tâche puis fourni à l'algorithme adéquat pour l'induction de la grammaire d'extraction correspondante[1]. L'élaboration ou l'induction de cette grammaire peut par ailleurs faire usage de connaissances externes accesibles au système.

Durant la phase de déploiement (fig. 2.4, D1 à D4), le système d'EI repère et classifie les informations pertinentes dans de nouveaux textes, distincts du corpus d'entraînement. Le composant de prétraitement (fig. 2.4, D2) est aussi similaire que possible à celui de la phase d'entraînement (fig. 2.4, T2). Le texte est ensuite traité par le composant d'extraction (fig. 2.4, D3) qui repose sur la grammaire acquise lors de l'entraînement et pouvant se référer à d'éventuelles connaissances externes. Les éléments textuels pertinents sont ainsi extraits, organisés relativement aux classes définies pour la tâche puis retournés dans le format structuré correspondant (fig. 2.4, D4). Les études sur l'EI portent usuellement moins sur le déploiement réel et concret d'un système que sur son développement et sur les tests associés ; la phase de déploiement correspond alors le plus souvent à une évaluation du système. Cette architecture générale se trouve notamment réalisée dans le système d'EI GATE [Cun+11b], développé depuis 1995 principalement à l'Université de Sheffield et toujours maintenu.

Après ces observations générales sur l'évolution historique de la discipline et ses caractéristiques principales, il nous est possible d'arrêter une définition stable de l'EI sur laquelle une réflexion concernant sa place dans le processus d'annotation sémantique peut s'appuyer. En suivant une nouvelle fois les formulations de Moens et De Busser [Moe06], cette définition peut prendre la forme suivante :

> [2.1] L'Extraction d'Information consiste en une opération de repérage et de structuration en classes sémantiques, par classification consécutive ou simultanée, d'éléments informatifs spécifiques présents dans des données non structurées, notamment textuelles, menée dans le but de donner à l'information une forme adéquate pour des traitements automatiques.

Comme évoqué plus haut (cf. *supra* 1.3) et en regard de cette définition, la fonctionnalité de repérage des éléments informatifs pertinents en fonction d'un domaine donné confère à l'EI un rôle indispensable pour une automatisation de l'Annotation Sémantique. Dans cette perspective, la relation au domaine constitue un aspect crucial de l'EI et en détermine le modèle de structuration sous-jacent.

2.3 Une sémantique par classification : des formulaires aux ontologies

La présentation de l'EI et de sa relation à l'Annotation Sémantique a souligné, dans le début de ce chapitre, son lien fondamental avec la définition préalable d'un domaine. De cette association découle une structuration de l'information reflétant une conceptualisation du domaine considéré et de la vue particulière adoptée à son égard dans le cadre de l'opération d'EI. La spécification de la conceptualisation adéquate est accomplie en amont du déploiement d'un système d'EI, celui-ci devant s'y conformer afin de produire les résultats attendus en termes de structuration

1. Le corpus peut être seulement partiellement annoté ou ne recevoir aucune annotation dans le cas de méthodes d'apprentissage faiblement ou non supervisé.

des éléments informatifs extraits. Les modalités de reconnaissance de ces éléments — règles d'extraction élaborées manuellement ou motifs acquis automatiquement à partir d'annotations — sont en effet conçues en fonction de la structuration donnée.

L'effort de conceptualisation associé au domaine traité rend compte, par la sélection des éléments à extraire ainsi que de leurs relations, du sens que l'on souhaite attribuer à l'information ainsi rendue disponible. La structuration correspondant à cette conceptualisation est en effet le moyen par lequel les traitements envisagés en aval de la collecte effectuée par l'EI peuvent s'appliquer. Concrètement, une extraction similaire à l'exemple de la table 2.1 permet, par le typage des données en ORGANISATION ou ANNÉE de conduire de façon systématique des opérations liées à ces types : l'identification des recrutements effectués par une entreprise entre deux années données, par exemple, peut être obtenue grâce à la formulation d'une requête lancée sur l'ensemble des informations extraites et construite en référence aux types de données spécifiés dans la conceptualisation sous-jacente. Ceci implique que l'on dispose pour ces types d'une algèbre définie, avec par exemple les opérateurs leq et geq pouvant s'appliquer aux données de type DATE, mais pas ORGANIZATION.

La forme prototypique de la conceptualisation en EI est celle du formulaire, introduite en tant que structure fondamentale par les campagnes MUC. Comme évoqué précédemment, un formulaire correspond à un scénario ou à un élément central du domaine, tel que les entités ; il présente un ensemble de champs correspondant aux caractéristiques définitoires de ces scénarios et entités, choisies en fonction de la perspective particulière adoptée sur le domaine considéré. Les formulaires de scénarios représentent un ensemble informatif complexe, dans lequel interviennent des entités mises en relation par des prédicats pertinents pour sa description. La figure 2.5 illustre la décomposition des scénarios en champs ou *slots*, qui renvoient à des caractéristiques de l'événement ainsi modélisé — sa date, par exemple —, ainsi qu'à des entités dont la relation avec l'événement est exprimée par le champ en question — l'auteur d'un attentat, par exemple. Les formulaires d'entités correspondent en revanche à des unités informatives auxquelles ils

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

FIGURE 2.5 : Formulaire pour les scénarios d'actes terroristes (MUC-3, extrait de [Ehr08]).

attribuent une formalisation et ainsi un *type* ou une *classe*. Chaque champ correspond à une propriété considérée comme inhérente à la classe ainsi définie, comme le montre la figure 2.6 où les formulaires pour les personnes et les organisations conçus pour MUC-6 présentent les types de valeurs acceptées par chaque champ. Les formulaires de ce type constituent une norme dans le cadre de l'EI défini par les campagnes MUC, mais s'inscrivent également dans le mode de représentation des structures de connaissances proposé par la théorie des cadres, évoquée en 2.1 et reposant sur des champs aux valeurs typées ainsi que sur une organisation en réseau formant

```
<ORGANIZATION> :=
        ORG_NAME:               "NAME"-
        ORG_ALIAS:              "ALIAS"*
        ORG_DESCRIPTOR:         "DESCRIPTOR"-
        ORG_TYPE:               {GOVERNMENT, COMPANY, OTHER}^
        ORG_LOCALE:             LOCALE-STRING {{LOC_TYPE}} *
        ORG_COUNTRY:            NORMALIZED-COUNTRY | COUNTRY-STRING *
        ORG_NATIONALITY:        NORMALIZED-COUNTRY-or-REGION | COUNTRY-or-REGION-STRING *
        OBJ_STATUS:             {OPTIONAL}-
        COMMENT:                " "-

<PERSON> :=
        PER_NAME:               "NAME"^
        PER_ALIAS:              "ALIAS"*
        PER_TITLE:              "TITLE"*
        OBJ_STATUS:             {OPTIONAL}-
        COMMENT:                " "-
```

Figure 2.6 : Formulaires d'entités (MUC-6, d'après [Gril2]).

un ensemble de description sémantique.

L'analogie avec les cadres souligne la fonction conceptuelle des formulaires, destinés à représenter les éléments informatifs d'un domaine donné. S'ils constituent le mode usuel de structuration de l'information en EI, les formulaires constituent seulement un des moyens d'ancrer les descriptions dans une sémantique définie. Plus généralement, il s'agit de spécifier une conceptualisation relativement à une vue particulière d'un domaine, en vue d'en trouver des réalisations sous forme textuelle. Comme cela a été évoqué au chapitre 1, une telle conceptualisation peut prendre la forme d'une taxonomie établissant un ensemble de classes conceptuelles correspondant aux éléments structurants du domaine. Sous une forme hiérarchisée, c'est-à-dire munie de relations de sous-classes, avec la spécification d'attributs relatifs aux classes définies et de relations conceptuelles entre classes, une taxonomie prend le tour d'une ontologie, telle que définie au chapitre 1 (section 1.2.2).

L'utilisation d'ontologies en tant que moyen de spécification conceptuelle et de structuration sous-jacent à la tâche d'EI repose sur leur adéquation quant à la formalisation des connaissances et de la sémantique qui leur est associée dans la perspective d'une interprétabilité des informations extraites. Un système d'EI lié à une ontologie procède donc au repérage des éléments informatifs dans les contenus traités, puis à leur classification au sein de l'ontologie adoptée. Chaque élément ainsi classifié est vu comme la réalisation linguistique du concept sélectionné, ou d'une instance de ce concept, selon que le segment textuel en question en constitue une dénotation générique ou particulière — ce second cas concernant les entités, auxquelles nous nous intéresserons plus précisément dans la suite de ce travail. Les attributs de classes ainsi que les relations entre concepts peuvent également faire l'objet d'une association avec le texte si la tâche d'EI en définit les règles.

L'apport majeur de l'ancrage de contenus dans une modélisation ontologique réside dans la formalisation logique caractérisant les connaissances ainsi représentées. À cet égard, l'EI ainsi envisagée peut être reformulée en une opération de *population d'ontologie*, comme cela est proposé par Nédellec et al. [NNB09] dans une étude consacrée à cette relation. Il y est également souligné que cette relation permet de rassembler et de structurer des connaissances sur un domaine *a priori*, et de les rendre ainsi disponibles aux outils d'EI au cours du processus d'extraction lui-même. L'EI ainsi guidée par une ontologie, réalisée notamment par Buitelaar et al. [Bui+08], constitue un mode d'accès à l'information reposant sur des connaissances non exclusivement linguistiques et surfaciques, mais également de type conceptuel.

L'ancrage de l'EI dans une modélisation de type ontologique s'illustre notamment dans des campagnes d'évaluation telles que TAC (Text Analysis Conference), que Grishman qualifie de successeur de MUC [Gril2] tout en en soulignant le caractère novateur porté par la tâche de population

de bases de connaissances (Knowledge Base Population, KBP). Contrairement aux composants de MUC, cette tâche dépasse l'unité documentaire dans le processus de collecte d'informations, qu'il s'agit d'agréger à partir d'un ensemble documentaire avec pour cible d'ancrage une base rassemblant les entités d'intérêt en une structure de représentation unifiée et ontologique. La tâche KBP de TAC s'apparente en effet à l'EI en tant qu'elle vise à une structuration de l'information à partir de données textuelles, mais elle s'en distingue également par la nature et l'étendue des éléments qu'il s'agit de repérer ; ces aspects particuliers feront l'objet d'un examen dans la suite de ce travail, au chapitre 3 (section 3).

La vue générale de l'EI proposée ici se situe dans la perspective d'une intégration dans le processus d'Annotation Sémantique. Les modalités d'accès à l'information permises par l'EI ont donc été mises en avant afin d'expliquer dans quelle mesure cette intégration est pertinente et sur quels aspects particuliers l'Annotation Sémantique peut s'appuyer pour faire émerger de façon explicite les éléments informatifs pertinents relativement à l'objectif d'enrichissement de contenus textuels. À cet égard, il convient d'orienter cette présentation vers le cas des entités, qui constituent la cible première de l'enrichissement dans notre contexte de travail. En tant qu'objets manipulés par l'EI, elles présentent des problèmes spécifiques quant à la définition de la sémantique qui leur est attribuée. La question de l'adéquation de cette sémantique, déterminée par la tâche d'EI et le TAL, aux fonctionnalités attendues par l'enrichissement se pose de façon centrale et permet de délimiter le rôle de l'EI dans l'Annotation Sémantique, qui pourra ensuite faire l'objet d'une définition plus avancée (chapitre 3).

3 Entités et entités nommées

Les entités ont été évoquées à plusieurs reprises au cours de la description de la tâche d'EI, dans laquelle elles occupent de fait une place centrale. En tant qu'ensemble regroupant notamment des personnes, lieux ou organisations, leurs réalisations linguistiques dans les contenus textuels, que désigne le terme *entités nommées*, constituent les éléments informatifs principaux des structures qu'il s'agit de dériver des textes par l'EI. Quel que soit le domaine traité, les entités sont en effet les actants constitutifs des événements rapportés. L'intérêt qui leur est porté dans le cadre de l'enrichissement de contenus est d'ailleurs lié à cette place centrale.

Le repérage des entités nommées dans un système d'EI correspond ainsi en premier lieu à une nécessité méthodologique pour la compositions des structures informatives attendues, mais acquiert également le statut de module indépendant, voire de tâche autonome motivée par l'intérêt informationnel des entités elles-mêmes. Conjointement à la place qui leur est faite en EI, centrée sur leur reconnaissance et leur classification, les entités nommées sont un objet d'étude et de questionnement pour le TAL en général, en tant que leur définition et les modalités de leur repérage se situent à la croisée de la linguistique et de la représentation des connaissances. La réalisation linguistique des entités sous la forme d'entités nommées souligne par ailleurs de façon prégnante le problème pris en charge par l'EI du passage entre texte et représentation, d'autant que les phénomènes d'ambiguïté qu'elle présente sont particulièrement prégnants et polymorphes. Ceux-ci interrogent en particulier le statut référentiel des entités nommées, renvoyant à des éléments extra-linguistiques — les entités elles-mêmes —, auxquels l'EI ne s'intéresse pas directement. Au-delà de la reconnaissance des entités nommées et de la structuration par classification inhérente à l'EI, les entités posent ainsi le problème de la référence et de sa représentation, notamment dans la perspective de l'Annotation Sémantique et de l'acquisition de métadonnées.

3.1 La Reconnaissance d'Entités Nommées

En tant qu'éléments informatifs constitutifs des structures visées par l'EI — scénarios, événements, prédications pertinentes pour le domaine considéré —, les entités se dégagent de la tâche globale pour former une cible d'analyse particulière. Il s'agit de l'illustration la plus probante du caractère modulaire atteint par l'EI à l'issue de ses principaux développements dans les années 1990 : la reconnaissance des entités dans les contenus analysés constitue en effet une étape incontournable dans le processus de structuration, et c'est à ce titre que la campagne MUC propose à partir de sa sixième édition en 1995 une sous-tâche de *Reconnaissance d'Entités Nommées* (ci-après REN). Celle-ci se présente alors non seulement un composant de l'EI, de ceux que ses initiateurs cherchent à identifier comme modules génériques et intégrables dans une tâche globale, quelle que soit sa nature et le domaine traité, mais elle devient également une tâche à part entière, dont les résultats peuvent être exploités pour eux-mêmes. Cette tâche est caractérisée par le repérage puis le typage de segments textuels dénotant des entités dont la nature est préalablement définie. Évalués sur l'anglais, plusieurs systèmes de REN participant à MUC-6 obtiennent une F-mesure supérieure à 0.90 et la meilleure performance atteint 0.96 — qualité qui peut être relativisée par la taille réduite, le caractère homogène et la régularité rédactionnelle du corpus utilisé, comme le rappelle Ehrmann [Ehr08] à la suite de Sundheim ; Grishman et Sundheim [Sun95 ; GS96]. [2]

À la suite de l'apparition et du succès de la REN dans le cadre de MUC, la tâche suscite un intérêt grandissant — deux des systèmes de MUC-6 sont commercialisés [Ehr08] — et donne lieu à des campagnes d'évaluation dédiées :

MET (*Multilingual Entity Task*) se tient parallèlement à MUC-6 et MUC-7 et est consacrée à la REN dans des langues autres que l'anglais : l'espagnol, le japonais et le chinois [MOC96].

ACE De 2000 à 2008, les campagnes ACE (*Automatic Content Extraction*) [Dod+04] succèdent à MUC avec des tâches de reconnaissance d'événements, de relations, de scénarios et surtout de REN. L'effort se concentre dans le cadre d'ACE sur la mise au point de technologies innovantes et fiables plutôt que sur le caractère applicatif de la tâche, à la manière de MUC. Une perspective plus sémantique s'ouvre dans la caractérisation faite des entités, comme le soulignent Maynard et al. [MBC03]. Leur identification est recherchée au niveau conceptuel et non plus seulement au niveau surfacique des chaînes de caractère, avec notamment une sous-tâche de résolution de coréférence pour la constitution de chaînes référentielles.

CoNLL Deux éditions (2002 et 2003) de CoNLL (*Conference on Natural Language Learning*) proposent une tâche de REN pour l'espagnol, le hollandais, l'anglais et l'allemand [TKS02 ; TKSDM03].

ESTER Dans le cadre du projet d'évaluation des technologies de la langue Evalda [3], la campagne ESTER a été menée de 2002 à 2006 pour l'évaluation des systèmes de transcription sur des corpus d'émissions radiophoniques en français et l'enrichissement des transcriptions à l'aide d'informations telles que les EN.

Quae<ro Dans le cadre du programme de recherche et d'innovation industrielle Quaero [4], une définition étendue et renouvelée des entités nommées est proposée, associée à une attention particulière portée à leur structuration interne ainsi qu'à leur annotation et l'évaluation de leur reconnaissance pour le français [Ros+11].

2. La métrique utilisée pour le calcul de ces performances, introduite lors de MUC-5, accorde par ailleurs des scores positifs aux résultats partiellement corrects — frontières erronées et/ou type incorrect — au lieu de les considérer comme faux, ce qui contribue à l'obtention de taux de réussite élevés.

3. `http://www.elda.org/rubrique69.html`

4. `http://www.quaero.org/`

En tant que composant autonome, la REN devient par ailleurs un sujet de recherche et de développement important en TAL. L'identification des segments correspondant à des EN se présente en effet comme une étape utile à des traitements de plus large portée. Ehrmann [Ehr08] rend compte des différentes tâches pouvant bénéficier d'une intégration de la REN :

- L'analyse syntaxique obtient ainsi des informations de segmentation et d'étiquetage au niveau des parties du discours permettant d'éviter des analyses non pertinentes : par exemple, l'analyse de *loan* en verbe transitif dans *Dinosaur Savings & Loan* ; plus généralement, la REN indique le caractère syntaxique atomique de certaines entités nommées, sont il s'agit de ne pas analyser la structure interne à ce niveau. Elle peut également s'appuyer sur la catégorisation des EN dans la dérivation de dépendances syntaxiques.

- La résolution de coréférence s'appuie sur la REN grâce à l'identification d'une partie des éléments de la chaîne référentielle, ceux-ci pouvant consister en des noms propres, des groupes nominaux ou des pronoms, et à la classification des EN : un typage d'EN en PERSONNE fournit, par exemple, une information utile à la résolution anaphorique de pronoms. Dans l'énoncé suivant, l'ambiguïté de référence du pronom souligné peut être levée grâce au type sémantique des arguments attendus par le verbe de la seconde phrase :

  > La ministre Christiane Taubira a défendu la loi sur le mariage pour tous à la tribune de l'Assemblée. Elle a été votée le 12 février.

- La REN peut assister la désambiguïsation lexicale [IV98] dans l'analyse des restrictions de sélection : la classification des EN lui fournit des indications sur le type sémantique des arguments relatifs aux prédicats (verbes, noms...) dont il s'agit d'identifier le sens. Les deux sens de *quitter* peuvent être distingués dans l'expression *quitter Paris*, ainsi que dans *quitter l'UMP*, si le type de son argument dans chacun des deux cas est pris en compte (exemple d'Ehrmann [Ehr08]).

- La traduction automatique peut opérer une distinction entre segments à translittérer et segments à traduire à partir des indications d'EN et de leur type, comme pour le nom *Jack London* pour lequel une traduction par *Jack Londres* est non pertinente (exemple d'Ehrmann [Ehr08]).

On peut observer de façon générale que la REN, grâce au repérage des segments textuels correspondant à des EN, permet aux autres composants de tâches de TAL de disposer d'une segmentation adéquate du texte donné en entrée, et ainsi d'optimiser les opérations en aval en évitant un certain nombre de redondances et d'ambiguïtés au niveau de l'analyse et de la reconnaissance.

L'intérêt du TAL pour la problématique des EN se porte également sur le problème de définition qu'elles posent, dès lors qu'il s'agit, comme dans toute tâche d'EI, de déterminer les types d'objets à d'obtenir par structuration des données textuelles. Trois grandes catégories d'EN sont généralement identifiées : noms, quantités, dates et durées. La question du périmètre définitoire des EN a par ailleurs donné lieu à des études approfondies, principalement celle d'Ehrmann [Ehr08], notamment autour de la notion de sens attribuée aux éléments désignés ou non comme EN selon les tâches et applications. La résolution de cette question demeure hors de notre sujet d'étude, pour lequel les EN sont réduites aux noms propres et donc à l'ensemble d'entités pour lesquelles la dénotation peut fonctionner à l'aide de noms propres. Cette restriction n'est cependant pas étrangère aux pratiques générales en REN, où sont au moins considérés les types PERSONNE, ORGANISATION et LIEU. On peut également observer que certaines tâches de REN visent la reconnaissance des entités sous la forme de descriptions définies, telles que *le président de la République française*, tandis que notre étude se limite aux noms propres, que l'on peut considérer comme la forme prototypique d'EN, en tout cas en ce qui concerne les trois types de base retenus.

Au niveau méthodologique, la REN hérite des propriétés principales de l'EI. Il s'agit d'une part de repérer les segments jugés pertinents, ici les EN, et d'autre part de les classifier selon un modèle défini au préalable. Celui-ci correspond en REN à des classes permettant de définir et de distinguer les différents types d'EN auxquelles la tâche s'intéresse. En termes de TAL, la REN s'apparente à l'étiquetage en parties du discours — assignation d'une étiquette choisie parmi un ensemble défini (nom, verbe, adjectif, préposition...) aux unités obtenues par segmentation du texte —, ou au chunking — segmentation du texte en constituants et étiquetage des constituants selon leur catégorie syntaxique. Plus généralement, on peut la voir comme une opération d'étiquetage de séquences au sein d'un signal linguistique textuel, éventuellement transcrit de l'oral. Dans ce type de tâche, les indications exploitées dans les données traitées sont de deux types :

- La forme surfacique des segments est examinée. Pour les EN, la casse typographique est à ce titre pertinente dans les langues où les majuscules initiales signalent généralement un nom propre, sauf en début de phrase (français, anglais...) ; cette indication est moins utile, voire hors de propos, dans des langues comme l'allemand, où tout nom, propre ou commun, est capitalisé à l'initiale, ou l'arabe, qui n'opère pas de distinction entre minuscules et majuscules dans sa typographie. La classification des EN peut également relever de la forme des segments : un nom de personne se présente par exemple régulièrement sous la forme d'un prénom suivi d'un nom ; une séquence de deux mots capitalisés à l'initiale peut ainsi orienter la classification vers le type PERSONNE. Dans un segment comme *Time Bank Inc.*, la présence du token « Inc. » indique l'appartenance à une classe rassemblant les organisations.

- Le contexte des segments, plus ou moins immédiat, permet de renforcer ou d'écarter des possibilités d'étiquetage : des marqueurs lexicaux identifiés, tels que les titres (*Monsieur*, abrégé en *M.*, ou *Dr*) ou les noms de fonction (par exemple *le président* dans « le président Hollande »), signalent avec un fort degré de certitude la présence d'une EN à leur droite, du moins en français.

La REN s'appuie souvent, en plus de ces *indices internes* (forme des segments) et *externes* (marqueurs contextuels, notamment lexicaux), selon la formulation de McDonald [McD96], sur un ensemble de ressources généralement constituées d'un ou plusieurs *lexiques*, également nommés par le terme anglais *gazeteer*. Ceux-ci peuvent rassembler, sous forme de liste, des quantités variables de noms, collectés à partir de ressources externes. Les éléments ainsi listés sont par ailleurs typiquement munis d'une indication de type, correspondant *a priori* aux classes recherchées dans une tâche donnée. Les lexiques peuvent également rassembler des éléments partiels utiles à la REN, notamment dans le cas de lexiques de prénoms.

Comme en EI, la mise en œuvre de la REN peut s'appuyer sur des méthodes :

- symboliques, où les règles de reconnaissance et de classification sont élaborées par un spécialiste humain ; elles prennent le plus souvent la forme de grammaires locales reposant sur des automates et transducteurs ou des grammaires non contextuelles. Le système GATE [Cun+11b], par exemple, implémente la REN à l'aide de transducteurs couplés à un lexique d'EN. Dans l'approche d'Ehrmann [Ehr08], les EN sont repérées *via* une analyse syntaxique : tout nom ou groupe nominal dont la tête est capitalisée à l'initiale constitue une EN potentielle.

- numériques, où le processus d'extraction repose sur un apprentissage automatique probabiliste, le plus souvent supervisé, c'est-à-dire utilisant des données annotées. Le système LIANE [BC10] en est un exemple, entraîné sur le corpus de la campagne ESTER évoquée plus haut et reposant sur le modèle statistique des CRF [LMP01] pour l'apprentissage. Le

système Stanford NER [FGM05], reposant également sur les CRF, connaît une large diffusion et fonctionne sur l'anglais, l'allemand et le chinois.

- hybrides, avec notamment l'intégration de données quantifiées dans le processus de génération de règles [Lin98 ; Noul2]

L'identification claire de la tâche de REN dans le cadre de campagnes d'évaluation, mais également les nombreux développements en la matière sortant du seul champ académique et pris en charge au niveau industriel, sont accompagnés et encouragés par de bons résultats. La F-mesure correspondante dépasse 90% pour l'anglais lors de MUC-7 [MP98], quand les autres langues donnent lieu à des scores relativement moins bons mais toujours satisfaisants pour l'efficacité attendue (86% pour le japonais lors de la campagne IREX [SI99], 75% environ pour le français lors de la campagne ESTER 2[5] [GGC09]). Ce succès témoigne de l'utilité, voire de la nécessité d'un recours à la REN automatique lorsqu'il s'agit de mettre en œuvre un accès aux entités constituant une part essentielle des connaissances véhiculées par les contenus textuels.

Bénéficiant de traitements efficaces et éprouvés, la tâche de REN est en grande partie définie par des problèmes auxquels les techniques de TAL cherchent à apporter une solution : ces problèmes, centrés sur la question du repérage et de la classification sémantique, constituent donc les points principaux de développement et d'innovation proposés en REN. On peut caractériser la problématique générale comme un cas d'ambiguïté double, que la REN vise à lever, le plus souvent de façon jointe. En effet, le repérage des EN consiste en premier lieu en une segmentation adéquate du texte donné en entrée, c'est-à-dire en l'indication des frontières sur l'axe syntagmatique au sein desquelles une EN est présente. Le typage de l'EN ainsi localisée linéairement relève d'un niveau de représentation différent du texte — l'axe paradigmatique des classes définies —, mais peut dériver d'informations communes au processus de localisation. Les règles d'extraction élaborées pour un système de REN peuvent ainsi définir concomitamment les critères de localisation et de typage, voire être elles-mêmes typées (en formulant par exemple la présence d'une EN de type PERSONNE lorsque deux mots inconnus du lexique se suivent et sont capitalisés à l'initiale, ou la présence à droite du marqueur lexicalisé « la ville de » d'une EN de type LIEU). L'ambiguïté découle ici, d'une part, du fait qu'un même segment de texte peut recouvrir plus d'une EN, ou être partiellement commun à plus d'une EN et, d'autre part, du typage possiblement non univoque des segments repérés. L'ambiguïté de segmentation, apparaissant lorsque plusieurs règles de reconnaissance sont applicables à une même région de texte, peut être levée par une pondération des règles, donnant la priorité à l'une d'elles ; dans un système numérique, cette priorisation est souvent inhérente au modèle résultant de l'apprentissage. L'ambiguïté de type subsiste en revanche même après la levée d'une ambiguïté de segmentation, lorsqu'un segment unique peut être classifié de façon multiple. L'exemple suivant illustre cette ambiguïté double :

(1) Le maire d'Orange Alain Labé enseignait l'histoire et la géographie.

Si la grammaire d'extraction correspondante définit une règle de reconnaissance des noms de personne par une succession de plusieurs mots capitalisés à l'initiale — pour assurer le repérage de noms tels que *Célimène Mater Durand*[6] —, ainsi qu'un lexique indiquant que la chaîne de caractères *Orange* peut correspondre à une EN de type LIEU ou à une EN de type ORGANISATION, la région soulignée peut alors donner lieu à plusieurs découpages :

5. Les différentes performances de REN selon les langues ne sont comparables que de façon limitée, puisque les données d'évaluation ainsi que les consignes d'annotation et de conduite de la tâche diffèrent pour chacune d'elles et pour chaque campagne.

6. Exemple extrait d'un article du blog `http://www.maitre-eolas.fr/` sur le problème juridique du double nom de famille

- [Orange Alain Labé], un nom de personne
- [Orange] et [Alain Labé], un nom de lieu, suivi d'un nom de personne
- [Orange] et [Alain Labé], un nom d'organisation, suivi d'un nom de personne

L'ambiguïté de découpage dans cet exemple correspond à plusieurs analyses possibles, dont une seule s'avère correcte en termes d'interprétation. Elle peut en revanche relever de l'interprétation elle-même, lorsque plusieurs découpages sont possibles ; les frontières considérées comme correctes dépendent alors de critères posés *a priori* pour la tâche courante. Les exemples suivants illustrent ce type d'ambiguïté, qui peut être provoqué par la coordination d'EN (2) ou l'imbrication d'EN (3) :

(2) Bill and Hillary Clinton are to visit Northern Ireland on Friday [7]

(3) l'Université de Corte [8]

Pour prendre en charge de tels cas, il est nécessaire de déterminer au préalable si une seule ou deux EN doivent être extraites dans *Bill and Hillary Clinton* (exemple 2), et si la cible d'extraction est l'EN englobante ou l'EN imbriquée (exemple 3).

L'ambiguïté relevant du type à assigner aux EN repérées pose quant à elle un problème inhérent à la notion d'entité et à son rapport avec la réalisation linguistique que constituent les EN. Bien que le contexte surfacique d'occurrence des EN puisse dans certains cas déterminer leur type, notamment lorsqu'un marqueur externe agit comme classificateur (par exemple « la ville » dans *la ville d'Orange*), la classification dont il s'agit porte sur les EN en tant qu'elles dénotent des entités, c'est-à-dire des objets extra-linguistiques, et non en tant que formes pour elles-mêmes.

3.2 Portée et limites de la sémantique typologique des entités

L'aspect structurant de l'EI se trouve réalisé en REN dans la tâche de classification des EN repérées, à l'aide de catégories qualifiées de *sémantiques*. Ces catégories enrichissent les extractions d'indications à même de les rendre interprétables en aval. Ces catégories font l'objet d'une explicitation préalable, la classification ainsi réalisée faisant office de modèle pour la tâche d'EI considérée. Les entités constituant des éléments centraux dans le processus de traitement de l'information, le choix des types correspondant occupe une place importante dans les développements de la tâche de REN.

Il s'agit en premier lieu de déterminer les types d'entités répondant de la façon la plus immédiate au besoin sous-jacent à l'accomplissement de la REN, c'est-à-dire ceux dont la valeur informative est incontournable. Les différentes éditions de MUC proposent ainsi de s'intéresser aux noms de personnes, de lieux et d'organisations, ainsi qu'aux dates, expressions temporelles, valeurs monétaires et pourcentages. La campagne ACE ajoute notamment les noms de bâtiments, véhicules et armes ; les noms de produits sont considérés dans la campagne IREX [SI99], dédiée au japonais. Mais la définition de la classification répond également à un objectif plus large et distinct de la tâche elle-même, consistant en une volonté de modélisation exacte et pertinente du monde tel qu'il peut être envisagé à travers les catégories existantes d'entités manipulées dans le langage. La tâche de REN étant largement définie les campagnes d'évaluation, la nécessité de produire des données annotées manuellement encourage cette réflexion sur les choix à opérer en termes de classification, et les cas de doute, d'hésitation et ou de désaccord y participent d'autant

7. Exemple extrait du journal en ligne Belfast Telegraph, 29 novembre 2012
8. Exemple emprunté à [Ehr08]

plus. Un exemple de difficulté d'annotation notoire est celui des noms de lieux pouvant également désigner des entités de type institutionnel, agissant comme des organisations. Les noms de pays, notamment, peuvent être employés dans des prédications attachées à l'entité gouvernante du pays en question :

(4) La France signe le traité de Versailles en 1919.

Ce cas conduit à la création du type GPE (*geo-political entities*) pour la tâche de REN de la campagne ACE, ainsi que du type GSP (*groupe géo-socio-politique*) pour l'annotation du corpus ESTER. En 2002, Sekine et al. [SSN02] proposent un modèle de classification d'EN destiné à la REN consistant en une hiérarchie de 150 types, communs et généraux (personne, organisation, événement...) d'une part, précis et de granularité conceptuelle fine d'autre part (monnaie, journal, parti politique, crime...).

Comme cela a été évoqué, l'effort de modélisation, dans lequel s'inscrit la classification des EN, et par là même des entités, est nécessairement circonscrit à une vision particulière du monde ou d'un domaine et non exhaustif quant à sa fidélité aux catégories existantes, qu'une opération d'inventaire ne saurait identifier définitivement. On peut par ailleurs observer qu'une part de la complexité inhérente à l'établissement d'une classification revient au problème de typages non triviaux, dans des cas comme celui des GPE mentionné plus haut. Une fois les catégories établies, le choix de classification pour les EN relevant de tels cas se heurte à la difficulté de l'ambiguïté, qu'un système de REN doit être à même de lever.

L'ambiguïté touchant l'assignation d'un type aux EN repérées peut être considérée, si l'on suit entre autres l'analyse d'Ehrmann [Ehr08], comme un problème de polysémie, qui découle, selon les cas :

- d'une *homonymie*, lorsque plusieurs entités peuvent être dénotées par la même expression linguistique, de façon coïncidentelle ; c'est le cas de la ville française d'Orange et de l'entreprise de télécommunications Orange, qui n'entretiennent aucune relation sémantique ou extra-linguistique ;
- d'une *métonymie*, lorsque le nom d'une entité est employé pour en dénoter une autre, les deux entités en question entretenant une relation (cause et effet, contenant et contenu, créateur et artefact, lieu et occupant du lieu, lieu et institution qui y est hébergée...). Dans

  (5) Marseille a gagné 2-0 en finale de la Ligue des Champions

  la chaîne *Marseille* désigne non la ville française mais l'équipe de football locale, l'*Olympique de Marseille*.
- de *facettes* pouvant être attachées à une entité, lorsque celle-ci peut être vue de différentes façons, notamment dans le cas de différentes fonctions occupées par une personne dans une organisation : *François Hollande* peut par exemple être une expression de même sens que *Premier Secrétaire du PS* ou *Président de la République française*, selon la date d'énonciation ; on retrouve ici la distinction entre *sens* et *dénotation* formulée par Frege [Fre92 ; FI71], qui permet de considérer les facettes d'entités comme cause de polysémie.

L'homonymie est en jeu dans l'exemple 1 (section 3.1), où il importe non seulement de repérer l'EN correspondant au segment textuel *Orange*, mais également de lui attribuer le type adéquat, choisi entre LIEU et ORGANISATION, si ces deux types sont définis dans le modèle sous-jacent. La REN vise à rendre compte de la métonymie en assignant à l'EN le type correspondant à l'entité effectivement dénotée, et non à celle dont le nom est employé. Homonymie et métonymie constituent ainsi les

deux objets principaux des nombreux travaux attenants à la REN portant sur la *désambiguïsation* d'EN ; ce terme est en effet employé pour désigner le processus qui revient de fait à sélectionner le type adéquat pour une occurrence d'EN donnée. Quant aux facettes, elles peuvent donner lieu, plutôt qu'à une désambiguïsation à proprement parler, à un raffinement de typage. Une EN extraite pourra alors recevoir une classe de portée plus fine que les classes génériques habituelles (acteur ou chanteur plutôt que personne), et ainsi enrichir l'information apportée par la tâche de REN, comme le proposentt notamment Ehrmann [Ehr08] ou Fleischman et Hovy [FH02].

La notion de désambiguïsation dont la REN use pour caractériser le repérage des EN en cas d'ambiguïté entre types est dépendante de la définition des EN en TAL : comme le formule Ehrmann dans une thèse en grande partie consacrée à cette définition [Ehr08], les EN sont des expressions mono-référentielles étant donné un modèle défini. Le type à assigner à une EN correspond donc à celui de l'entité à laquelle il est considéré qu'elle réfère, ce qui justifie de classifier l'EN *Orange* en lieu ou organisation selon qu'il s'agit de la ville ou de l'entreprise. La modélisation adoptée en REN pour prendre en compte les ambiguïtés relevant de la polysémie est examinée par Ehrmann [Ehr08] : en cas d'homonymie ou de métonymie, il s'agit pour le système considéré de réaliser la désambiguïsation nécessaire en fonction des classes qui lui sont rendues disponibles. En cas d'ambiguïté entre deux types, comme dans le cas d'*Orange*, il ne peut y avoir désambiguïsation que si ces deux types sont définis. Dans le cas contraire, l'EN est considérée comme monosémique. La classe GPE proposée dans le cadre de la campagne ACE fournit un moyen de contournement du problème en opérant une fusion entre les types lieu et organisation, ce qui revient à modéliser le phénomène de métonymie touchant de façon régulière certains ensembles d'EN, notamment les pays et capitales.

On peut observer, dans la méthodologie généralement adoptée en REN pour aborder le problème de l'ambiguïté, que plusieurs niveaux d'analyse coexistent et peuvent, dans une certaine mesure, révéler une limitation quant à l'accomplissement de la tâche visée. En effet, si les EN peuvent être définies comme des *expressions mono-référentielles*, le traitement qui en fait en REN s'attache à une notion du sens associée au principe de la classification. En assignant un type à une EN, la REN lui attribue un sens, ce qui peut nécessiter une levée d'ambiguïté lorsque plusieurs types, autrement dit plusieurs sens, sont envisageables pour une seule forme de surface. Il convient de noter que l'association d'une EN à une classe sémantique est rendue possible par le lien dénotationnel existant entre une EN et l'entité à laquelle elle réfère, selon la définition adoptée, mais que la sémantique ainsi exprimée par la REN demeure celle de la classe considérée, et non celle de l'entité dénotée. La REN se présente ainsi, en tant que tâche relevant de l'EI, comme un moyen d'accès à l'information dont le niveau d'analyse franchit la frontière entre surface et connaissances extra-linguistiques, mais de façon partielle. Le phénomène dénotationnel n'est en effet pas pris en compte en tant que tel dans l'approche des entités par la REN, qui s'intéresse de façon privilégiée à leurs réalisations linguistiques — les EN. Les éléments informatifs ainsi retournés par la REN se présentent en effet sous la forme de chaînes de caractères, associées à un type sémantique et éventuellement regroupées en variantes d'une même forme canonique : ils ne se distinguent donc pas du niveau textuel, même s'ils sont extraits de la chaîne syntagmatique et participent de la structuration de l'information attendue à l'issue d'une tâche de REN.

La relation entretenue entre entité et EN, qui relève de la dénotation, se distingue de l'appartenance d'une EN à une classe sémantique en tant qu'elle met en jeu l'existence d'un *référent*, dont l'EN est une expression linguistique possible. La dénotation opère grâce à la possibilité de nommer les objets du monde dans le langage, mais ne véhicule pas par elle-même l'élément de connaissance constitué par l'entité considérée ; il s'agit d'un processus d'évocation, par lequel l'auditeur ou le lecteur est amené à accéder à sa propre connaissance de cette entité afin d'interpréter les énoncés qui lui sont soumis. Si la REN permet un degré d'accès à l'information en isolant et

en typant sémantiquement les éléments pertinents dans des données textuelles, elle se limite en revanche à la description de leur comportement linguistique et informatif. Il est intéressant d'observer que la majeure partie des travaux en REN approchent le concept de l'entité, à travers les EN, en s'efforçant de spécifier de la façon la plus pertinente possible le type sémantique qui peut leur être attaché, et en proposant des structures informatives complexes et détaillées — notamment par le biais des formulaires introduits par MUC — dont les éléments sont des attributs descriptifs des entités dénotées. La relation entre EN et entité n'est cependant pas clairement établie dans la mesure où la désambiguïsation opérée en REN se limite au type sémantique. Dans le cas de l'EN *Orange*, un système peut retourner l'un des deux types LIEU ou ORGANISATION en adéquation avec le contexte d'occurrence, mais ce comportement ignore un second niveau d'homonymie, puisqu'il existe dans le monde une vingtaine de villes dont le nom est *Orange*, ainsi que plusieurs entreprises du même nom. Cette approche rappelle ainsi que le terme de *désambiguïsation*, ainsi que la sémantique adoptée en REN, s'entendent au niveau classificatoire et non référentiel.

En privilégiant une analyse des entités portant sur leur réalisation linguistique et relevant d'une représentation sémantique typologique, la REN ne fait donc pas intervenir de façon explicite et intégrée leur aspect référentiel. La dichotomie entre niveaux linguistique et extra-linguistique, qui caractérise les entités nommées et les entités dans le phénomène dénotationnel, n'y est en effet pas spécifiquement représentée, tout comme les entités en elles-mêmes en tant que référents extra-linguistiques n'y trouvent pas de modélisation explicite.

3.3 Des entités nommées aux entités : prise en charge référentielle

Si la REN s'intéresse particulièrement à la réalisation linguistique des entités, d'autres tâches définies dans le cadre de l'EI, dans son prolongement ou indépendamment sont quant à elles tournées vers leur aspect référentiel. Il s'agit alors de tenir compte de l'existence de l'objet extra-linguistique, autrement dit de l'entité, qu'implique une mention par une entité nommée dans des données textuelles. L'accès à l'information et sa représentation intègrent alors la notion de référent, que celui-ci soit modélisé de façon explicite ou non. Ces variations de la prise en charge des entités, bien que ne relevant pas de la REN telle que définie précédemment, entretiennent des liens indéniables avec le cadre général de l'EI dans la mesure où elles visent une représentation structurée des éléments informatifs issus de données textuelles. C'est principalement au niveau du modèle de structuration défini pour chaque tâche que cette représentation tend à se distinguer du paradigme typologique de l'EI, pour se rapprocher davantage d'objectifs similaires à ceux de l'Annotation Sémantique dont il sera plus précisément question dans la suite de ce travail.

3.3.1 Référence interne et implicite : résolution de coréférence

À partir de la sixième édition de MUC (1995) et dans le cadre des campagnes ACE, la tâche de résolution de coréférence s'inscrit de façon précise et durable dans le cadre de l'EI. Étroitement liée à la tâche de REN, elle porte sur la mise en relation de mentions d'entités au travers des données textuelles par l'établissement de liens de coréférence existant entre ces mentions. Plus spécifiquement, la résolution de coréférence vise à regrouper un ensemble d'expressions, typiquement nominales, formant une *chaîne de coréférence* au sein d'un document ou à travers plusieurs documents d'un corpus déterminé. Elle implique l'existence d'un *référent*, correspondant à l'entité mentionnée par chacune des expressions d'un même groupe. Ce référent ne donne pas lieu à une modélisation distincte des contenus concernés : le groupe ou *cluster* résultant du processus de résolution en donne lui-même la représentation ; on peut ainsi parler, en résolution de coréférence, de référent *interne* ou *implicite*.

La résolution de coréférence partage avec la tâche de résolution d'anaphore, étudiée et traitée de plus longue date ([Hob86 ; LL94 ; GHC98 ; Mit02]), la problématique de la relation référentielle unissant des expressions linguistiques distinctes. La résolution d'anaphore est cependant restreinte aux cas asymétriques de cette relation, mettant en jeu une expression dans le rôle d'antécédent, et une autre dans le rôle d'anaphore. Cette dernière revêt ce statut en tant qu'elle contient une référence identique à celle de son antécédent, mais également dans la mesure où elle n'est pas interprétable en tant que telle, en dehors de sa relation avec l'antécédent. Les expressions anaphoriques qu'il s'agit de traiter sont ainsi typiquement des pronoms personnels. La résolution de coréférence concerne quant à elle également les expressions coréférentes dans une relation d'équivalence, symétrique et transitive, formant ainsi indifféremment un groupe ou une chaîne. La figure 2.7 illustre cette tâche, qui aboutit en un certain nombre de groupes représentant les référents (e_0, e_1, ...) mentionnés dans le texte d'entrée.

[Le député Pascal Cherki]$_{\mathbf{0}}$, qui a adressé son "mépris républicain" au [Medef]$_{\mathbf{1}}$ dans ses voeux de nouvelle année, a eu une réaction "épidermique et compréhensible", estime [le ministre Benoît Hamon]$_{\mathbf{2}}$, lui aussi classé à l'aile gauche du [PS]$_{\mathbf{3}}$, dans un entretien au [Parisien]$_{\mathbf{4}}$ dimanche.
Est-ce qu'un tel propos vous choque ? "Les élus de terrain sont confrontés tous les jours aux conséquences de la crise", répond [le ministre délégué à l'Economie sociale et solidaire]$_{\mathbf{5}}$, citant le mal-logement, le chômage et le surendettement. "On aurait tort de croire la [France]$_{\mathbf{6}}$ silencieuse. [Elle]$_{\mathbf{7}}$ gronde".
Alors, poursuit-[il]$_{\mathbf{8}}$, "quand certains patrons de grands groupes qui - crise ou pas crise - gagnent de mieux en mieux leur vie passent leur journée à répéter partout que la [France]$_{\mathbf{9}}$ est trop rigide, qu'il faut supprimer les 35 heures ou modérer les salaires... forcément ça provoque des réactions épidermiques et compréhensibles".
Mercredi, le message de voeux lancé son compte [Twitter]$_{\mathbf{10}}$ par [le député-maire PS du XIVe arrondissement de Paris]$_{\mathbf{11}}$ avait suscité l'indignation de plusieurs élus de droite et du centre, rappelant en substance que les entreprises sont nécessaires pour créer des emplois et de la richesse en [France]$_{\mathbf{12}}$.
"Bonne année à toutes et tous sauf aux exilés fiscaux et au [MEDEF]$_{\mathbf{13}}$ qui ne méritent que notre mépris républicain", écrivait dans ce tweet [Pascal Cherki]$_{\mathbf{14}}$, qui est un proche de [M. Hamon]$_{\mathbf{15}}$.
(AFP, 6 janvier 2013)

e$_0$: {Le député Pascal Cherki$_{[0]}$, le député-maire PS du XIVe arrondissement de Paris$_{[11]}$, Pascal Cherki$_{[14]}$}
e$_1$: {Medef$_{[1]}$, Medef$_{[13]}$}
e$_2$: {le ministre Benoît Hamon$_{[2]}$, le ministre délégué à l'Economie sociale et solidaire$_{[5]}$, il$_{[8]}$, M. Hamon$_{[15]}$}
e$_3$: {PS$_{[3]}$,} **e$_4$** : {Parisien$_{[4]}$} **e$_5$** : {France$_{[6]}$, Elle$_{[7]}$, France$_{[9]}$, France$_{[12]}$,} **e$_{[6]}$** : {Twitter$_{[10]}$}

Figure 2.7 : Tâche de résolution de coréférence.

Mise en œuvre par des méthodes reposant largement sur l'apprentissage supervisé (synthèse de Ng [Ng10]) ou non supervisé ([BR04 ; Ng08 ; HK10]), la résolution de coréférence est confrontée à la difficulté de l'établissement des possibles référents représentés par les mentions qu'il s'agit de regrouper. Au sein d'un document donné, l'ensemble des mentions d'entités (entités nommées ou descriptions définies) constitue un espace de recherche au sein duquel tout sous-ensemble de mentions correspond potentiellement à un référent ; il y a donc autant de référents potentiels que de partitions de cet ensemble, comme le souligne notamment Denis [Den07]. On peut ajouter à cela la relation de dépendance et de succession existant entre la tâche de résolution de coréférence et la REN : il est en effet nécessaire de disposer des mentions d'entités d'un document ou d'un corpus, sous la forme de noms propres, descriptions définies ou pronoms personnels, avant d'en déterminer les liens de coréférence mutuelle et d'en obtenir les regroupements représentant les

référents de façon implicite. La tâche de résolution dépend alors en partie des résultats de la REN, les regroupements référentiels pouvant être bruités ou incomplets en fonction de la qualité des mentions fournies. Cette difficulté supplémentaire ne s'entend cependant que dans le cas où les deux tâches sont envisagées dans le même contexte applicatif ; dans le contexte des campagnes d'évaluation, la résolution est souvent évaluée en tant que telle à partir d'un ensemble de mentions issu des données de référence établies pour la tâche. La relation de succession entre reconnaissance de mentions d'entités et résolution de coréférence souligne néanmoins la distinction existant entre ces deux tâches quant à la prise en charge du caractère référentiel des entités nommées : la première considère les mentions pour elles-mêmes, en tant que segments textuels porteurs de dénotation, quand la seconde intègre pleinement la notion de dénotation en proposant une représentation implicite du référent sous la forme d'un ensemble de mentions.

3.3.2 Discrimination d'entités nommées

De façon similaire à la résolution de coréférence, la tâche de discrimination d'entités nommées intègre l'existence d'un référent sous-jacent au processus de mention. Elle se concentre en revanche plus particulièrement sur le phénomène de l'ambiguïté dénotationnelle, par lequel un même nom ou un même variante lexicale peut référer à plusieurs entités distinctes. Il s'agit alors également d'opérer des regroupements de mentions, cette fois parmi un ensemble d'occurrences de mentions identiques, chaque groupe représentant, comme en résolution de coréférence, un référent distinct. La tâche de discrimination est généralement mise en œuvre selon des méthodes non supervisées de partitionnement ou de *clustering* portant sur les contextes d'occurrence de mentions ([Ped+06 ; PPK05]). Comme en résolution de coréférence, les référents représentés par les regroupements de mentions demeurent implicitement spécifiés. On peut observer que la discrimination d'entités nommées constitue un cas particulier de la discrimination lexicale, où le partitionnement contextuel cherche à distinguer les différents sens possibles d'un même mot ([Sch98 ; PP04]).

3.3.3 Normalisation : variation surfacique et référence

Le terme *normalisation* revêt deux sens principaux dans le contexte de traitements associés aux entités et aux entités nommées. Le premier correspond à un processus de manipulation lexicale par lequel des mentions d'entités, obtenues notamment à l'issue d'une tâche de REN, sont modifiées en des formes considérées comme canoniques. Le second fait intervenir la notion de référents sous-jacents aux mentions d'entités, et ce de façon explicite et externe aux contenus traités, contrairement à la représentation qui en est donnée en résolution de coréférence et en discrimination. Il s'agit alors de faire correspondre une mention ou variante lexicale à un référent, qui doit être défini parmi un ensemble préalablement constitué.

Dans le premier sens de la normalisation, les mentions sont généralement mises en correspondance avec des formes canoniques identifiées parmi l'ensemble de mentions retournées pour un même document. L'opération concerne alors principalement les noms de personnes et d'organisation, dont les variations surfaciques sont en partie prédictibles par les phénomènes d'abréviation et d'acronymie : dans un document présentant les mentions *Hillary R. Clinton* et *Clinton*, la seconde mention peut ainsi être normalisée en *Hillary R. Clinton*, forme plus complète du même nom ; pour les mentions *ONU* et *Organisation des Nations Unies*, la même opération permet de normaliser la première sur la seconde forme. Dans les deux cas, la mention identique à la forme complète est normalisée sur elle-même. La normalisation intervient généralement à l'issue d'une tâche de REN afin d'en présenter des résultats sous une forme homogénéisée. On peut noter que, si une telle opération ne résulte pas en la spécification de référents sous-jacents aux mentions, même normalisées, elle s'appuie de façon implicite sur la supposition que deux mentions normalisées sur la même forme sont coréférentes, du moins si elles apparaissent dans

le même document. Il serait en effet problématique, dans le cas contraire, de normaliser la mention *Clinton* sur *Hillary R. Clinton*, puisque le nom *Clinton* peut constituer la dénotation d'entités pour lesquelles *Hillary R. Clinton* n'est pas une mention possible (*Bill Clinton*). Cette approche de la normalisation intègre donc partiellement et implicitement le phénomène de variation dénotationnelle touchant les entités, qui peuvent être mentionnées par plusieurs formes lexicales distinctes, dont une peut être considérée comme une forme canonique. Dans certains contextes, le regroupement de mentions par forme canonique est par ailleurs considéré comme équivalent à l'établissement d'un groupe co-référentiel, et donc à un référent représenté implicitement. On peut enfin observer que la normalisation de mentions sur une variante commune constitue un élément utile à la résolution de coréférence évoquée précédemment.

La normalisation faisant explicitement état de référents externes tient compte de façon nette de la distinction entre entités nommées et entités, les premières devant être systématiquement reliées aux secondes dans cette tâche. Bien que non limitée au niveau surfacique comme la normalisation dans le premier sens du terme, la problématique abordée ici part de la variation dénotationnelle, considérée comme un obstacle à l'accès direct aux entités mentionnées, qu'il s'agit donc de lever. Cette normalisation peut être vue comme immédiatement dérivée de la première, en supposant qu'une entité correspond de façon univoque à un nom canonique — il n'y aurait, par exemple, qu'une seule entité dont le nom canonique est *Hillary Rodham Clinton*, et toutes les variations surfaciques normalisées sur ce nom correspondraient à des dénotations de la même entité. Mais les travaux en normalisation d'entités nommées, tels que ceux de Li et al. [Li+02], Khalid et al. [KJDR08] ou Jijkoun et al. [Jij+08], tiennent généralement compte des phénomènes de synonymie et d'ambiguïté à l'œuvre dans la dénotation. La correspondance entre variante lexicale et référent est alors établie par des moyens étendus, notamment le recensement des variantes possibles pour une entité donnée à partir de ressources encyclopédiques, Wikipedia [9] figurant parmi les plus utilisées.

On peut observer à l'égard de la normalisation entendue dans son second sens que la correspondance établie entre mentions et référents, définis préalablement et séparément des contenus, se ramène à une identification d'entités à partir de leurs occurrences au sein de contenus textuels. Il s'agit précisément de la tâche à accomplir dans le cadre du présent travail par la mise en œuvre de l'Annotation Sémantique ; celle-ci implique cependant une spécification formelle des références, notamment par le biais de ressources ontologiques. Cette caractéristique de l'Annotation Sémantique, évoquée au chapitre 1 et abordée plus spécifiquement dans le chapitre suivant, reste hors du champ de la normalisation, ainsi que d'autres travaux dans lesquelles le niveau surfacique et textuel des mentions est également mis en relation avec une représentation explicite de référents.

On peut constater, dans certains de ces travaux (par exemple [Liu+12]), que la distinction entre les deux sens de la normalisation n'est pas toujours nette, notamment lorsqu'un ensemble de référents disponibles pour la mise en correspondance des mentions d'entités n'est pas spécifié, et que le processus de regroupement de mentions en fonction d'une forme canonique est considéré comme équivalent à l'établissement d'un référent. Une définition générale, unifiée et claire de la tâche associée à ces différents travaux est proposée dans le cadre de la campagne d'évaluation TAC-KBP (*Text Analysis Conference*), tâche de Population de Bases de Connaissances ou *Knowledge Base Population*), qui fait également l'objet d'une étude dans le chapitre suivant.

3.3.4 Résolution d'entités

La résolution d'entités est concernée par leur représentation hors des contenus textuels, sous la forme prototypique de bases de données. De telles bases impliquent que chaque entité ainsi

9. `http://www.wikipedia.org`

représentée constitue une entrée unique. Leur constitution est ainsi confrontée au problème des doublons, lorsque plusieurs entrées correspondent à une même entité. La tâche de résolution consiste ainsi à identifier ces doublons afin de maintenir la cohérence des données et des informations agrégées à leur endroit. Plus généralement, étant donné une ressource sous la forme d'une base de données ou plus généralement structurée, de nouvelles entrées peuvent être présentées afin d'y être intégrées. Il s'agit alors de déterminer si l'entrée candidate correspond à une entité déjà recensée par la base ou si elle doit faire l'objet d'une nouvelle entrée. Le même processus intervient lors de la fusion de plusieurs ressources : pour chaque entrée de chaque ressource, la résolution détermine si les autres ressources disposent de l'entrée équivalente ou non, dans l'objectif d'aboutir à une ressource unifiée et sans doublons.

La résolution d'entités, aussi appelée *record linkage* et distincte de la tâche de *linking* dont il sera question dans la suite de ce travail, s'appuie sur des méthodes empiriques consistant à déterminer les critères de comparaison pertinents étant donné une ressource et un ensemble d'entrée candidates, ou un ensemble de ressources à fusionner. Ces critères dépendent du schéma de structuration des ressources concernées. Les entités représentées de façon structurée sont généralement associées à un attribut correspondant à leur nom, qui peut être comparé au nom d'entités candidates ; la décision de fusion ou de création d'une entrée distincte s'appuie alors souvent sur le calcul de distances d'édition. Tous les attributs des entités d'une base (âge, coordonnées, salaire, nationalité...) peuvent ainsi être comparés un à un avec ceux des entrées candidates, selon la méthode la plus adéquate étant donné leur nature. Une synthèse des méthodes de résolution est notamment proposée par Brizan et Tansel [BT06]. Bien que définie hors du paradigme de l'EI, la résolution d'entités peut faire intervenir une analyse de contenus textuels dans le processus de comparaison, notamment par le calcul de similarités contextuelles entre entités. La figure 2.8 illustre par des exemples de ressources et d'entrées candidates la problématique et la méthode générale de la résolution d'entités.

On peut enfin noter que le terme *résolution d'entités* est parfois employé dans le cadre d'autres types de travaux, également en relation avec le problème de l'identité, entendue dans ses deux acceptions — déterminer de quelle référent ou entité il s'agit, notamment à partir de mentions textuelles, ou reconnaître deux entrées ou entités comme identiques. Le processus de mise en correspondance entre mentions d'entités et référent externe, qualifié de normalisation par certains (cf. *supra*, 3.3.3), prend le nom de résolution par exemple chez Pilz et Paaß [PP09].

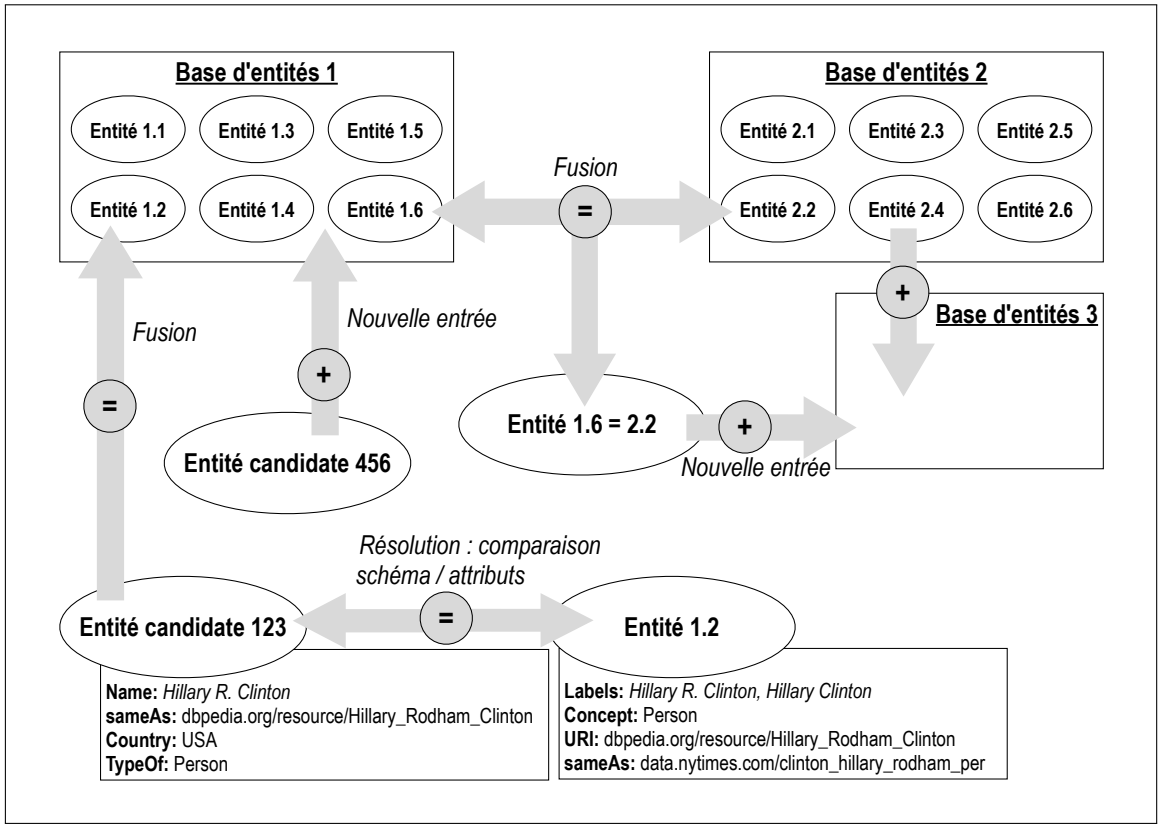


Figure 2.8 : Résolution d'entités.

Chapitre 3

Annotation Sémantique et identification d'entités

L'acquisition de métadonnées à partir des entités mentionnées dans les contenus textuels nécessite l'adoption d'une méthodologie dépassant le cadre de l'Extraction d'Information et permettant d'accéder pleinement à la sémantique référentielle de ces entités. Le paradigme du Web Sémantique rejoint de nouveau cette recherche méthodologique dans la mesure où c'est également dans ce cadre qu'est définie l'Annotation Sémantique. Ce processus rassemble dans sa formalisation et son résultat les aspects liés à l'enrichissement — l'adjonction d'information aux contenus — ainsi qu'à la sémantique devant caractériser cet enrichissement — les annotations produites se présentant comme métadonnées potentielles, telles que le chapitre 1 (section 2) a pu en évoquer la forme et l'usage. L'Annotation Sémantique se présente à cet égard comme le complément productif et formel à la suite de l'Extraction d'Information, dont elle permet de transposer la structuration vers une sémantique formalisée en termes d'identification, notamment en ce qui concerne les entités. C'est en effet à la condition d'un traitement des entités en tant qu'individus référencés et objets de descriptions formelles que des métadonnées proprement dites peuvent être envisagées.

En tant que moyen de réalisation de l'enrichissement de contenus textuels à l'aide de métadonnées, l'Annotation Sémantique doit donc donner lieu à un examen à la fois définitoire et méthodologique. La formalisation de cette tâche s'inscrit dans le cadre et les objectifs du Web Sémantique, qu'il s'agit de mettre en relation avec la notion de métadonnées (section 1.1), avant de centrer notre étude sur le cas des entités et de la sémantique qui doit leur être associée (section 1.2). Une description des ressources disponibles ou nécessaires, émergeant également des activités liées au Web Sémantique, se joint à la présentation générale de la tâche d'Annotation Sémantique (section 1.3). En pratique, l'Annotation Sémantique fait l'objet de plusieurs systèmes et cadres de développement, reposant sur des approches méthodologiques variées. C'est notamment à travers l'observation de ces approches que la place de l'Extraction d'Information sera discutée (section 2).

Si le fonctionnement méthodologique général de l'Annotation Sémantique convient à l'objectif d'un enrichissement à l'aide de métadonnées, le processus d'identification d'entités demande en revanche une spécification précise, notamment au regard du traitement du phénomène dénotationnel qui en constitue le cœur. Le Liage d'entités, défini dans la tâche de Population de Bases de Connaissances de la conférence TAC, se présente comme une réponse adéquate pour ce processus et sera détaillé à la section 3.

1 La tâche d'Annotation Sémantique pour l'enrichissement de contenus textuels

1.1 Du cadre du Web Sémantique à l'acquisition de métadonnées

1.1.1 Enrichissement de contenus et Web Sémantique

L'enrichissement de contenus textuels à l'aide de métadonnées a été évoqué à plusieurs reprises dans les précédents chapitres comme un objectif de publication documentaire venant dépasser le seul niveau surfacique des textes et en faire émerger les connaissances de façon formelle et explicite. Il s'inscrit ainsi dans une perspective de traitement de l'information sophistiqué, principalement défini par les notions de partage, d'intégration et d'interprétabilité, notamment automatique, des connaissances véhiculées. À titre d'illustration, le fragment de document suivant :

(6) Vendredi, François Hollande a tenu à rappeler que sa décision de retirer les troupes françaises combattantes d'Afghanistan n'était "pas négociable".[1]

mentionne, dans les segments soulignés, une personne et un pays, deux éléments pouvant être considérés comme centraux et pertinents au niveau du contenu informatif. Une version enrichie de ce fragment :

(7) Vendredi, `<metadata type="Person" doc_id="123987">`François Hollande`</metadata>` a tenu à rappeler que sa décision de retirer les troupes françaises combattantes d'`<metadata type="Country" doc_id="456321">`Afghanistan`</metadata>` n'était "pas négociable".

promeut les mentions considérées au statut de métadonnées. Non seulement clairement délimitées et ainsi mises en valeur dans la consultation du document, ces mentions fournissent une information explicite sur son contenu. Dans cet exemple, cette information relève d'une association, par l'attribut `doc_id` de la balise `metadata`, entre les mentions considérées et une ressource externe, pouvant par exemple être mise à disposition par un service de documentation. L'espace de connaissances couvert par le document est ainsi étendu par une telle association, un programme pouvant automatiser l'accès à cette ressource à partir des métadonnées ajoutées au document. La manipulation de cet espace informatif peut alors donner lieu à plusieurs types de traitements qualifiés de *sémantiques*, dont les documents et métadonnées constituent les éléments de base : à des fins de recherche d'information ou de classification automatique, les métadonnées fonctionnent comme descripteurs formels des documents, permettant une indexation structurée ainsi qu'une sélection documentaire guidée par des éléments informatifs explicitement caractérisés. Ainsi, un utilisateur peut bénéficier d'un enrichissement de documents par la possibilité d'exprimer une requête de recherche à partir du pays *Afghanistan* et d'accéder en retour à l'ensemble des documents disponibles en faisant mention. Une spécification plus détaillée des usages centrés autour de métadonnées munies d'une telle sémantique sera proposée dans le chapitre 4, en relation avec les besoins industriels connus par la presse en termes de traitement de l'information, et par l'AFP en particulier.

L'enrichissement à l'aide de métadonnées se présente ainsi comme une application concrète, visant les usages réels correspondant aux idées de fonctionnement général du Web Sémantique. Il en constitue une formulation tangible, émanant d'acteurs concernés au plus près par les problématiques de renouvellement de traitement de l'information qui s'y trouvent exprimées. L'Annotation Sémantique joue à ce titre le rôle de producteur des métadonnées d'enrichissement, à travers une méthodologie construite par les différentes communautés de développement du Web Sémantique.

1. Exemple extrait du site Web `http://www.lemonde.fr` daté du 10 mai 2012

Elle résulte ainsi de travaux moins déterminés par des problématiques de recherche que par des visées d'ingénierie applicatives à l'image du Web Sémantique lui-même, comme l'illustrent des contributions telles que celle de Domingue et al. [DFH11].

1.1.2 Annotation Sémantique : définition

L'Annotation Sémantique (AS) a été abordée dans le premier chapitre du présent travail (section 1.3) en tant que nécessité au niveau opérationnel, découlant des attendus exprimés par le Web Sémantique en termes d'intégration et d'interprétabilité de l'information contenue dans les documents du Web. Elle constitue en effet le processus par lequel un niveau de représentation formelle de connaissances peut émerger du niveau textuel [2]. Si le Web Sémantique propose une représentation du monde — réel et global ou réduit et scindé en domaines — reposant sur des modèles de base formelle logique, l'AS en est le pendant opérationnel, qui effectue l'association entre les connaissances distribuées sous forme linguistique et les modèles choisis. Ainsi, à partir d'un ensemble documentaire et d'une ontologie modélisant le domaine considéré, l'augmentation des documents par AS permet une représentation des connaissances exprimées au niveau textuel par le truchement de liens établis avec le modèle. La figure 3.1 illustre une telle configuration, à partir d'articles journalistiques sur le thème de la vie politique européenne et d'une ontologie associée.

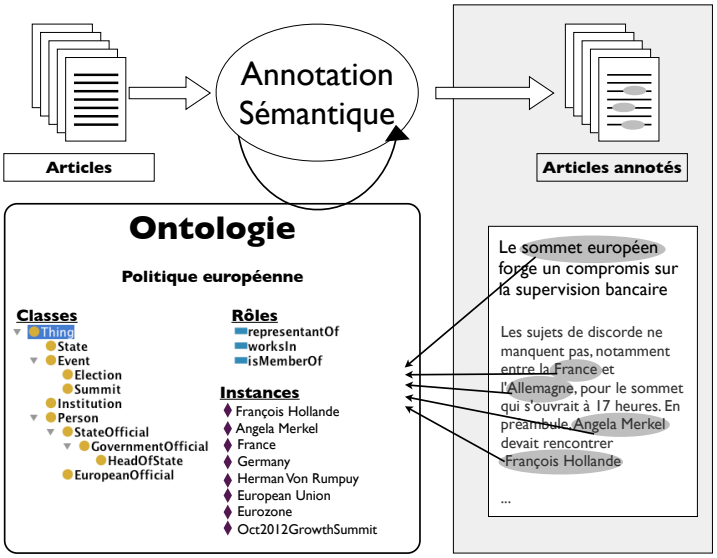


Figure 3.1 : Exemple de schéma général d'Annotation Sémantique.

L'AS se définit comme :

> [3.1] La mise en relation explicite d'un support informatif, sous la forme d'un segment textuel de document, et d'un modèle de connaissances à l'aide de marqueurs indiquant les éléments informatifs sélectionnés pour l'établissement de ce lien et encodant formellement leur relation référentielle au modèle.

2. l'AS peut s'envisager au niveau de données non textuelles : parole ou autre données sonores, image, video, infographie sont d'autres formes d'information présentes sur le Web pouvant donner lieu à des traitements d'ordre sémantique. Nous nous limitons cependant dans le présent travail aux données textuelles et au problème particulier de la relation entre niveau linguistique et représentation.

Elle repose donc sur les éléments constitutifs suivants :

Documents L'AS est une opération centrée autour des documents : ceux-ci constituent le support de l'information et des connaissances recherchées ; les annotations se définissent en relation avec le document.

Modèle Le caractère sémantique des annotations repose sur leur association avec un modèle ontologique, qui doit être spécifié au préalable, accessible et identifiable par le mécanisme des URI. L'adoption du formalisme ontologique répond aux impératifs de spécification conceptuelle et de partage de cette spécification énoncés au chapitre 1 (section 1.2.2), ainsi qu'aux insuffisances formelles des modèles non liés usuellement employés en Extraction d'Information.

Langage Un langage de balisage permet l'insertion des marqueurs d'annotation au niveau textuel, ainsi que l'encodage formel des associations entre annotations et modèle. La syntaxe du langage XML est généralement adoptée, couplée à l'utilisation d'URI pour les références au modèle.

Au-delà du processus d'AS, tout agent humain ou automatique peut accéder aux connaissances véhiculées par le biais des annotations ainsi produites et des URI renseignées. L'interprétabilité rendue possible par la modélisation permet par suite d'associer à ces annotations des actions programmatiques tenant compte de la sémantique définie par l'ontologie sous-jacente.

La relation de référence entre segment textuel et modèle peut viser tout élément constitutif de l'ontologie :

Concept Annotation d'un segment du document, typiquement un terme ou une expression nominale, comme mention d'un concept de l'ontologie ; dans l'exemple 7, le nom commun *troupes* peut être associé au concept correspondant s'il est défini dans l'ontologie.

Rôle Annotation d'un segment du document, typiquement verbal, comme mention d'un rôle [3] ou d'une instance derôle ; dans l'exemple 7, le verbe *retirer* peut être associé au rôle correspondant s'il est défini dans l'ontologie, tandis qu'une analyse plus complète et profonde de la phrase elle-même peut identifier une instance de ce rôle, avec pour domaine *François Hollande* et pour portée *les troupes françaises*.

Instance Annotation d'un segment du document, typiquement nominal, comme expression référant à une instance de l'ontologie ; dans l'exemple 7, les noms propres *François Hollande* et *Afghanistan* peuvent ainsi être associés aux instances de concepts représentant les personnes et pays, respectivement.

Dans chacune de ces situations d'annotation, il est utile de souligner que seuls des éléments définis au préalable dans l'ontologie peuvent constituer une cible pour le marqueur correspondant. Cette antécédence du modèle sur les contenus en termes d'éléments qu'il est possible de référencer distingue l'AS de la tâche d'acquisition ou de population d'ontologie, notamment, qui mettent également en jeu une relation entre modèle et contenus textuels informatifs. Dans ces tâches, la structure conceptuelle ainsi que leurs membres instanciés peuvent être dérivés à partir des contenus : c'est donc la mention des concepts, relations et instances dans un ensemble documentaire qui préside à leur introduction en tant qu'élément du modèle. Dans le cadre de l'AS, les éléments informatifs issus des contenus et non représentés au sein du modèle sont donc démunis

3. Le terme *rôle* est employé dans la description formelle d'ontologies et correspond à la notion de relation conceptuelle ; le terme *relation* peut ainsi lui être substitué, ainsi que le terme *propriété*, plus proche de la terminologie propre au langage OWL.

de la relation de référence nécessaire à l'établissement d'une annotation. Le possible complément ainsi exclu de l'ensemble d'annotations peut donner lieu à des traitements particuliers, visant notamment à l'enrichissement du modèle selon un processus cyclique : les outils déployés aux fins d'AS peuvent ainsi, par exemple, constituer au fil de l'annotation un ensemble de candidats non définis dans le modèle, puis le proposer en retour à un module de gestion de l'ontologie ; celle-ci peut, dans le cycle suivant, rendre disponibles les références correspondantes lors d'une nouvelle annotation. Une telle fonctionnalité implique, au niveau des outils d'AS, une capacité à repérer les éléments informatifs absents du modèle. Ceci ne constitue pas une opération triviale, comme l'illustrera l'examen des méthodologies effectives d'AS dans la suite de ce travail.

1.1.3 Constitution et acquisition des métadonnées

Contrairement à la situation usuelle en Extraction d'Information, le déploiement de l'AS n'entretient pas de lien de dépendance par rapport à une tâche particulière qui serait envisagée en aval sur les contenus traités. Elle est au contraire conçue pour fonctionner de façon générale dans l'espace du Web Sémantique, indépendamment des applications exploitant les informations représentées sous forme d'annotations. Elle repose ainsi sur le principe de non spécialisation en fonction d'usages, qui ne peuvent faire l'objet d'une prédiction étant donné le caractère ouvert, distribué et diversifié du Web. L'AS peut donc être envisagée comme une tâche autonome, organisée autour de ses éléments constitutifs — documents, modèle, langage d'annotation. Cette autonomie dérive de l'aspect standardisé de l'AS, qui caractérise de façon générale le Web Sémantique ; c'est en effet par la standardisation que sont envisageables des traitements non définis d'avance, dès lors que ces derniers sont également intégrés aux standards du Web Sémantique. C'est ainsi au double titre de composant à part entière du Web Sémantique et de tâche autonome que l'AS se présente comme la méthode de mise en œuvre de l'enrichissement de contenus textuels à l'aide de métadonnées.

L'enrichissement de contenus peut en effet être formulé par des producteurs et éditeurs de contenus — agence de presse comme l'AFP, auteurs de blogs ou gestionnaire de site Web d'entreprise — en tant que besoin générique, visant à une intégration dans le paradigme du Web Sémantique et ne préjugeant pas nécessairement de ses usages possibles. Ceux-ci relèvent en effet des producteurs eux-mêmes, pour qui des contenus enrichis permettent d'envisager une gamme d'applications, elles aussi encadrées par les pratiques du Web Sémantique, mais également des utilisateurs extérieurs au processus de production — clients d'une agence de presse, abonnés d'un blog ou public d'un site Web — et dont les intentions quant à l'usage de ces contenus n'est pas contraint d'avance. La généricité du besoin d'enrichissement concerne donc son caractère formel, défini par l'emploi de l'AS comme méthode primordiale de mise en œuvre.

Une certaine spécialisation intervient en revanche au niveau des contenus visés par l'enrichissement, en tant qu'ils sont généralement constitués autour d'un domaine — champ thématique ou centre d'intérêt d'ordre communautaire. L'information qu'il s'agit ainsi de représenter formellement est donc quant à elle concernée par les notions de pertinence et de sélection, en relation avec le domaine considéré : à partir d'un ensemble documentaire, la question se pose de savoir quels types d'éléments informatifs sont visés par l'enrichissement. Ainsi, un producteur de contenus spécialisés sur la vie politique européenne peut proposer ces contenus au public sous une forme enrichie par un processus d'AS, sans contraindre ni figer *a priori* leurs usages ultérieurs. Les cibles de l'enrichissement sont en revanche délimitées en relation avec le domaine traité dans ces contenus : il pourra s'agir, comme l'illustre la figure d'exemple 3.1, d'institutions, de pays ou de personnalités liées à ce domaine, ainsi que des relations qu'ils entretiennent. Les contours de cette relation ne sont pas déterminés *a priori* par le domaine lui-même et relèvent de modalités de décisions propre à la définition d'un modèle, pouvant varier d'un point de vue et d'une situation à l'autre.

La relation entre domaine et cibles de l'enrichissement justifie l'intégration de l'AS au processus rédactionnel, déjà évoquée précédemment (chapitre 1, section 1.3.1. Les métadonnées constituant cet enrichissement sont en effet issues d'une sélection d'éléments informatifs considérés comme pertinents en regard du contenu traité. Celui-ci peut donc être vu comme enrichi de façon concomitante à sa production, même si les deux opérations ne prennent pas place simultanément au niveau temporel. La validité de la sélection, c'est-à-dire la capacité pour un élément informatif à donner lieu à une métadonnée de document, est quant à elle garantie par la possibilité d'une relation de référence au modèle défini.

La fonction méthodologique de l'AS pour l'enrichissement de contenus peut donc être systématisée de la façon suivante :

- L'objectif d'enrichissement est formulé à l'égard d'un ensemble documentaire portant sur un domaine. Il implique une sélection d'éléments informatifs au travers de ces contenus, selon un critère de pertinence associé au processus rédactionnel.

- Le domaine considéré ainsi que les éléments informatifs pertinents donnent lieu à une modélisation conceptuelle, sous la forme d'une ontologie. Celle-ci spécifie les concepts relatifs au domaine selon la vue choisie, ainsi que les éventuelles relations entrant dans cette modélisation. L'ontologie peut également être peuplée, c'est-à-dire comporter un ensemble d'instances des concepts spécifiés.

- Les éléments informatifs rencontrant les critères de pertinence adéquats donnent lieu à des annotations, sous forme de marqueurs les délimitant au sein des contenus et indiquant leur lien référentiel avec l'ontologie sous-jacente, lorsque celui-ci existe. Les trois types d'éléments ontologiques — concepts, relations et instances — constituent les cibles potentielles de ce lien.

- Les métadonnées de document correspondent à l'ensemble de ces annotations. Chacune d'elles est constituée d'une référence vers un élément de l'ontologie employée, permettant aux traitements utilisateurs d'accéder à sa description, ainsi que d'une association avec le document sous la forme de sa localisation dans le texte.

Dans l'exemple figuré par le schéma 3.1, l'enrichissement des documents traitant de politique européenne peut se présenter comme l'illustre la figure suivante (3.2), où un extrait de contenu textuel annoté est reproduit en regard du modèle ontologique. L'URI de l'ontologie utilisée, `www.semanticweb.org/ontologies/2012/euroPol`, est abrégée en `euroPol` dans les balises d'annotation nommées `metadata`.

1.2 Sémantique des entités comme métadonnées

Dans la perspective d'une sélection des métadonnées considérées comme pertinentes relativement à un domaine et un modèle donnés, un producteur de contenus tel que l'AFP s'intéresse en premier lieu aux entités — personnalités, lieux, organisations — mentionnées dans les fils d'actualité. Cet ensemble peut être étendu à d'autres éléments, tel que les événements dans lesquelles ces entités interviennent ou les relations qu'elles entretiennent, mais est visé en priorité par le processus d'enrichissement [4]. Cette attention particulière accordée aux entités correspond à la place essentielle qu'elles occupent dans l'espace informatif, notamment s'agissant d'actualité et de contenus journalistiques.

4. Le modèle employé pour l'enrichissement des contenus AFP est présenté au chapitre 7 (section 1.1).

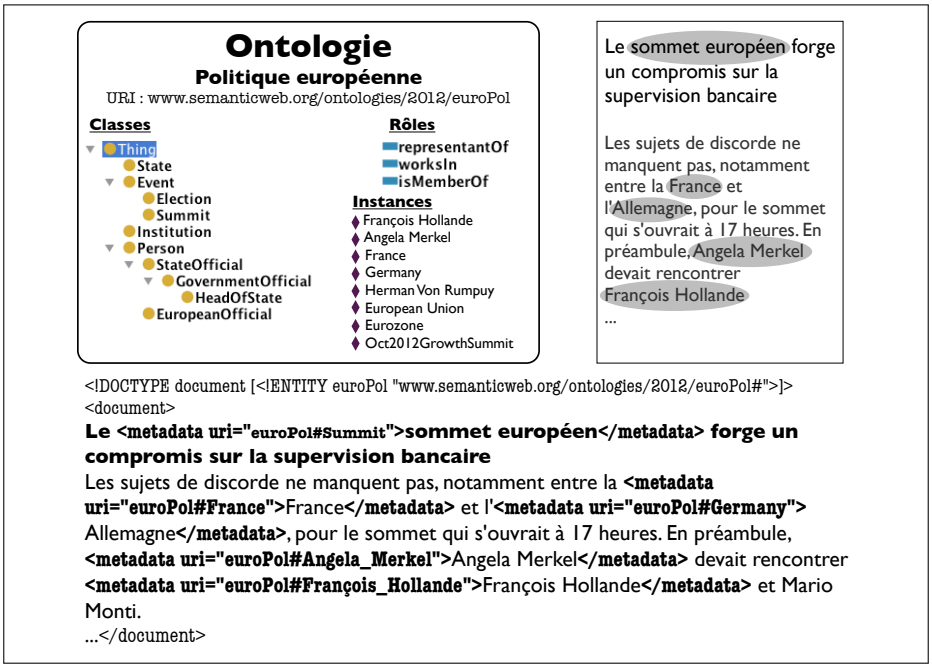


Figure 3.2 : Enrichissement de documents : Politique européenne.

L'étude de la Reconnaissance d'Entités Nommées (REN) proposée au chapitre 2 (section 3.1) évoque le rôle central des entités dans les tâches s'intéressant aux connaissances véhiculées par les contenus textuels et délimite la sémantique attribuée aux entités par la REN sur le mode de la classification. Un enrichissement de contenus textuels pourrait être envisagé sur la base typologique proposée par la plupart des systèmes de REN : les métadonnées de documents ainsi traités porteraient alors les informations de type — personne ou lieu — définies dans un modèle et associées à des segments correspondant à des dénotations d'entités. Un tel enrichissement peut s'appuyer sur un système d'Extraction d'Information dont le modèle sous-jacent est de nature ontologique, comme l'a présenté la section 2.3 du chapitre 2. Il s'agirait alors d'une forme d'AS dans laquelle les mentions d'entités sont associées à des classes ontologiques, comme le sont les mentions de concepts.

Le paradigme général du Web Sémantique et l'organisation des connaissances qu'il propose, auxquels l'enrichissement de contenus à l'aide de métadonnées sémantiques cherche à se rattacher, conduisent cependant à étendre la notion d'ancrage sémantique des entités au-delà de la classification typologique. En tant qu'objets caractérisés par la notion d'*individus*, les entités sont en effet modélisées dans l'espace du Web Sémantique comme instances ontologiques, identifiables de façon unique. Cette modélisation correspond au caractère représentationnel de ces instances, qui tiennent lieu d'approximation pour ces objets existant par ailleurs. Les instances ontologiques représentant des entités en fournissent donc une identité. Chacune d'elle constitue ce que le Web nomme une *ressource*, qui peut être accompagnée d'une description. L'appartenance d'une instance à une classe conceptuelle ontologique en fournit une description minimale, et l'ontologie à laquelle elle se rattache peut augmenter cette description, notamment par l'instanciation de relations et l'assignation d'attributs. Parallèlement, le rôle attendu des entités en tant que métadonnées sémantiques dans le cadre de l'enrichissement est également fortement attaché à la notion d'entité : il s'agit concrètement d'expliciter, au fil des contenus, de qui ou de quoi il s'agit, de façon à permettre, d'une part, un accès à la ressource correspondante dans un but de documentation, et, d'autre part, une mise en relation systématique et immédiate de ces

contenus avec d'autres, sur la base des entités qui y sont mentionnées. Si l'on reprend l'idée de métadonnées dont la sémantique est celle du type des entités mentionnées, de telles ouvertures de l'espace informatif des documents traités seraient limitées à ces types : il serait alors possible d'accéder à une définition du type en question et non de l'entité elle-même, et la mise en relation de contenus retournerait un ensemble de documents mentionnant des entités du même type, par exemple PERSONNE.

L'enrichissement de contenus textuels visant les entités comme métadonnées nécessite donc que celles-ci prennent en charge la modélisation des entités telle que la définit le Web Sémantique et les standards correspondant. Les métadonnées considérées comme valables dans ce cadre doivent ainsi comporter une référence à une instance ontologique représentant une entité, éventuellement munie d'une description ; cette référence prend la forme d'une URI permettant de localiser et surtout d'identifier de façon univoque l'entité dont il s'agit. La figure 3.2 présente des exemples de telles métadonnées : les marqueurs

```
<metadata uri="euroPol#François_Hollande">François Hollande</metadata>
```

et

```
<metadata uri="euroPol#Angela_Merkel">Angela Merkel</metadata>
```

renvoient à deux instances de l'ontologie identifiée par l'URI

```
www.semanticweb.org/ontologies/euroPol
```

abrégée en `euroPol#`. Le mécanisme de référencement des URI associé à la syntaxe OWL permet par suite d'accéder aux connaissances concernant ces instances, notamment leur classe conceptuelle d'appartenance et les relations qui les lient à d'autres instances de l'ontologie. Dans l'exemple correspondant à la figure 3.2, de telles connaissances pourront se présenter sous la forme du graphe 3.3, à partir de l'instance *François Hollande* et des classes, relations et autres instances définies dans l'ontologie `euroPol#`.

La sémantique ainsi attribuée aux entités est donc liée aux notions de référence et d'identité ; elle peut en ce sens être qualifiée de *référentielle*, comme cela a été proposé précédemment (chapitre 2, section 3.3). L'interprétation qu'elle permet relève en effet de la relation de référence établie entre un segment textuel et une entité représentée dans un modèle. Sur le plan opérationnel, l'établissement d'une telle relation constitue un processus d'*identification*. En termes d'AS, cette identification restreint les segments textuels à annoter aux dénotations d'entités d'une part, et les cibles d'annotation aux instances ontologiques membres de classes conceptuelles représentant des entités d'autre part.

Sur ce dernier point, on peut observer que de telles classes posent la question de la modélisation des entités, c'est-à-dire de leur possible classification conceptuelle. La REN s'intéresse particulièrement aux différentes modalités d'organisation des entités en classes, en relation avec des tâches spécifiques mais également dans un souci plus général de modélisation du monde au travers des entités (cf. chapitre 2, section 3.2). Plusieurs typologies d'entités ont à ce titre été proposées dans le cadre de campagnes et de système de REN et il est aisé de les considérer comme base conceptuelle valide pour une représentation ontologique dans le cadre de l'AS. Les classes communes à la plupart de ces typologies correspondent en effet aux types PERSONNE, LIEU, avec des distinctions possibles entre lieux proprement géographiques et entités géopolitiques, et ORGANISATION, ce dernier comprenant généralement le type ENTREPRISE. Ces types recouvrent ainsi un ensemble pertinent pour le traitement de contenus, notamment journalistiques, et dont le

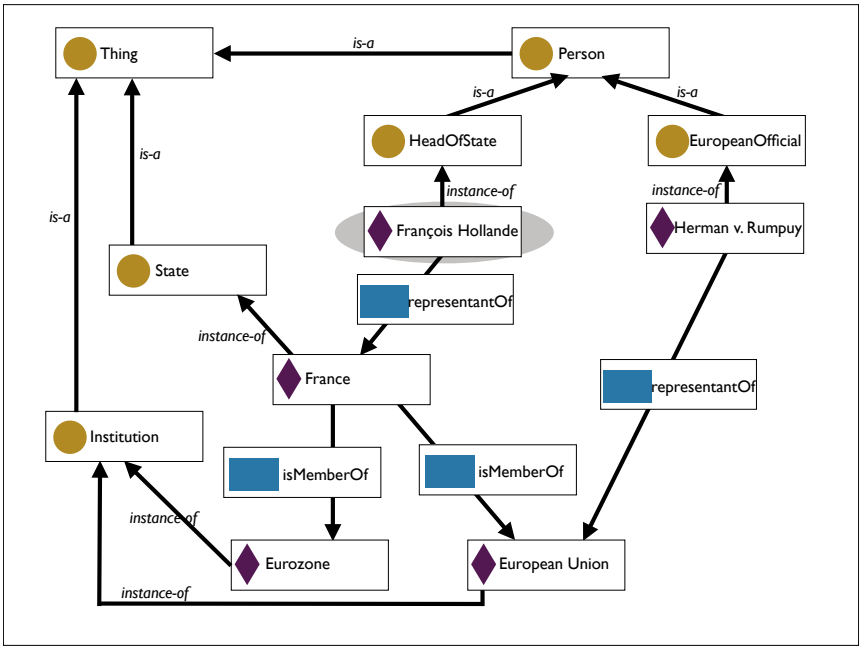


Figure 3.3 : Exemple de graphe de connaissances accessibles à partir d'une instance ontologique.

degré de généralité est de nature à s'adapter à de nombreux contextes et tâches. La question de la modélisation et de sa place dans l'AS pourra être discutée plus en avant avec l'examen des ressources nécessaires à sa mise en œuvre. Celles-ci pourront également illustrer les modalités de représentation de l'identité des entités visées par l'enrichissement à l'aide de métadonnées.

1.3 Ressources pour l'Annotation Sémantique

La définition de la tâche d'AS comme une mise en relation de segments textuels au sein de documents et d'un modèle implique la mise à disposition de ressources adéquates. Il s'agit en premier lieu du modèle en question, envisagé sous la forme d'une ontologie. Les éléments constitutifs de cette ontologie forment l'ensemble des cibles possibles de l'AS, pour lesquelles sont fournies les descriptions formelles nécessaires. Ces éléments relèvent de l'un des types d'objets ontologiques — concept, rôle ou instance — et donnent lieu aux types d'annotations correspondants, tels que décrits précédemment (section 1.1.2). Il est utile d'observer que l'ontologie adoptée comme ressource doit être *peuplée* afin de permettre des annotations ciblant non seulement des concepts et relations abstraites, mais également des instances de concepts et de relations. Dans le cadre d'un enrichissement de contenus visant principalement les entités, comme cela a été discuté dans la section précédente, la disponibilité d'instances ontologiques dont les classes conceptuelles définissent de telles entités est incontournable. L'ensemble des instances définies dans une ontologie, ainsi que des relations qui y sont instanciées est désigné par le terme de *population*.

Les ressources d'AS ainsi définies peuvent être regroupées en deux types généraux. Le premier ensemble considéré est celui ressources développées dans le cadre du Web Sémantique et plus particulièrement des Linked Data (LD), dont une description générale a été proposée au chapitre 1 (section 1.2.3). Les LD se caractérisent principalement par une mise à disposition publique et un accès systématisé [5], par le biais de l'architecture ouverte et distribuée du Web ainsi que du mécanisme des URI. Ces données sont typiquement structurées selon un schéma ontologique

5. `http://datahub.io/`

définissant au moins un ensemble de concepts et le plus souvent un ensemble d'instances membres de ces classes conceptuelles. Les LD comptent un nombre croissant d'ensembles de données, produites par divers agents et communautés inscrivant leur activité dans le cadre du Web Sémantique. Ces ensembles connaissent une organisation en réseau, reposant sur la définition mutuelle de liens de synonymie entre les ensembles de données concernés ; chacun d'entre eux constitue un nœud du graphe obtenu. Ces nœuds correspondent ainsi à la modélisation de différents domaines, activités et centres d'intérêt, et plus particulièrement à leur représentation sous forme de populations d'instances ontologiques dont le nombre et la diversité conceptuelle reflète le degré de couverture et la vue adoptée sur chaque domaine considéré.

Parallèlement à ces ressources préexistantes et publiquement disponibles, une tâche d'AS particulière peut envisager la constitution *ad hoc* ou privée des modèles et populations nécessaires. L'agent à l'origine d'une AS sur un ensemble documentaire peut en effet privilégier une forte adéquation des ressources employées au domaine traité, en raison de l'absence de telles données parmi les nœuds existant dans les LD ou d'un impératif de non publicité des données. Dans le premier cas, il est important de noter qu'une telle constitution de données est souvent envisagée comme une première étape vers un ancrage des ressources ainsi créées dans le réseau des LD. Ces ressources peuvent par ailleurs s'appuyer sur les LD en important tout ou partie de leur population des ensembles de données existants, en définissant le plus souvent les équivalences conceptuelles nécessaires à la mise en correspondance d'un modèle vers l'autre.

À ces deux types généraux s'ajoutent certaines ressources hybrides, partageant le caractère libre des nœuds publics des LD sans nécessairement faire l'objet d'un référencement formel sur ce réseau. Ces ressources s'apparentent également au second type en tant qu'elles peuvent faire l'objet d'un développement en regard d'une tâche et d'un contexte d'application particuliers. Ce développement peut intégrer à divers degrés un souci de généricité et de compatibilité avec les LD, dont elles pourront ainsi constituer de nouveaux nœuds à l'occasion de développements ultérieurs.

Afin d'illustrer les requis formels et pratiques touchant les ressources d'AS, un certain nombre d'ensembles de données correspondant aux catégories évoquées peuvent faire l'objet d'un examen particulier. Il s'agit pour chacun d'entre eux d'identifier le modèle ontologique adopté ainsi que le mode de description formelle associé aux instances de ce modèle. Le choix de ces exemples reflète les différents aspects pouvant mener à leur adoption pour une tâche d'AS : généricité ou spécialité du domaine, caractère public ou privé, couverture et formalisation des descriptions d'entités. Ce dernier critère concerne plus particulièrement l'enrichissement de contenus à l'aide de métadonnées ciblant principalement les entités.

DBpedia

DBpedia [6] est issu d'un effort communautaire dédié à l'extraction et à la structuration des informations contenues dans l'encyclopédie en ligne Wikipedia [7] et est disponible sous licence GPL [8]. Bien que Wikipedia existe en tant que nœud des LD, auquel il est possible de faire référence par le mécanisme des URI, DBpedia répond à un besoin de systématisation d'accès et de représentation, notamment par l'adoption des standards du Web Sémantique tels que RDF. Le processus de conversion à l'œuvre entre le corpus formé par Wikipedia et l'ensemble de données résultant dans DBpedia est présenté en détails dans [Biz+09] ; après une présentation schématique de l'encyclopédie en ligne Wikipedia, nous donnons ici une vue synthétique de ce processus et de son résultat.

6. `http://dbpedia.org/About`
7. `http://www.wikipedia.org`
8. `http://www.gnu.org/licenses/gpl.html`

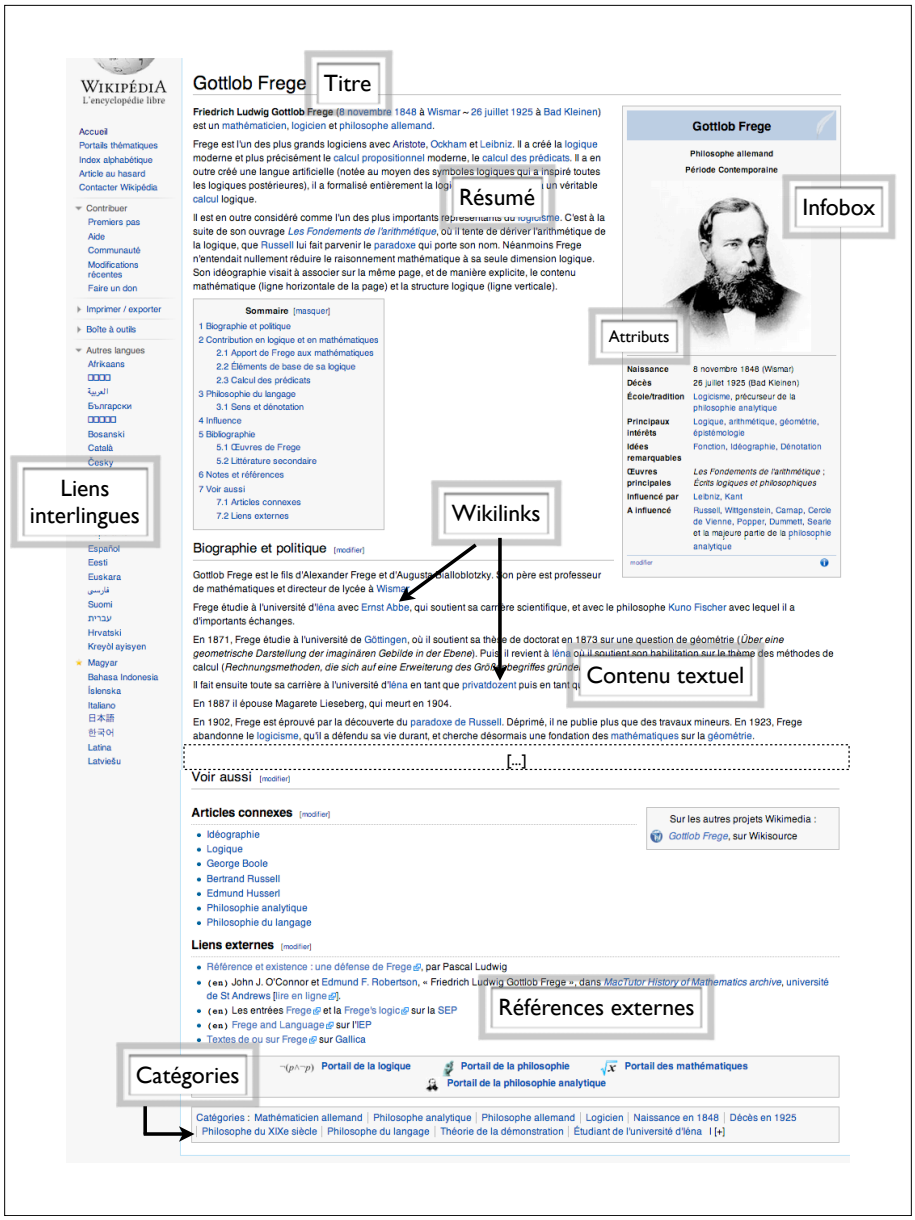


Figure 3.4 : Schéma d'un article Wikipedia muni d'une infobox.

Wikipedia La modélisation ainsi que la structuration des données de DBpedia s'appuient sur le schéma fondamental de Wikipedia qui, en tant qu'encyclopédie, est composées d'*articles*, portant sur un large éventail de domaines. Chaque article, sous la forme d'une page Web, réfère à un concept, une notion, un événement, ou une entité — personnalité, lieu, organisation, etc. Un article présente du texte libre, ainsi que des éléments de structuration :

Titre Nom canonique du sujet traité par l'article, éventuellement suivi d'un terme parenthésé donnant une indication de désambiguïsation dans le cas d'homonymies. L'article concernant le joueur de basketball Michael Jordan a ainsi pour titre `Michael Jordan`, tandis que celui portant sur le joueur de football britannique du même nom a pour titre

`Michael Jordan (footballer)`.

Résumé Court texte introductif placé en tête de l'article.

Catégories Une ou plusieurs catégories sont assignées à chaque article. L'information portée par les catégories peut être d'ordre thématique (*basketball*), typologique (*joueur de basketball*), historique ou événementiel (*JO 2012*), relatif à un type d'article (*naissance en 1963)*... Les catégories sont constituées en listes légèrement hiérarchisées mais ne correspondant pas à une modélisation conceptuelle systématique comme le serait une ontologie. On trouve par exemple la catégorie *Biographie* qui comprend les sous-catégories *Autobiographie*, *Film biographique* ou *Récit de voyage*.

Infobox Un sous-ensemble d'articles de Wikipedia présente des infobox, placées en regard du corps de l'article. Une infobox comprend des informations sur le sujet de l'article sous la forme d'un ensemble de couples d'attributs et de valeurs. Un article portant sur une personnalité peut par exemple présenter une infobox indiquant sa date et son lieu de naissance, ainsi que sa profession ou sa fonction dans une organisation. La valeur d'un attribut peut ainsi consister en un lien interne renvoyant à l'article de Wikipedia correspondant à cette valeur. Il est important de noter que les attributs choisis pour le remplissage des infobox, ainsi que les types d'infobox eux-mêmes, sont laissés au libre choix des éditeurs d'articles sans qu'un schéma systématique ne leur soit attaché. Différents noms d'attributs peuvent ainsi renvoyer au même type d'information (`birthPlace` et `placeOfBirth`, par exemple) ; différents schémas d'infobox peuvent par ailleurs exister pour la description d'entités de même type (`city_japan` et `swiss_town`, par exemple).

Wikilinks Des liens internes, renvoyant à d'autres articles de Wikipedia, peuvent être insérés au texte de l'article. Ces liens se situent au niveau des mentions textuelles des concepts ou entités ainsi jugés pertinents pour leur relation avec le sujet traité dans l'article courant. Les wikilinks constituent ainsi un des moyens principaux d'exploration et de valorisation de l'information mise à disposition dans l'ensemble de l'encyclopédie.

Liens externes Des liens pointant vers des ressources distinctes de Wikipedia peuvent être insérés à la fin des articles afin d'étendre l'espace informatif relatif au sujet traité.

Liens interlingues L'encyclopédie Wikipedia faisant l'objet de plusieurs éditions linguistiques (285 langues donnent lieu à une édition de Wikipedia à ce jour), un article dont le même sujet est traité dans une ou plusieurs autres éditions linguistiques présente un lien permettant d'accéder à chaque édition.

La figure 3.4 donne une description schématique d'un article Wikipedia muni d'une infobox [9].

En dehors des articles, Wikipedia fournit également des liens de *redirections* ainsi que des pages de *désambiguïsation*. Les premiers correspondent aux variantes lexicales pouvant désigner le sujet d'un article et pointent vers l'article en question. Les redirections prennent notamment en charge le phénomène des alias, pseudonymes et changements de nom au cours du temps (une requête avec la chaîne *Ali le chimique* déclenche ainsi une redirection vers l'article intitulé `Ali Hassan al-Majid` ; *Carla Bruni* est redirigé vers l'article `Carla Bruni-Sarkozy`), ainsi que des variations dues à des erreurs orthographiques ou des formes incomplètes (*Hilary Clinton* est ainsi redirigé vers `Hillary Rodham Clinton`). Les pages de désambiguïsation sont présentées à l'utilisateur lorsque le sujet recherché est concerné par l'homonymie ; les homonymes possibles sont ainsi listés avec une courte description permettant une désambiguïsation et un accès à l'article

9. La version française de Wikipedia comporte un peu moins de 50% d'articles munis d'une infobox

adéquat, comme l'illustre l'extrait d'une page de désambiguïsation de Wikipedia, représenté à la figure 3.5.

Figure 3.5 : Page de désambiguïsation de Wikipedia pour *Michael Jordan*. Les noms listés correspondent aux liens vers les articles concernés.

À partir des contenus de Wikipedia ainsi mis à disposition, DBpedia opère une conversion systématique aboutissant à une *base de connaissances*, selon le terme employé par les auteurs à l'origine de ce projet. Une base de connaissances est en effet un regroupement d'informations concernant un domaine — ici conçu comme général —, structuré de façon à en dériver des connaissances et de forme exploitable automatiquement. Les éléments principaux de DBpedia correspondent aux sujets faisant l'objet d'articles dans Wikipedia. Il s'agit de concepts et d'entités munis d'un identifiant unique, qui leur donne ainsi le statut de ressources Web et les rend accessibles au titre des LD. Des connaissances relatives à ces concepts et entités sont également dérivées de Wikipedia. Une ontologie élaborée manuellement par les développeurs de DBpedia à partir des 350 types d'infobox les plus courants dans Wikipedia, constituée de 170 classes peu hiérarchisées et 720 relations, permet une catégorisation des concepts et entités ainsi qu'une représentation formelle des connaissances les concernant. À cette ontologie s'ajoutent une modélisation selon les catégories Wikipedia et deux schémas externes, Yago[10] et UMBEL[11]. Enfin, des liens externes sont définis à partir de DBpedia vers d'autres ressources d'information sur le Web, tandis que DBpedia fait l'objet de liens issus de nœuds des LD, pointant vers les concepts et entités qui y sont référencés. Ces liens entrants et sortants contribuent à placer DBpedia au cœur des LD, lui conférant un statut de pivot essentiel à leur fonctionnement et venant s'ajouter à la pertinence de DBpedia pour de nombreux domaines et contextes applicatifs, dérivée de la large couverture de Wikipedia. Fin 2012, la présentation[12] de la base de connaissances de DBpedia indique pour sa version anglaise 3,77 millions de ressources, dont 2,35 millions sont catégorisées selon l'ontologie correspondante. Les catégories modélisant les personnes, lieux et organisations comptent respectivement 764 000, 573 000 et 192 000 ressources. DBpedia est par ailleurs développée en 111 autres langues, donnant lieu à autant de versions qui comptent au total 20,8 millions de ressources, dont 10,5 millions comportent des liens avec des ressources de la version anglaise. DBpedia est en constante évolution et augmentation, en raison de sa synchronisation avec les changements et additions réalisées dans l'encyclopédie Wikipedia et de l'organisation de son développement et de sa maintenance reposant sur la méthode d'externalisation ouverte (*crowdsourcing* en anglais).

10. `http://www.mpi-inf.mpg.de/yago-naga/yago/`
11. `http://www.umbel.org/`
12. Description disponible et mise à jour sur `http://dbpedia.org/About`

L'opération de conversion de Wikipedia vers DBpedia concerne d'une part les éléments communs à tout article, et d'autre part la représentation dans DBpedia des informations contenues dans les infobox, relatives aux sujets d'articles pour lesquels une infobox est définie. Cette seconde conversion se fait selon deux procédés : l'un transfère directement les informations d'infobox de Wikipedia à DBpedia, l'autre intègre une mise en correspondance entre le format des infobox et le modèle ontologique de DBpedia. La table 3.1 rend compte de ce processus de conversion et liste les principaux éléments constitutifs de la représentation formelle adoptée dans DBpedia. La figure 3.6 présente un extrait de la description obtenue pour une ressource de DBpedia *via* le Web. Les tables 3.3 et 3.2 illustrent la position de DBpedia au centre des LD, *via* les liens entrants et sortants établis avec d'autres sources de données, que les auteurs évaluent à 3,1 millions et 4,9 millions, respectivement.

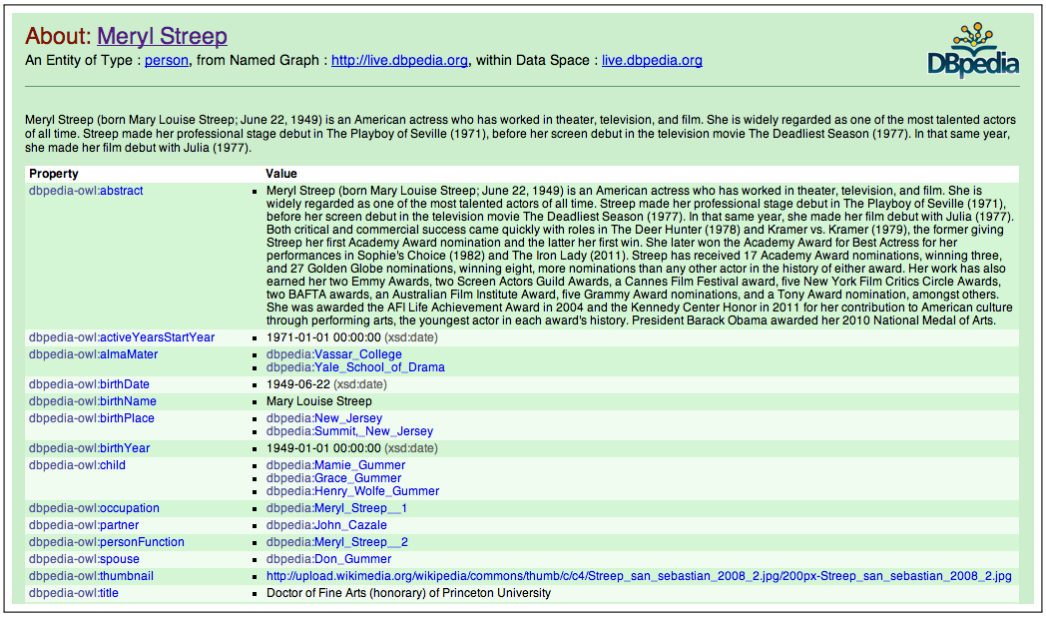
Élément DBpedia	**Description**	**Élément d'article Wikipedia utilisé**
URI	Identifiant unique de ressource	URI de DBpedia + titre (version anglaise)
`rdfs:label`	Variantes dénotationnelles	Titre
		Liens interlingues
		Liens de redirection
		Éléments de pages de désambiguïsation
`rdfs:comment`	Description courte	Premiers mots du résumé
dbpdia :abstract	Résumé	Résumé
`dbpedia:reference`	Références externes	Liens externes
`dbpedia:wikilink`	Lien Wikipedia interne	Wikilinks
`concept`	Catégories Wikipedia	Catégorie
`rdf:type`	Classe ontologique	Type de l'infobox
`dbpedia-owl:`*attribute*	Attribut de ressource	Attribut d'infobox
`dbpedia-owl:`*role*	Relation entre ressources	Attribut d'infobox

TABLE 3.1 : Construction de DBpedia à partir de Wikipedia.

GeoNames

Dans le réseau des LD, GeoNames[13] se présente comme le principal ensemble de données géographiques, disponibles sous licence libre[14]. Organisée en 9 types de lieux principaux, sous-divisés en 645 sous-types, la base de données fournie par GeoNames comprend 2,8 millions d'entrées et compte plus de 8 millions de noms différents associés à ces lieux, avec notamment des variantes linguistiques. Chaque entrée est identifiée de façon unique par une URI et directement accessible en tant qu'instance d'un des types ou sous-types définis. Une ontologie reprend ce modèle typologique et intègre pour chaque ressource ses relations de subsomption avec les autres ressources de GeoNames, ainsi que des liens de synonymie vers Wikipedia lorsqu'une telle association est possible. La figure 3.7 présente quelques types et sous-types utilisés dans le modèle de GeoNames. La figure 3.8 donne un exemple de ressource telle qu'elle est modélisée dans l'ontologie de GeoNames au format RDF.

13. `http://www.geonames.org`
14. Licence Creative Attribution 3.0 — `http://creativecommons.org/licenses/by/3.0/`



About: Meryl Streep

An Entity of Type : person, from Named Graph : http://live.dbpedia.org, within Data Space : live.dbpedia.org

Meryl Streep (born Mary Louise Streep; June 22, 1949) is an American actress who has worked in theater, television, and film. She is widely regarded as one of the most talented actors of all time. Streep made her professional stage debut in The Playboy of Seville (1971), before her screen debut in the television movie The Deadliest Season (1977). In that same year, she made her film debut with Julia (1977).

Property	Value
dbpedia-owl:abstract	Meryl Streep (born Mary Louise Streep; June 22, 1949) is an American actress who has worked in theater, television, and film. She is widely regarded as one of the most talented actors of all time. Streep made her professional stage debut in The Playboy of Seville (1971), before her screen debut in the television movie The Deadliest Season (1977). In that same year, she made her film debut with Julia (1977). Both critical and commercial success came quickly with roles in The Deer Hunter (1978) and Kramer vs. Kramer (1979), the former giving Streep her first Academy Award nomination and the latter her first win. She later won the Academy Award for Best Actress for her performances in Sophie's Choice (1982) and The Iron Lady (2011). Streep has received 17 Academy Award nominations, winning three, and 27 Golden Globe nominations, winning eight, more nominations than any other actor in the history of either award. Her work has also earned her two Emmy Awards, two Screen Actors Guild Awards, a Cannes Film Festival award, five New York Film Critics Circle Awards, two BAFTA awards, an Australian Film Institute Award, five Grammy Award nominations, and a Tony Award nomination, amongst others. She was awarded the AFI Life Achievement Award in 2004 and the Kennedy Center Honor in 2011 for her contribution to American culture through performing arts, the youngest actor in each award's history. President Barack Obama awarded her 2010 National Medal of Arts.
dbpedia-owl:activeYearsStartYear	1971-01-01 00:00:00 (xsd:date)
dbpedia-owl:almaMater	dbpedia:Vassar_College dbpedia:Yale_School_of_Drama
dbpedia-owl:birthDate	1949-06-22 (xsd:date)
dbpedia-owl:birthName	Mary Louise Streep
dbpedia-owl:birthPlace	dbpedia:New_Jersey dbpedia:Summit,_New_Jersey
dbpedia-owl:birthYear	1949-01-01 00:00:00 (xsd:date)
dbpedia-owl:child	dbpedia:Mamie_Gummer dbpedia:Grace_Gummer dbpedia:Henry_Wolfe_Gummer
dbpedia-owl:occupation	dbpedia:Meryl_Streep__1
dbpedia-owl:partner	dbpedia:John_Cazale
dbpedia-owl:personFunction	dbpedia:Meryl_Streep__2
dbpedia-owl:spouse	dbpedia:Don_Gummer
dbpedia-owl:thumbnail	http://upload.wikimedia.org/wikipedia/commons/thumb/c/c4/Streep_san_sebastian_2008_2.jpg/200px-Streep_san_sebastian_2008_2.jpg
dbpedia-owl:title	Doctor of Fine Arts (honorary) of Princeton University

Figure 3.6 : Extrait d'une description de ressource dans DBpedia (format de visualisation). Le préfixe `dbpedia` indique un lien vers une autre ressource.

Source	# Liens
Freebase	2 400 000
flickr wrappr	1 950 000
WordNet	330 000
GeoNames	85 000
OpenCyc	60 000
UMBEL	20 000
Bio2RDF	25 000
WikiCompany	25 000
MusicBrainz	23 000
Book Mashup	7 000
Project Gutenberg	2 500
DBLP Bibliography	200
CIA World Factbook	200
EuroStat	200

Table 3.2 : Distribution des liens de DBpedia pointant vers d'autres sources de données (table reproduite à partir de [Biz+09].

Source	Classes
BBC Music	musiciens, groupes
Bio2RDF	gènes, protéines
CrunchBase	entreprises
Diseasome	maladies
flickr wrappr	*classes diverses*
FOAF	*classes diverses*
GeoNames	lieux
GeoSpecies	espèces
LIBRIS	auteurs
LinkedMDB	films
Lingvoj	langues
OpenCyc	*classes diverses*
OpenCalais	lieux, personnes
UMBEL	*classes diverses*

Table 3.3 : Sources de données publiant des liens pointant vers DBpedia (table reproduite à partir de [Biz+09].

NLGbAse

NLGbAse[15], dont le processus de construction est décrit dans [CTM10] et [CG12], repose sur une dérivation d'information à partir de Wikipedia et indique pour chaque ressource référencée l'URI de la ressource DBpedia correspondante. Chaque ressource de NLGbAse comprend un type, un ensemble de variantes surfaciques possibles ainsi qu'un sac de mots dérivé de l'article Wikipedia

15. `http://www.nlgbase.org/`

A country, state, region,...		
ADM1	first-order administrative division	a primary administrative division of a country, such as a state in the United States
ADM1H	historical first-order administrative division	a former first-order administrative division
ADM2	second-order administrative division	a subdivision of a first-order administrative division
H stream, lake, ...		
AIRS	seaplane landing area	a place on a waterbody where floatplanes land and take off
ANCH	anchorage	an area where vessels may anchor
L parks,area, ...		
AGRC	agricultural colony	a tract of land set aside for agricultural settlement
P city, village,...		
PPL	populated place	a city, town, village, or other agglomeration of buildings where people live and work
PPLA	seat of a first-order administrative division	seat of a first-order administrative division (PPLC takes precedence over PPLA)
PPLA2	seat of a second-order administrative division	
PPLA3	seat of a third-order administrative division	
PPLA4	seat of a fourth-order administrative division	
PPLC	capital of a political entity	
PPLCH	historical capital of a political entity	a former capital of a political entity
S spot, building, farm		
ADMF	administrative facility	a government building
AGRF	agricultural facility	a building and/or tract of land used for improving agriculture

FIGURE 3.7 : Quelques types et sous-types de données géographiques dans GeoNames.

```
<rdf:RDF xmlns:cc="http://creativecommons.org/ns#" xmlns:dcterms="http://purl.org/dc/terms/" xmlns:foaf="http://xmlns.com/foaf/0.1/"
  xmlns:gn="http://www.geonames.org/ontology#" xmlns:owl="http://www.w3.org/2002/07/owl#" xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#" xmlns:wgs84_pos="http://www.w3.org/2003/01/geo/wgs84_pos#">
  <gn:Feature rdf:about="http://sws.geonames.org/3020251/">
    <rdfs:isDefinedBy rdf:resource="http://sws.geonames.org/3020251/about.rdf"/>
    <gn:name>Embrun</gn:name>
    <gn:alternateName xml:lang="oc">Ambrun</gn:alternateName>
    <gn:featureClass rdf:resource="http://www.geonames.org/ontology#P"/>
    <gn:featureCode rdf:resource="http://www.geonames.org/ontology#P.PPL"/>
    <gn:countryCode>FR</gn:countryCode>
    <gn:population>7069</gn:population><gn:postalCode>05200</gn:postalCode><gn:postalCode>05201</gn:postalCode>
    <gn:postalCode>05202</gn:postalCode><gn:postalCode>05208</gn:postalCode><gn:postalCode>05209</gn:postalCode>
    <wgs84_pos:lat>44.56387</wgs84_pos:lat><wgs84_pos:long>6.49526</wgs84_pos:long>
    <gn:parentFeature rdf:resource="http://sws.geonames.org/6446638/"/>
    <gn:parentCountry rdf:resource="http://sws.geonames.org/3017382/"/>
    <gn:parentADM1 rdf:resource="http://sws.geonames.org/2985244/"/>
    <gn:parentADM2 rdf:resource="http://sws.geonames.org/3013738/"/>
    <gn:parentADM3 rdf:resource="http://sws.geonames.org/3016701/"/>
    <gn:parentADM4 rdf:resource="http://sws.geonames.org/6446638/"/>
    <gn:nearbyFeatures rdf:resource="http://sws.geonames.org/3020251/nearby.rdf"/>
    <gn:locationMap rdf:resource="http://www.geonames.org/3020251/embrun.html"/>
    <gn:wikipediaArticle rdf:resource="http://de.wikipedia.org/wiki/Embrun"/>
    <gn:wikipediaArticle rdf:resource="http://en.wikipedia.org/wiki/Embrun%2C_Hautes-Alpes"/>
    <owl:seeAlso rdf:resource="http://dbpedia.org/resource/Embrun%2C_Hautes-Alpes"/>
    <gn:wikipediaArticle rdf:resource="http://fr.wikipedia.org/wiki/Embrun_%28Hautes-Alpes%29"/>
    <gn:wikipediaArticle rdf:resource="http://it.wikipedia.org/wiki/Embrun"/>
    <gn:wikipediaArticle rdf:resource="http://nl.wikipedia.org/wiki/Embrun"/>
    <gn:wikipediaArticle rdf:resource="http://oc.wikipedia.org/wiki/Ambrun"/>
    <gn:wikipediaArticle rdf:resource="http://pl.wikipedia.org/wiki/Embrun"/>
    <gn:wikipediaArticle rdf:resource="http://vo.wikipedia.org/wiki/Embrun_%28Hautes-Alpes%29"/>
  </gn:Feature>
</rdf:RDF>
```

FIGURE 3.8 : Description de ressource dans GeoNames au format RDF.

correspondant. La classification des ressources est obtenue à partir de Wikipedia selon un processus d'apprentissage. Les variantes surfaciques sont collectées à partir de cinq éditions linguistiques de Wikipedia (français, anglais, espagnol, allemand, italien), afin de prévoir des traitements multilingues, des pages de redirections, des pages de désambiguïsation, ainsi que du contenu textuel marqué par des wikilinks au sein des articles de l'encyclopédie. Le sac de mots associé à chaque ressource est constitué des mots de l'article Wikipedia correspondant ; une valeur TFIDF est assignée à chaque mot pour chaque édition linguistique utilisée. La table 3.4 présente la distribution des ressources de NLGbAse selon leur type pour le français, l'anglais et l'espagnol.

NLGbAse comprend donc à la fois un inventaire typé de ressources, dont les entités de type PERSON, ORGANIZATION et LOCATION représentent 62% et 64%, pour le français et l'anglais

	Person	Organization	Location	Product	Function	Time	Encyclopedic
FR	232 027	87 052	183 729	96 571	1 588	18 871	130 530
EN	754 586	305 706	565 941	326 155	3 783	13 575	468 829
ES	84 623	58 600	93 030	51 427	41	2 048	92 462

Table 3.4 : Distribution des ressources de NLGbAse par type pour le français, l'anglais et l'espagnol (table reproduite à partir de [CG12]).

respectivement, ainsi que des informations d'ordre contextuel et lexical sur ces ressources. Ces informations sont collectées dans une perspective d'exploitation par un système d'AS, qui sera décrit plus loin dans ce chapitre. La figure 3.9 illustre un exemple de ressource de NLGbAse dans le format de visualisation retourné par une requête en ligne sur `http://www.nlgbase.org/`. Le nom canonique (*Meryl Streep*), le type (pers.hum), les formes de surface, l'ancrage dans les LD à partir de DBpedia ainsi que les premiers mots associés à la ressource selon leur poids tfidf sont présentés dans ce format.

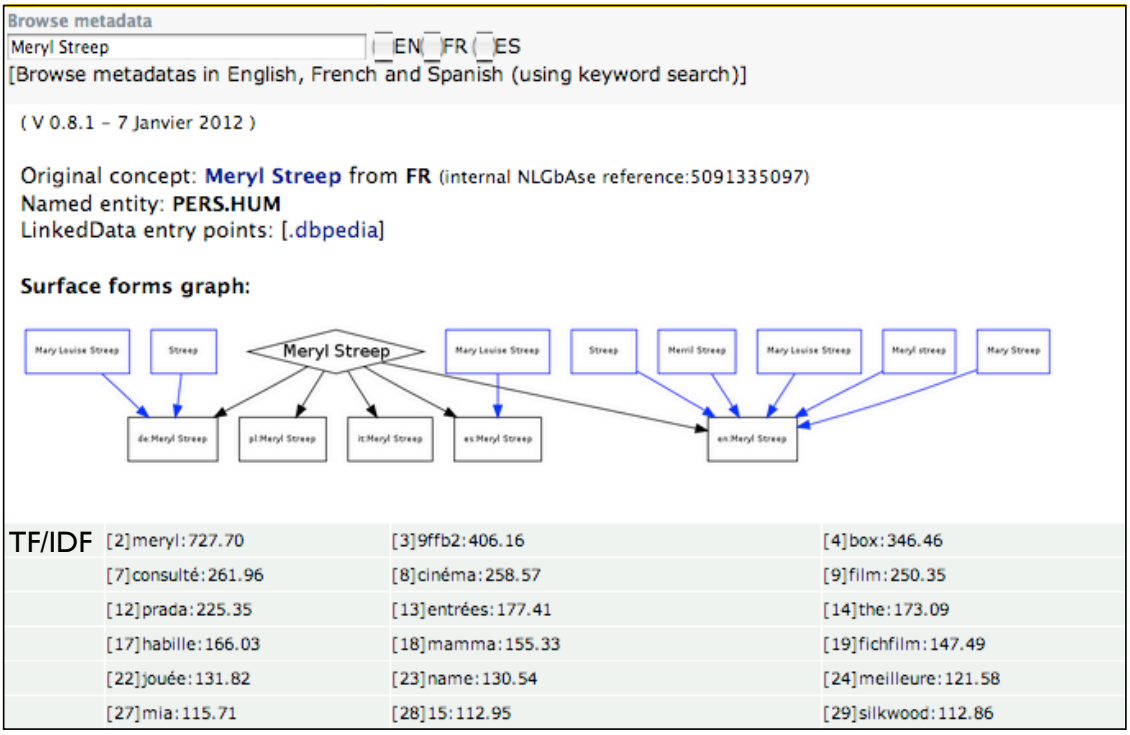


Figure 3.9 : Visualisation d'une ressource dans NLGbAse.

Aleda

Aleda[16] [SS12a] est issu d'une dérivation à partir de Wikipedia et de GeoNames, destinée au regroupement de ressources selon une classification comprenant les types d'entités courants. De façon comparable à NLGbAse, Aleda intègre les entités de type personne, organisation, et

16. Aleda est librement disponible sous licence LGPL-LR (`http://www.ida.liu.se/~sarst/bitse/lgpllr.html`) à l'adresse `https://gforge.inria.fr/frs/download.php/30598/aleda-0.5.tar.gz`, dans le cadre de la plateforme de modélisation et d'acquisition d'informations lexicales Alexina (`http://gforge.inria.fr/projects/alexina/`)

ENTREPRISE à partir de Wikipedia [17], tandis que les lieux sont obtenus à partir de GeoNames. Aleda opère par ailleurs cette dérivation exclusivement au niveau des entités, et ne comprend donc pas de ressources de type encyclopédique — concepts, notions — ni d'événement ou de dates. À partir de Wikipedia et de GeoNames, Aleda résulte ainsi en une base d'entités, disponible sous la forme d'une base de données classique, dont le schéma correspond à la modélisation adoptée. Aux types d'entités mentionnés s'ajoutent les œuvres (films, romans...) et produits. La table 3.5 indique le nombre de ressources pour chaque type dans la version française d'Aleda, autrement dit la dérivation effectuée à partir de l'édition française de Wikipedia et des ressources GeoNames pour lesquelles un label en français est renseigné [18]. Des versions anglaise, espagnole et allemande sont également en développement. Pour les lieux, le typage des ressources est immédiatement déduit de leur origine dans GeoNames. Au type LIEU s'ajoutent de possibles sous-types correspondant à ceux qui peuvent être définis dans GeoNames (cf. figure 3.7) [19]. La table 3.6 présente la correspondance entre sous-types d'Aleda et ceux de GeoNames. Pour les autres types, un processus semi-automatique prenant en compte les catégories d'articles Wikipedia ainsi que les éventuelles infobox d'articles permet de dériver une classe pour les ressources concernées. Décrit en détail dans [SS12a], ce processus s'appuie sur :

- Une correspondance établie manuellement entre les modèles d'infobox les plus fréquents et un type prévu dans Aleda : les modèles d'infobox *Person* ou *Infobox_officeholder* sont ainsi associés au type PERSON dans Aleda.
- Le nombre d'occurrences, dans une édition de Wikipedia donnée, de l'association entre chaque catégorie Wikipedia et chaque modèle d'infobox déjà typé, chaque article présentant cette association venant augmenter ce nombre. Par exemple, la catégorie *Naissance en 1978* apparaît dans 2 777 articles dont une majorité comporte une infobox dont le modèle a été associé au type PERSON d'Aleda ; cette catégorie est ainsi associée à ce même type.
- L'assignation d'un type aux articles Wikipedia dont les catégories et infobox, ou les catégories seules en l'absence d'infobox, ont été associés à ce type ; ainsi, tous les articles pour lesquels la catégorie *Naissance en 1978* est spécifiée donnent lieu à la création d'une ressource de type PERSON dans Aleda.

PERSONNE	ORGANISATION	ENTREPRISE	LIEU
304 158	41 543	18 109	465 926
PRODUIT	**ŒUVRE**	**FICTIONCHAR**	**Total**
2 526	83 713	6 729	**922 704**

TABLE 3.5 : Distribution des ressources par type dans la version française d'Aleda.

Chaque ressource est associée à des attributs typologiques et descriptifs, dont la table 3.7 illustre la correspondance avec Wikipedia et GeoNames. L'attribut *poids* est prévu pour modéliser l'aspect de popularité ou d'importance relative des entités, qui peut être utile dans des traitements comme l'AS où un tel critère peut conduire à privilégier une entité sur les autres. Ce poids est dérivé de la taille des articles Wikipedia, qui peut être vue comme le reflet de la popularité ou de l'importance

17. Des entités de type ŒUVRE , PRODUIT ainsi que FICTIONCHAR (personnages de fiction tels qu'Arsène Lupin) sont également extraites à partir de Wikipedia.

18. Lorsqu'une ressource de GeoNames ne présente aucune indication de langue pour son ou ses labels, ceux-ci sont également importés dans Aleda. Par ailleurs, ne sont importés dans Aleda que les lieux pour lesquels GeoNames indique une population supérieure à 200 habitants.

19. Une description complète de ces sous-types est disponible sur `http://www.geonames.org/export/codes.html`

Aleda	GeoNames	Exemples (nom canonique)
Region	L.RGN, L.CONT	Southeast Asia, Europe, Pays Basque
Territory	A.TERR	Western Sahara
Administrative	A.PCLS	Hong Kong Special Administrative Region, Palestine
Country	A.PCLI	Hellenic Republic
CountryDivision	A.ADMD, A.ADM1, A.ADM2, A.ADM3, A.PCLD	Département de l'Yonne, Texas, Los Angeles County
Capital	P.PPLC	Baghdad, Nicosia
City	P.PPL, P.PPLA, P.PPLA2, P.PPLA3, P.PPLA4, P.PPLA5, P.PPLG	Marseille, Hums
Stadium	S.STDM	Stade de France
CitySection	P.PPLX	Champs-Élysées, Chinatown
Airport	S.AIRP	Paris-Orly, Heraklion Airport
MilitaryBase	L.MILB	Air Force Flight Test Center
AtomicCenter	S.CTRA	Castle Bravo H-Bomb Test on Reef
Museum	S.MUS	Musée du Louvre
Monument	S.MNMT	Sphinx, Tour Eiffel
IsraeliSettlement	P.STLMT	Har Adar
SpaceCenter	S.CTRS	Centre Spatial Guyanais

Table 3.6 : Correspondance entre sous-types d'Aleda et de GeoNames.

d'une entité, puisque Wikipedia est construite sur un mode collaboratif, public et ouvert. Pour les ressources issues de GeoNames, il correspond au nombre d'habitants d'un lieu donné. La ville de Paris, capitale de la France, a ainsi un poids plus élevé (2 138 551) que la ville de Paris dans l'État du Texas aux États-Unis (25 171).

Attribut Aleda	Attribut Wikipedia	Attribut GeoNames
Identifiant	[nombre entier unique]	[nombre entier unique]
Nom canonique	titre d'article	nom canonique (souvent nom en anglais)
Type	typage Wikipedia	GeoNames → type location
Poids	taille de l'article	nombre d'habitants
Description	premiers mots du résumé	-
Lien	URI Wikipedia	URI GeoNames
Sous-type	-	correspondance (cf. table 3.6)
Code pays	-	code pays
Longitude	-	longitude
Latitude	-	latitude

Table 3.7 : Correspondance entre attributs Aleda et Wikipedia ou GeoNames.

La base de données d'Aleda définit également pour chaque ressource un ensemble de variantes surfaciques, obtenues :

- à partir du nom canonique attribué à la ressource GeoNames ou du titre de l'article Wikipedia correspondants,
- à partir des liens de redirections et des pages de désambiguïsation associés aux articles Wikipedia,

- à partir des labels renseignés pour chaque ressource dans GeoNames, avec une indication de langue correspondant à la version d'Aleda considérée,
- pour les personnes, par génération à partir du nom canonique : le nom *Hillary Rodham Clinton* donne ainsi lieu aux variantes *H. Clinton* (capitalisation du prénom et omission du « middle name »), *H. R. Clinton* (capitalisation du prénom et du « middle name »), *Hillary Clinton* (omission du « middle name »), *Clinton* (nom de famille seul), etc.

2 Mise en œuvre de l'Annotation Sémantique

La mise en œuvre de l'AS repose sur une méthodologie qu'il s'agit de dégager à partir de compte-rendus et de descriptions de systèmes présentés dans la littérature comme complètement ou partiellement dédiés à cette tâche. Cette méthodologie ne donne en effet pas lieu à une explicitation systématique, en raison d'un intérêt plus grand porté sur l'objectif d'intégration dans les LD par des communautés relevant d'activités d'ingénierie, ou plus généralement du caractère récent et principalement applicatif des contextes dans lesquels l'AS est envisagée. Une exception notoire à cette faible explicitation de la tâche, notamment en termes méthodologiques, se trouve dans les travaux relatifs à la plateforme KIM [Kir+04 ; Pop+03]. La définition de l'AS que nous proposons ici, ainsi que la formulation des différents composants et problèmes qui lui sont associés, s'inscrivent dans une perspective similaire à ces travaux. Cette parenté pourra être davantage explorée lors d'une présentation consacrée à KIM (cf. *infra*, section 2.2).

Les opérations constitutives d'une méthodologie générale d'AS seront en premier lieu identifiées à partir de la définition proposée en 1.1.2, notamment en tant qu'elles donnent lieu à des approches variées de la tâche, selon des critères à examiner. L'automatisation de l'AS est le lieu premier de mise en œuvre de ces différentes approches, qui seront illustrées par la présentation de quelques systèmes d'AS automatique. Le statut particulier des entités dans la tâche d'AS nécessite par ailleurs une explicitation afin d'orienter la réflexion méthodologique à mener sur l'accomplissement de l'AS dans notre cadre de travail.

2.1 Méthodologie

2.1.1 Opérations

La définition de l'AS proposé à la section 1.1.2 de ce chapitre peut être reprise en tenant compte de façon plus spécifique des ressources envisagées pour cette tâche telles qu'elles ont été décrites précédemment :

> [3.2] L'Annotation Sémantique consiste en la mise en relation explicite d'un support informatif, sous la forme d'un segment textuel de document, et d'un modèle de connaissances à l'aide de marqueurs indiquant les éléments informatifs sélectionnés pour l'établissement de ce lien et encodant formellement leur relation référentielle au modèle, représenté sous forme d'ontologie ancrée dans le réseau des Linked Data. Dans le cas particulier des segments textuels constituant des mentions d'entités, les liens établis par l'annotation peuvent cibler des instances ontologiques de classes modélisant les types d'entités considérés et correspondent alors à une identification.

Les deux opérations principales à effectuer afin d'accomplir une AS conformément à une telle définition relèvent d'un processus de sélection, appliqué à différents niveaux :

1. Les segments textuels destinés à être annotés doivent être sélectionnés dans l'ensemble du contenu documentaire considéré. Cette sélection peut couvrir tout ou partie de ce contenu : chaque mot ou terme d'un document peut en effet donner lieu à une annotation.

2. Pour chaque segment à annoter, un élément de l'ontologie adoptée pour la tâche doit être sélectionné. Cette sélection est encodée par l'insertion de l'URI de cet élément ontologique — concept, relation ou instance — dans le marqueur indiquant le segment annoté.

L'opération 1 est à considérer comme composante intégrante du processus rédactionnel à l'origine de la publication documentaire : la sélection dont il s'agit repose en effet sur un critère de pertinence lié à la valeur informative recherchée pour les documents publiés, puisque l'AS et plus particulièrement l'enrichissement de contenus à l'aide de métadonnées sont destinés à un ancrage dans un réseau informatif plus large ; les points d'entrée vers ce réseau, que constituent les annotations et métadonnées qui en dérivent, doivent ainsi présenter un intérêt en regard du contenu considéré, dont son auteur est à même de décider.

L'opération 2 implique, lors de l'accomplissement de la tâche d'AS, qu'un accès immédiat soit possible afin de sélectionner la ressource ontologique adéquate parmi l'ensemble disponible et d'en importer l'URI devant être insérée au marqueur d'annotation. Cet accès est à intégrer dans le fonctionnement des outils, notamment les *content management systems* (CMS) évoqués au chapitre 1 (section 3.3), mis à la disposition des auteurs et annotateurs. L'intégration et le support des modèles ontologiques constituent un pré-requis nécessaire, selon la description d'une plateforme d'AS proposée par Uren et al. [Ure+06]. Les descriptions associées aux ressources ontologiques dans le modèle adopté peuvent ainsi être consultées par les producteurs afin d'appuyer le processus de sélection. L'accès aux ressources peut par ailleurs inclure une fonctionnalité de contrôle et de restriction sur les éléments sélectionnés, évitant ainsi l'insertion de liens erronés ou inexistants et d'erreurs de syntaxe dans les marqueurs d'annotation.

Cette distribution opérationnelle de l'AS peut s'envisager dans un accomplissement manuel, concomitant au processus de production : à la rédaction de chaque document, l'interface d'édition permet une sélection des segments à annoter selon le choix du rédacteur, ainsi qu'un accès à l'ontologie adoptée dans laquelle le rédacteur sélectionne les ressources à lier à ces segments. Une telle configuration d'AS présente cependant des limitations inhérentes, déjà évoquées au chapitre 1 (section 1.3.1) : une annotation manuelle de chaque document produit au sein d'une organisation est nécessairement coûteuse en temps et peut représenter une charge de travail peu valorisée. Les manipulations textuelles et logicielles successives à effectuer pour chaque annotation peuvent en effet comporter un aspect laborieux et aride, rendant la tâche difficile à envisager dans la durée d'un point de vue à la fois pratique et intellectuel. Des freins de ce type sont évoqués dès 2001 dans la description d'une migration entre AS manuelle et semi-automatique par Erdmann et al. [Erd+00]. Le traitement d'archives, c'est-à-dire de corpus documentaires déjà publiés, qu'ils soient anciens ou récents, ajoute un obstacle majeur à l'AS manuelle, qui ne peut prendre en charge des quantités de données dépassant quelques documents par jour et par rédacteur ou annotateur. L'AS de contenus préexistants constitue pourtant un volet incontournable du traitement de l'information suivant le paradigme et l'architecture du Web Sémantique, dans lequel un tel traitement peut émaner de tout agent et porter sur tout contenu disponible. On peut rappeler ici l'argumentaire de Wilks et Brewster [WB09] selon lequel le TAL est indispensable à la réalisation de l'AS notamment pour cette raison.

2.1.2 Automatisation

L'AS se présente donc comme une tâche à automatiser, dans un but d'allègement de la charge de travail au niveau du processus rédactionnel d'une part, et de systématisation d'autre part. La nécessité de l'automatisation de l'AS est notamment affirmée par Uren et al. [Ure+06]. Dans l'inventaire de systèmes proposé par Reeve et Han [RH05], l'AS n'est envisagée que sous sa forme automatisée, dont les systèmes Armadillo [Cir+04], KIM [Kir+04] ou SemTag [Dil+03] sont des réalisations proposées au début des années 2000.

Les aspects de contrôle et de restriction de sélection des ressources évoqués précédemment sont des facteurs de cohérence et peuvent constituer une forme de semi-automatisation de l'AS à partir d'une configuration manuelle. Une automatisation plus étendue voire totale portant sur les opérations réalisées lors de l'AS permet cependant une systématisation de la chaîne de traitement dans son ensemble. Un niveau de traitement semi-automatique peut alors être introduit par la définition de certaines interventions humaines dans cette chaîne, principalement la validation manuelle des résultats retournés par l'outil automatique employé dans la tâche d'AS considérée, comme c'est le cas dans le système PANKOW [CHS04].

2.1.3 Points de variation dans les approches d'Annotation Sémantique

Les systèmes d'AS existants, dont certains seront étudiés dans la section suivante (2.2) à titre d'illustration de l'état de l'art dans ce domaine, peuvent être comparés et examinés selon leur prise en charge de l'automatisation au niveau des points méthodologiques évoqués précédemment.

Dans la plupart des systèmes dont le fonctionnement est rapporté dans la littérature, cette prise en charge se formule, pour la première opération de sélection effectuée en AS, en fonction de l'approche adoptée vis-à-vis du repérage automatique des éléments informatifs. Deux postures s'opposent à ce niveau, selon que le repérage en question repose ou non sur la discipline, les méthodes et les outils développés en Extraction d'Information.

Pour Uren et al. [Ure+06], même si les résultats obtenus peuvent présenter un taux d'erreur supérieur à un processus manuel en termes de précision et de rappel, l'intégration d'un composant d'Extraction d'Information dans un système d'AS est présentée comme indispensable. Les systèmes reposant pour la première opération sur un composant d'Extraction d'Information sont listés par Uren et al. [Ure+06] et Reeve et Han [RH05]. Il s'agit notamment de Armadillo [Cir+04], KIM [Kir+04] ou Mímir [Cun+11a] qui s'appuient sur l'architecture et les outils d'Extraction d'Information de GATE [Cun+11b], évoqués au cours du chapitre précédent. PANKOW [CHS04] repère les éléments à annoter à l'aide d'un étiqueteur en parties du discours, capable d'identifier les occurrences de noms propres.

Les systèmes SemTag [Dil+03] et Spotlight [Men+11a] sélectionnent les éléments à annoter à partir d'un lexique de formes lexicales associées aux ressources ontologiques dont ils disposent : toute forme appartenant au lexique repérée dans le texte est ainsi présentée à la seconde opération d'AS. La méthode adoptée par ces systèmes ne repose donc pas sur les techniques d'Extraction d'Information existantes. L'approche du système Cerno [Kiy+09] emploie des méthodes de repérage explicitement décrites comme ne relevant pas de l'Extraction d'Information ni du TAL, considérés comme trop lourds en termes de moyens et ressources informatiques, mais du processus de rétroingénierie pour la conception logicielle. Celui-ci peut être ramené à un repérage sur la base de motifs textuels qui, bien que considérée par les auteurs comme légère et peu sophistiquée, n'est cependant pas une technique étrangère à de nombreuses architectures d'Extraction d'Information.

Pour la seconde opération, l'établissement des relations entre mentions textuelles et ressources ontologiques nécessite un décodage qui peut être vu de façon plus ou moins naïve : les mentions peuvent en effet être directement associées aux ressources au vu de leur forme de surface, qui peut correspondre au nom donné à la ressource dénotée ou à un ensemble de formes lexicales définies pour cette ressource. Ainsi, la mention *troupes* dans l'exemple 6 peut être mise en relation avec un concept ontologique nommé `Troupes` dans l'ontologie, ou `Armée`, si ce concept définit cette forme lexicale comme pouvant lui être associée. De même, la mention *François Hollande* peut être mise en relation avec le concept `Personne` ou avec une instance de ce concept pour laquelle le label équivalent est défini. Mais dans la plupart des systèmes d'AS décrits, ce décodage tient compte du problème de l'ambiguïté touchant la relation entre texte et représentation logique : plusieurs ressources ontologiques peuvent en effet correspondre à une même forme lexicale, qu'il s'agit alors de lier à la ressource adéquate. L'établissement de cette

relation repose sur un ensemble d'heuristiques dans le système Kim [Kir+04], tandis que d'autres approches modélisent le problème de l'ambiguïté par une recherche de similarité entre le contexte textuel des occurrences de mentions et la description ontologique des ressources, comme le font les systèmes SemTag [Dil+03], Spotlight [Men+11a] ou Wikimeta [CGO11].

On constate également une variation parmi les différentes approches d'AS au niveau du type de ressource ontologique ciblé pour les annotations d'entités. Il a en effet été évoqué la possibilité de lier une mention d'entité, particulièrement les noms de personnes, de lieux ou d'organisation, à une instance ontologique membre d'une classe modélisant le concept correspondant. Cette possibilité s'oppose à l'établissement d'un tel lien avec la classe elle-même plutôt qu'une instance : dans ce cas, il s'agit principalement de typer l'entité mentionnée de la façon la plus précise possible en regard des concepts modélisés dans l'ontologie et de leur granularité. Ainsi, la mention *Michael Jordan* peut être liée aux concepts `Sportif`, `Économiste` ou `Scientifique`, définis comme sous-classes de la classe PERSONNE dans l'ontologie considérée. De tels liens sont notamment établis par le système Pankow [CHS04], produisant ainsi des documents annotés suivant les concepts de l'ontologie adoptée. Il est cependant à observer qu'une telle approche ramène aux problèmes de sémantique des entités évoqués au chapitre 2 (section 3) : un typage, aussi fin et précis soit-il, ne constitue pas l'établissement explicite d'un lien de référence entre une mention et une entité. Il ne peut donc être considéré comme l'équivalent d'une identification d'entité sous la forme d'une instance. Le système KIM, notamment, établit à l'inverse une relation systématique entre mention textuelle et instance ontologique d'entité, parallèlement à un typage conceptuel fin parmi les classes disponibles.

On constate de façon générale que le modèle ontologique employé en AS est de nature simple et légère, autrement dit que les modèles complexes, globaux et profonds tels que l'ontologie de haut niveau Cyc [20] n'ont pas la faveur des applications concrètes. Celles-ci reposent sur des ontologies définissant peu d'axiomes, restreints aux propriétés et attributs essentiels des concepts modélisés, eux-mêmes limités à des domaines d'ordre général. Cette tendance est illustrée par le mode de développement des LD, dont les modèles sous-jacents proposent des conceptualisations à large couverture mais de structure hiérarchique relativement plate, avec peu de modélisation relationnelle. Les ensembles de données des LD présentent en revanche souvent un grand nombre d'instances, et de fait un grand nombre d'interconnexions entre instances de différents ensembles de données. Ces éléments quantitatifs traduisent un intérêt davantage porté sur la mise en relation des informations elles-mêmes que sur la capacité à modéliser conceptuellement l'ensemble des informations véhiculées dans les contenus. Cette approche est un des traits caractéristiques des trois systèmes d'AS présentés dans la section suivante.

2.2 Exemples de systèmes d'Annotation Sémantique

Les systèmes d'AS se répartissent en fonction

- des choix méthodologiques tenant aux deux composants présentés ci-dessus : repérage automatique par Extraction d'Information ou non, degré de systématisation de la désambiguïsation lors du décodage de la relation entre texte et modèle,

- du type de modèle adopté : ressources ontologiques publiques, ancrées dans le réseau des LD, ou ontologie et population construites dans le cadre du système lui-même,

- à la place attribuée aux entités, selon qu'elles constituent le focus principal du système ou non.

20. www.cyc.com/

KIM (Ontotext)

La plateforme KIM [Kir+04 ; Pop+03], développée par la société Ontotext [21] est destinée à la mise en œuvre de l'AS ainsi qu'aux objectifs d'indexation et de recherche documentaire qui lui sont liés. KIM présente d'une part une ontologie (KIMO pour *KIM ontology*) et une base de connaissances constituée d'instances suivant le schéma de l'ontologie, et d'autre part une architecture d'Extraction d'Information dérivée de GATE [Cun+11b] et étendue à la tâche d'AS relativement à KIMO. Un serveur associé comprend les fonctionnalités d'indexation, de récupération de documents et d'interface avec l'utilisateur. KIM se concentre sur l'AS des entités, de façon similaire à notre objectif d'enrichissement de contenus textuels.

Les points principaux faisant de KIM un exemple intéressant de système d'AS sont :

1. une association non restrictive de la base de connaissances et de ses instances au schéma ontologique de KIMO,

2. une base de connaissances peuplée d'un grand nombre d'instances d'entités à partir de différentes sources,

3. un usage extensif des techniques de TAL et d'Extraction d'Information existantes, permettant la découverte de nouvelles instances lors de l'AS,

4. une sémantique des entités liée à leur description individuelle et non seulement à leur classe conceptuelle.

1 KIMO présente 250 classes et 100 propriétés modélisant les types d'entités les plus communs, d'après une exploration de corpus journalistique dont la nature est jugée pertinente pour l'obtention d'un degré de généralité satisfaisant dans le cadre de tâches d'AS variées. Elle présente un faible nombre d'axiomes, au titre d'un souci de simplicité formelle et algorithmique dans la perspective de traitements ultérieurs. La base de connaissances qui l'accompagne, que nous nommons KIM-KB pour *KIM knowledge base*, peut être considérée séparément : KIMO constitue en effet le schéma de modélisation adopté pour KIM-KB, sans que les informations stockées pour chaque instance ne doivent exactement lui correspondre. Autrement dit, une instance de KIM-KB peut être associée à des informations non prévues dans KIMO, ce qui permet la représentation d'individus dont l'association à l'ontologie est encore non effectuée ou impossible au vu des classes définies.

2 et 3 KIM-KB a fait l'objet d'une population initiale semi-automatique à partir de ressources librement disponibles, résultant en plus de 200 000 instances d'entités :

- 36 000 lieux, avec une hiérarchisation conceptuelle en sous-types comparable à celles de GeoNames et d'Aleda (cf. *infra* section 1.3),

- 147 000 organisations, dont les grandes organisations internationales (ONU, OTAN...), 140 000 entreprises de niveau international, et des informations de localisation les mettant en relation avec les instances de lieux,

- 6 000 personnes.

La population de KIM-KB est destinée à être enrichie au cours de l'AS elle-même : l'utilisation de techniques de TAL et d'Extraction d'Information permet en effet, notamment par l'application

21. www.ontotext.com/kim

de motifs surfaciques, de repérer des mentions d'entités ne présentant pas encore de ressource formelle dans KIM-KB, et ainsi de proposer de nouvelles entités pour cette population.

Chaque instance de KIM-KB comprend :

- une identification par URI,
- une association à l'une des classes de KIMO, avec un degré de spécification maximale,
- une description au format RDF,
- un ensemble d'interconnexions avec d'autres ressources,
- un ensemble de variantes lexicales pour l'anglais, le français et l'espagnol.

4 Les auteurs de la plateforme KIM insistent sur la nécessité, pour les entités, de recevoir un lien de nature référentielle vers une instance de KIM-KB afin que la sémantique recherchée par la tâche d'AS soit effectivement formalisée et exploitable. Une classification des mentions d'entités suivant la modélisation conceptuelle de l'ontologie est en effet présentée comme insuffisante à ce titre. L'intégration de techniques d'Extraction d'Information, au centre de la méthodologie de KIM, fait l'objet d'une adaptation importante destinée à les étendre au niveau sémantique. Le caractère non formel et non lié des modèles employés en Extraction d'Information traditionnelle est présenté comme la différence fondamentale avec une annotation véritablement sémantique qui, pour les entités, doit atteindre le niveau des instances en supplément d'un typage ontologique. Cet argument constitue un point important de notre approche du problème de l'identification d'entités, qui sera abordé de façon systématique dans la suite de ce travail.

Fonctionnement Partant de l'architecture de GATE, complétée par des extensions de niveau sémantique et sur le modèle ontologique de KIMO, KIM opère le repérage des mentions d'entités à annoter suivant une méthodologie traditionnelle de Reconnaissance d'Entités Nommées. Celle-ci repose sur

- un prétraitement textuel surfacique ;
- un automate représentant les motifs à reconnaître ;
- un lexique constitué des variantes lexicales définies pour les instances de KIMO ; ces variantes sont typées en fonction des instances qu'elles peuvent dénoter ;
- un ensemble d'informations sur lesquelles peuvent s'appuyer les règles de reconnaissance, par exemple une liste des suffixes de noms d'organisation (*Inc.*)

Les règles de reconnaissance sont adaptées à la tâche par un ensemble de spécifications relatives au modèle ontologique. Une règle pourra ainsi permettre la reconnaissance et la classification d'une entité de type montagne plutôt que lieu. Un lien est ensuite établi pour chaque mention repérée avec une instance de KIM-KB, sur la base de la correspondance entre variante lexicale et instance. Les ambiguïtés possibles à ce niveau — cas où une mention peut correspondre à plusieurs instances — sont évoquées par les auteurs, qui apportent la réponse suivante :

> *Il est difficile de répondre à ces questions dans un contexte général. KIM, comme de nombreux autres systèmes, implémente une série d'heuristiques afin d'y répondre avec une précision raisonnable.* (in [Kir+04], notre traduction)

Bien que des variantes lexicales soient définies pour les instances d'entités en anglais, français et espagnol, les langues effectivement traitées par KIM ne font pas l'objet d'une indication explicite dans les descriptions de la plateforme.

Évaluation KIM intègre de façon directe un composant de Reconnaissance d'Entités Nommées pour la mise en œuvre de l'AS à laquelle ce système est dédié. Les auteurs pointent l'absence de métrique établie pour l'AS [22], ainsi que celle de données de référence annotées manuellement selon le schéma ontologique dont il s'agit ici ou tout autre ontologie pouvant être mise en correspondance avec KIMO. L'évaluation de KIM porte donc sur la Reconnaissance d'Entités Nommées uniquement ; l'établissement des liens avec les instances d'entités, par ailleurs peu spécifié en termes méthodologique dans la présentation de KIM, n'en fait pas partie. Les résultats obtenus par KIM en Reconnaissance d'Entités Nommées sont reproduits dans la table 3.8 pour les types principaux. La F-mesure adoptée (colonne F1) accorde un poids égal à la précision et au rappel.

Type	Précision	Rappel	F1
PERSON	87,61	90,87	89,09
ORGANIZATION	82,29	71,30	76,03
LOCATION	92,77	89,77	91,23
Moyenne	87,56	83,98	85,45

TABLE 3.8 : Résultats de KIM en Reconnaissance d'Entités Nommées (adapté de [Kir+04]).

L'approche générale de l'AS par KIM repose ainsi sur un modèle et une base de connaissances pertinents pour le traitement des entités, ainsi que sur un usage des techniques de TAL et d'Extraction d'Information ayant fait preuve d'efficacité et de maturité après plusieurs décennies de recherches. Le problème de la mise en relation des mentions avec les instances d'entités auxquelles elles réfèrent est cependant abordé en peu de détails, bien qu'il puisse être considéré comme le centre des difficultés et des solutions à apporter pour la mise en œuvre de l'AS.

DBpedia Spotlight

Spotlight [23] s'inscrit dans l'effort communautaire de développement de DBpedia, qui constitue la ressource principale sur laquelle ce système s'appuie. La relation de DBpedia avec Wikipedia est également intégrée à Spotlight, où l'encyclopédie joue le rôle de corpus de données linguistiques dont diverses informations sont dérivées et exploitées dans le processus d'AS. Le modèle ontologique de Spotlight est donc celui de DBpedia, présenté à la section 1.3 et dont environ 62% des données sont classifiées en tant qu'instances de concepts. Contrairement à KIM, Spotlight cible ainsi directement le réseau des LD. L'AS réalisée par Spotlight concerne les concepts et entités communs et généraux. La langue traitée par Spotlight est l'anglais ; la mise au point de versions du système pour d'autres langues, à partir des éditions linguistiques correspondantes de DBpedia, relève d'initiatives libres de la communauté de développement autour de DBpedia, encouragée par les auteurs.

Mendes et al. [Men+11a] présentent Spotlight comme un système comparable à la méthodologie générale proposée précédemment. Il s'agit d'attribuer à toute occurrence de mention de concept ou d'instance de DBpedia l'URI correspondante. Spotlight s'appuie :

- sur la structuration des informations rendue disponible par DBpedia : chaque concept et instance est associé à un ensemble de variantes lexicales, obtenues à partir de Wikipedia (titre d'article, liens de redirection, pages de désambiguïsation, wikilinks — cf. section 1.3 et figure 3.4),

22. Le sujet de l'évaluation de la tâche d'Annotation Sémantique ainsi que des données annotées qui lui sont nécessaires sera abordé dans la suite de ce travail, lors de la présentation de notre système d'identification d'entités (chapitre cha :nomos).

23. spotlight.dbpedia.org

- sur la distribution des mentions dans le corpus d'articles de Wikipedia : pour chaque concept ou instance, chaque occurrence de l'une de ses mentions dans un wikilink incrémente un compteur général de fréquence, indiquant un niveau de notoriété et à ce titre comparable à l'attribut *poids* modélisé dans la base d'entités Aleda (cf. section 1.3) ; le paragraphe d'article dans lequel apparaît une de ces mentions est par ailleurs stocké sous forme de sac de mots, après tokenisation, stemming et filtrage par *stoplist*[24]. Chaque concept ou instance est ainsi associé à la modélisation d'un contexte lexical canonique.

Fonctionnement Spotlight effectue l'AS de contenus textuels en trois étapes principales :

Repérage (Spotting) Correspondant au premier composant méthodologique proposé dans la section précédente (2.1.1), ce repérage ne fait pas intervenir de techniques particulières relevant de l'Extraction d'Information. Il est réalisé à l'aide de l'outil de traitement de données textuelles LingPipe[25] dont le module *Exact Dictionary-Based Chunker* applique l'algorithme de repérage de chaîne de Aho et Corasick [AC75] sur la base du lexique de variantes fournis par DBpedia. Toute variante possible est ainsi repérée, avec une priorité à la chaîne la plus longue en cas de chevauchement ou d'imbrication de chaînes, et présentée à l'étape suivante.

Sélection de candidats Pour chaque mention repérée, Spotlight constitue un ensemble de candidats possibles pour l'établissement de sa relation référentielle au modèle. Ces candidats sont les concepts ou instances de DBpedia présentant la mention considérée dans leur ensemble de variantes lexicales. Cette étape permet également de définir un lien référentiel par défaut pour la mention, si l'on ne considère pas la seconde opération de sélection de la troisième étape : ce lien par défaut peut correspondre au candidat au niveau de notoriété maximal, tel que défini plus haut (nombre d'occurrence dans Wikipedia).

Désambiguïsation Le choix du candidat adéquat pour l'établissement de la relation référentielle entre mention et modèle est ramené à un problème de désambiguïsation. Celle-ci porte sur un ensemble de paires de contextes lexicaux, chaque paire représentant un candidat et la mention courante à partir (i) du sac de mots stocké pour chaque candidat à partir de ses occurrences dans Wikipedia, et (ii) du sac de mots correspondant au contexte d'occurrence de la mention courante, dérivé en sac de mots de façon identique au premier.
Chaque paire est ensuite caractérisée par une fonction de similarité : les contextes lexicaux donnent lieu à une modélisation vectorielle, à la manière d'un document en Recherche d'Information, chaque mot constituant un point de l'espace modélisé et recevant un score TFICF. Le poids TF (*term frequency*) mesure la pertinence locale d'un mot pour un document donné, tandis que le poids ICF (*inverse candidate frequency*, distinct du poids IDF (*inverse document frequency)* usuel en Recherche d'Information, mesure la pertinence d'un mot pour un candidat donné ; le pouvoir discriminant d'un mot parmi plusieurs candidats est ainsi vu comme inversement proportionnel au nombre de candidats *courants* auxquels il est associé, et non au nombre total d'instances auxquelles il est associé dans la totalité de DBpedia. Une fonction de similarité cosinus appliquée au vecteur de mots du candidat et celui de la mention assigne un score à cette paire. Toutes les paires (mention, candidat) sont ensuite ordonnées selon ce score de similarité contextuelle. Le candidat présent dans la paire maximisant cette similarité est sélectionné pour l'établissement du lien référentiel entre la ressource et le modèle.

24. Une *stoplist* est établie par inventaire de mots considérés comme non pertinents pour une analyse textuelle donnée, en général constitué des catégories de mots grammaticaux (prépositions, déterminants, conjonctions, etc.) ainsi que des mots les plus courants dans une langue donnée.

25. `http://alias-i.com/lingpipe/`

Spotlight définit par ailleurs un ensemble de paramètres de configuration destinés à modifier le comportement du système en fonctions de besoins particuliers d'utilisateurs. Il est ainsi possible de restreindre l'espace du modèle fourni par DBpedia afin de n'obtenir des annotations que sur un ensemble de concepts et d'entités correspondant au domaine traité. Cet espace peut également être réduit par l'élimination d'annotations mettant en jeu des ressources — concepts ou entités — peu communes ; le caractère commun et notoire des ressources est alors modélisé par un nombre minimal de wikilinks pointant vers une ressource considérée. Les ressources considérées comme peu pertinentes en regard du document peuvent être éliminées par la définition d'un seuil de similarité en-deçà duquel une annotation est éliminée. Enfin, au niveau de l'ambiguïté prise en charge par Spotlight, une attention particulière portée sur la précision des résultats peut conduire au rejet des annotations pour lesquelles plusieurs ressources semblent partager une forte similarité, indiquant ainsi une ambiguïté au sein du même domaine contextuel ; l'étape de désambiguïsation définit par ailleurs un paramètre de confiance entre 0 et 1 dont il est possible de fixer un seuil en-deçà duquel l'annotation peut être éliminée.

La figure 3.10 montre une capture d'écran de l'interface de démonstration de Spotlight [26], sur laquelle apparaît le formulaire de paramétrage de la tâche. Les liens obtenus par l'AS effectuée pointent vers les pages Web des ressources DBpedia correspondantes. Spotlight est librement disponible sous sa forme de code source ainsi que comme service accessible *via* le Web ou sur un serveur local [27].

Wikimeta

Comme Spotlight, Wikimeta, décrit dans [CGO11] aborde la tâche d'AS relativement aux LD en proposant des liens référentiels vers DBpedia. Ce système emploie cependant une modélisation intermédiaire des éléments de connaissance nécessaires à la réalisation de la seconde opératoin de l'AS. Wikimeta repose en effet sur la base NLGbAse, décrite à la section 1.3. NLGbAse dispose, pour chaque ressource, d'un ensemble de variantes lexicales, dont les labels pour l'anglais, le français, l'allemand, l'italien et l'espagnol, d'un ensemble de mots contenus dans l'article Wikipedia correspondant associés à leur poids TFIDF, ainsi que d'une URI l'associant aux LD *via* DBpedia.

L'opération de repérage des éléments à annoter est accomplie par Wikimeta à l'aide d'un module de Reconnaissance d'Entités Nommées, reposant sur le modèle statistique des CRF [LMP01] et intègre ainsi pleinement les capacités de reconnaissance de l'Extraction d'Information, de façon comparable à KIM, qui utilise des méthodes d'Extraction d'Information symbolique, et à la différence de Spotlight. La Reconnaissance d'Entités Nommées permet notamment de ne pas limiter les mentions reconnues à celles correspondant aux variantes lexicales fournies par NLGbAse. Un algorithme dédié à la seconde opération de l'AS, consistant à lier les mentions repérées à la ressource adoptée, identifie ce problème en termes de désambiguïsation entre plusieurs candidats, comme le fait Spotlight. L'ensemble à désambiguïser est constitué des informations de NLGbAse — sac de mots et leur poids TFIDF — pour tous les candidats dont la mention constitue une variante lexicale, ainsi que du contexte lexical d'occurrence de la mention, sous forme d'une fenêtre gauche et droite de n mots autour de la mention. Mais Wikimeta envisage trois configurations différentes pour l'étape de désambiguïsation, selon que :

- l'ensemble des candidats est vide : la mention repérée ne correspond à aucune variante lexicale fournie par NLGbAse,
- l'ensemble des candidats contient un seul élément,

26. `http://dbpedia-spotlight.github.com/demo/`
27. Code source : `https://github.com/dbpedia-spotlight/dbpedia-spotlight`
Service Web : `https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki/Web-service`
Serveur local : `https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki/Installation`

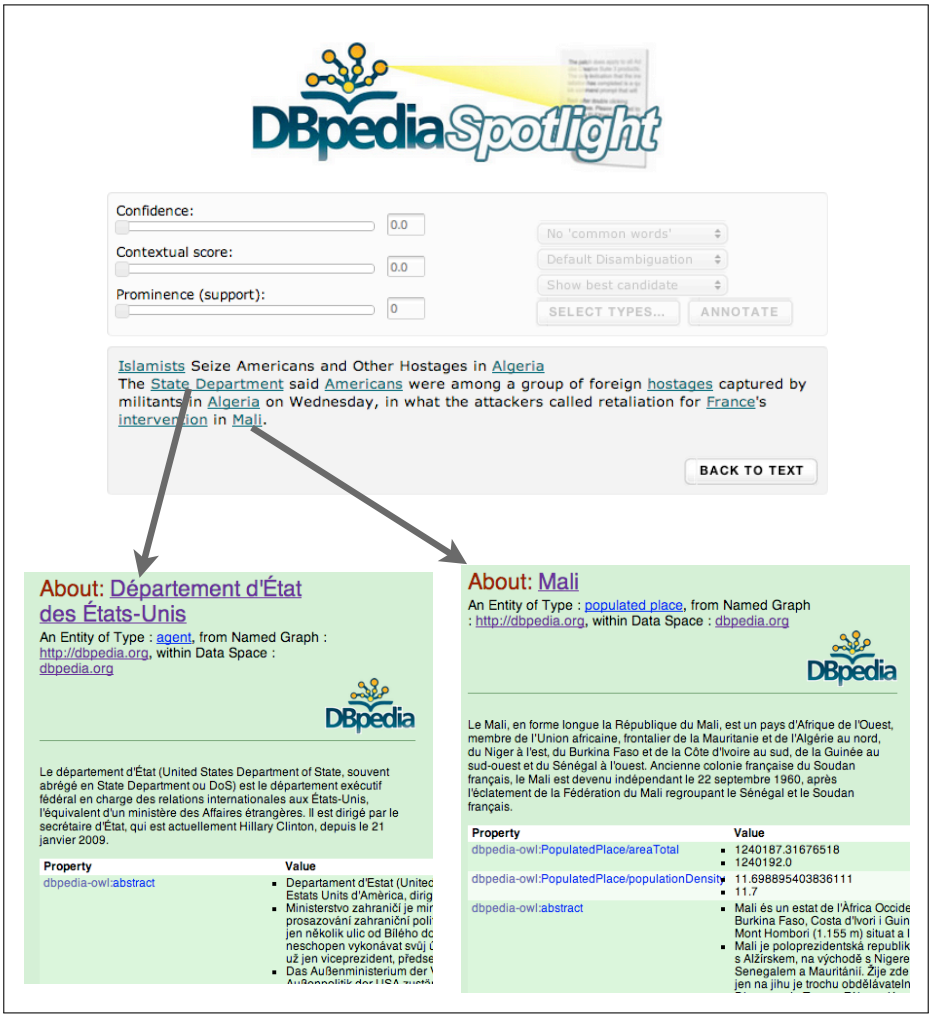


Figure 3.10 : Interface Web de démonstration du système Spotlight.

- l'ensemble des candidats contient plusieurs éléments.

Le premier cas est jugé trivial, en l'absence de lien à assigner à l'annotation. Dans les deuxième et troisième cas, Wikimeta applique une mesure de similarité cosinus entre le sac de mots associés à un poids TFIDF disponible pour chaque candidat et la fenêtre de contexte lexical de la mention, suivant une modélisation vectorielle de ces contextes comme c'est également le cas pour le système Spotlight. Un seuil de similarité minimale est défini, au-dessus duquel le candidat obtenant la similarité maximale avec la mention est sélectionné pour l'établissement du lien référentiel. Autrement dit, une annotation peut demeurer sans lien après cette étape si aucun des candidats ne dépasse ce seuil de similarité.

Évaluation Comme pour le système KIM, la description de Wikimeta donnée dans [CGO11] souligne l'absence de métriques ainsi que de données de référence disponibles pour la tâche d'AS. Wikimeta est évalué, pour le français et l'anglais, à l'aide de corpus annotés usuellement destinés à la tâche de Reconnaissance d'Entités Nommées. L'annotation de ces ressources est étendue à la tâche d'AS par les auteurs eux-mêmes, qui procèdent à l'augmentation nécessaire selon une procédure semi-automatique. Le corpus français ESTER (cf. chapitre 2, section 3.1) ainsi

que le corpus du Wall Street Journal[28] (WSJ), ont été annotés par le système Wikimeta, puis les reconnaissances de mentions ainsi que les insertions automatiques de liens ont été manuellement corrigées. La capacité du système à reconnaître les cas ne devant pas donner lieu à un lien vers le modèle est particulièrement importante au vu de la couverture de la ressource utilisée par rapport aux corpus à annoter. NLGbAse couvre en effet 83% des lieux mentionnés dans ESTER mais 44% des personnes ; dans le corpus WSJ, la couverture est de 96% et 62%, respectivement.

L'évaluation de Wikimeta porte sur deux fonctionnalités du système : *a*) la capacité à lier correctement une mention au modèle lorsqu'un tel lien existe, *b*) la capacité à identifier une absence de lien lorsque la mention ne correspond à aucune ressource du modèle. Le point d'évaluation a concerne donc uniquement l'ensemble des mentions correspondant effectivement à une ressource du modèle et la qualité de l'algorithme de désambiguïsation. Le point b évalue quant à lui la justesse du seuil de similarité minimale fixé pour l'étape désambiguïsation. Les performances de la REN réalisée en amont de l'AS ne sont quant à elles pas évaluées. L'évaluation du système pour l'anglais et le français porte donc uniquement sur l'ensemble des mentions d'entités correctement détectées et emploie la mesure du rappel r suivante :

$$r = \frac{\text{total des annotations correctes}}{\text{total de référence}}$$

Les résultats de cette évaluation sont reproduits à la table 3.9. L'ensemble des mentions de référence devant être associées au modèle est noté [a], les mentions de référence sans lien à établir avec le modèle sont notées [no a]

	Français				Anglais			
Mentions	[no a]	r	[a]	r	[no a]	r	[a]	r
PERSON	483	0,96	1 096	0,91	380	0,93	612	0,94
ORGANIZATION	764	0,91	1 204	0,90	1 129	0,85	1 608	0,86
LOCATION	1 017	0,94	1 218	0,92	709	0,84	739	0,82
PRODUCT	23	0,60	59	0,50	60	0,85	61	0,85
Total	2 287	0,93	3 577	0,90	2 278	0,86	3 020	0,86

TABLE 3.9 : Évaluation de Wikimeta sur les corpus ESTER (français) et WSJ (anglais) (reproduit à partir de [CGO11]).

2.3 Place et traitement des entités dans l'Annotation Sémantique

Les systèmes d'AS permettent la mise en relation automatique de contenus textuels avec des modèles sémantiques de domaine ou généralistes, par l'annotation de segments sélectionnés à travers ces contenus et porteurs d'une valeur informative pertinente. Parmi ces segments, les mentions d'entités et les instances ontologiques correspondantes font l'objet d'une attention plus ou moins centrale dans les systèmes évoqués et décrits. Cette attention, particulièrement mise en avant par KIM [Kir+04] ou Wikimeta [CGO11] notamment, reflète le rôle central accordé aux entités dans le traitement de l'information, comme l'a montré la présentation de l'Extraction d'Information au chapitre 2. L'AS se présente en ce sens comme un prolongement méthodologique de l'Extraction d'Information : son objectif de formalisation des connaissances et d'application générique, notamment *via* le Web et les LD, la distingue de l'Extraction d'Information, mais elle en constitue également une forme renouvelée en tant que moyen de représentation des connaissances sous une forme structurée et distincte du niveau textuel.

28. Dans sa version correspondant à la tâche d'évaluation (*Shared Task*) de la campagne CoNLL de 2008 [Sur+08].

Lorsque les entités sont concernées par l'AS, une distinction formelle importante est à observer quant à la relation établie avec le modèle employé : les mentions d'entités font en effet l'objet d'un lien vers une instance ontologique plutôt que vers un concept — bien qu'une telle possibilité soit envisagée par certains systèmes et revient alors à un typage sémantique plus ou moins fin. La correspondance formelle entre instance et entité peut en effet être vue comme plus manifeste que dans le cas d'autres classes conceptuelles : dans le passage

(8) un texte retentissant réclamant la fin de l'armement nucléaire [29]

le terme *armement nucléaire* peut en effet être mis en relation avec des concepts ontologiques tels que `Arme` ou `ArtefactMilitaire`, mais il serait difficile d'affirmer qu'il s'agit là d'une référence à un objet vu comme un individu, pouvant donner lieu à une instance ontologique de l'un de ces concepts. Ce terme en constituerait plutôt une mention textuelle que l'AS peut reconnaître comme telle, permettant ainsi une représentation conceptuelle de l'information contenue dans les données textuelles traitées. Dans le cas d'annotation de mentions d'entités mises en relation avec des instances, l'AS porte sur l'implication d'individus dans le contenu informatif, et non plus seulement sur une conceptualisation de l'information.

Le statut d'individus porté par les entités est appuyé par leur représentation ontologique elle-même, sous forme d'instances. Celles-ci sont en effet désignées par le terme *individu* dans la terminologie des logiques de description et peuvent être déclarées comme uniques et distinctes les unes des autres dans le langage ontologique OWL. Leur intégration dans l'espace du Web Sémantique et des LD renforce ce statut : elles y sont qualifiées de *descriptions* lorsqu'elles sont déclarées explicitement comme membres d'ensemble de données définis dans cet espace, tels que DBpedia. Cette désignation implique que l'accès à une instance *via* le Web et les LD permet également l'accès à des connaissances la concernant, définies dans l'ontologie dont elle est membre ou par un ensemble de triplets RDF. Les instances d'entités constituent ainsi une représentation de ces entités permettant de déclarer explicitement, *via* l'AS, de qui ou de quoi il est question dans un document.

Reconnaissance Le traitement des entités dans l'AS donne donc lieu à une forme de spécialisation, où *a*) le repérage des éléments à annoter correspond à la Reconnaissance d'Entités Nommées, telle que définie en Extraction d'Information (cf. chapitre 2, section 3.1), et où *b*) l'établissement d'une relation au modèle cible des instances de classes modélisant des entités.

a) La relation entre AS et Extraction d'Information ne relève pas uniquement de la parenté, évoquée précédemment, mais aussi de la méthodologie : l'AS en tant que telle ne définit pas de méthode d'accès aux éléments informatifs eux-mêmes. Elle repose pour cette étape nécessaire sur l'Extraction d'Information, de façon plus ou moins explicite et manifeste. Certaines applications d'AS n'en font ainsi pas mention (Spotlight), tandis que d'autres l'intègrent (Wikimeta) voire identifient la tâche d'AS à une forme adaptée d'Extraction d'Information (KIM). Même dans le premier cas, un processus de repérage est nécessaire et s'apparente indéniablement à l'Extraction d'Information. La Reconnaissance d'Entités Nommées, développée depuis plusieurs décennies spécifiquement pour le problème du repérage de mentions d'entités, se présente dès lors comme pertinente en tant que méthode éprouvée et aux résultats performants. Contrairement à des méthodes usant exclusivement de lexiques de mentions, s'appuyant uniquement sur la correspondance entre segment textuel et entrée du lexique, la REN permet de distinguer les segments effectivement dénotationnels des autres, y compris pour les segments présents dans un lexique. La REN dépasse également l'approche par recherche de chaînes définies dans un lexique par sa capacité à découvrir des mentions non préalablement listées comme telles et ainsi d'élargir la couverture des éléments informatifs repérés.

29. Exemple extrait du site `http://www.lepoint.fr` daté du 15 juillet 2012

b) En associant des mentions d'entités à des instances ontologiques plutôt qu'à des concepts, l'AS accomplit une explicitation de la relation dénotationnelle existant entre une mention et une entité. Elle dépasse ainsi les limitations de la sémantique typologique proposée par l'Extraction d'Information, qui ne porte pas sur l'entité elle-même, en tant qu'individu, mais sur son aspect conceptuel. Celui-ci est également traité par l'AS, qui en donne une définition explicite par l'ancrage dans un modèle ontologique, dont la formalisation s'oppose à des structures de modélisation non liées en Extraction d'Information.

Désambiguïsation La seconde opération d'AS consistant à établir un lien entre mention et modèle, qui peut être restreint aux instances dans le cas des entités, donne lieu à la formulation d'un problème d'ambiguïté, tel que le formulent les auteurs de KIM, Spotlight ou Wikimeta. L'ambiguïté est ici celle de variantes lexicales pouvant constituer les mentions de plusieurs instances du modèle. Elle se rapporte au problème plus général de la dénotation linguistique des entités, abordé au chapitre 2 (section 3.2), qui se traduit par une relation équivoque entre mentions et entités. La chaîne parlée et écrite permet en effet de mentionner un même référent par plusieurs expressions, donnant ainsi lieu à autant de variantes lexicales. Cet aspect de la dénotation est souvent pris en charge par les systèmes d'AS par l'usage de ressources définissant les possibles variantes correspondant aux différentes instances d'entités, comme cela a été décrit précédemment. Il est également possible qu'une unique variante puisse dénoter plusieurs entités distinctes, ce qui correspond au phénomène de l'homonymie. Celle-ci est au centre du problème d'ambiguïté formulé en AS, pour les systèmes le prenant explicitement en compte. C'est le cas de Spotlight et Wikimeta, dont l'approche consiste en une désambiguïsation sur la base de similarités contextuelles d'occurrence des mentions et des entités correspondantes.

On peut rapprocher cette prise en charge de l'ambiguïté dénotationnelle par l'AS de travaux réalisés au milieu des années 2000, portant sur la désambiguïsation des entités nommées. Il s'agit principalement des propositions de Bunescu et Pasca [BP06] et de Cucerzan [Cuc07], où la désambiguïsation repose sur les informations encyclopédiques de Wikipedia. Sans s'inscrire dans le paradigme du Web Sémantique et de l'AS, ni définir explicitement de distinction entre la tâche de Reconnaissance d'Entités Nommées et de mise en relation entre le niveau textuel des dénotations et une représentation formelle des entités, ces travaux posent les bases générales de la méthodologie à l'œuvre dans les systèmes d'AS décrits ici. Ils constatent en effet la nécessité dans certains contextes applicatifs, notamment la Recherche d'Information, d'extraire les entités nommées relativement aux différents sens qu'elles véhiculent, autrement dit aux entités en tant qu'objets identifiables de façon unique. La constitution de lexiques de variantes lexicales à partir de Wikipedia ainsi que la maximisation de la similarité entre contexte d'occurrence d'une mention et informations collectées pour chaque entité, au travers de sa représentation sous forme d'article Wikipedia, constituent le cœur de la méthode employée dans l'approche de désambiguïsation de Cucerzan [Cuc07] et de Bunescu et Pasca [BP06]. Cucerzan évoque les cas de mentions dénotant une entité absente de la collection d'articles Wikipedia à disposition ; ces cas sont exclus de l'évaluation du système, qui obtient un score d'exactitude de la désambiguïsation (*accuracy*) de 89,85% en moyenne. Cette évaluation porte autrement dit sur le nombre de désambiguïsations correctes parmi l'ensemble des mentions d'entités disposant d'un référent dans Wikipedia. Bunescu et Pasca intègrent explicitement les cas d'absence d'entités en définissant deux opérations principales à réaliser : (i) déterminer si la mention réfère à une entité de Wikipedia ou non et (ii) désambiguïser la mention parmi les différentes entités de Wikipedia qu'elle peut dénoter. L'exactitude (*accuracy*) ainsi mesurée atteint 84,8%.

On peut cependant observer, en AS comme chez Bunescu et Pasca et Cucerzan, que cette formulation du problème en tant que désambiguïsation des mentions d'entités est dans une large mesure relative à l'emploi de ressources déterminées. Le nombre d'instances d'entités présentant

une même variante lexicale est en effet incidentel et dépend du processus de collecte ayant abouti à la constitution de la ressource en question. Une ressource est en effet, quel que soit son type, à considérer comme toujours incomplète, dans la mesure où elle dépend de la mise en œuvre d'un processus d'acquisition, pouvant donner lieu à des erreurs et des lacunes. L'ensemble des entités potentiellement mentionnées dans les contenu traités ne peut, par ailleurs, donner lieu à une collecte exhaustive : de nouvelles entités peuvent en effet émerger dans ces contenus au gré de l'actualité et ne pas faire l'objet d'une intégration immédiate aux ressources ; d'autres entités peuvent être mentionnées sans pour autant bénéficier de la notoriété souvent adoptée comme critère pour une telle intégration. L'association établie entre ensembles de variantes lexicales et représentations d'entités par les systèmes évoqués est donc potentiellement incomplète : il existe d'une part le cas de variantes ne correspondant à aucune entité recensée, et d'autre part le cas de variantes liées à une ou plusieurs entités de la ressource. Dans le premier cas, il ne peut être établi avec certitude que l'entité dénotée est absente : l'ensemble des variantes associées à une entité peut être lacunaire. Dans le second cas, la variante peut, dans un contexte d'occurrence donné, dénoter une entité supplémentaire, non recensée dans la ressource. Il résulte que la non univocité de la dénotation ne relève pas uniquement des cas d'ambiguïté modélisés par une ressource, et doit également s'entendre au niveau des entités et variantes dénotationnelles qui en seraient absentes. C'est pourtant le terme de *désambiguïsation* qui se trouve consacré, dans les différents travaux portant sur les entités et notamment l'AS, au désavantage d'une qualification plus générale du phénomène traité.

Identification Ces observations sur l'approche de la relation entre mentions et entités en termes exclusifs de désambiguïsation amènent à proposer une formulation plus générale et systématique du problème posé par l'établissement d'une telle relation. Le phénomène dénotationnel mettant en jeu une mention et une entité peut donner lieu à une ambiguïté, due à la variation surfacique des mentions et à l'homonymie ; il s'agit cependant d'associer cette mention à l'une des entités formalisées au sein d'une ressource donnée, au-delà de la résolution de l'ambiguïté. L'établissement de cette relation tient, d'une part, à la mise en correspondance entre le niveau textuel et le niveau de représentations formelles, et d'autre part à la couverture de la ressource en termes d'entités potentiellement dénotées au niveau textuel.

Le terme d'*identification* est ainsi proposé pour qualifier de façon générale la tâche effectuée en AS, par laquelle les mentions textuelles d'entité sont associées à une représentation formelle pouvant être qualifiée d'*identité*. L'usage de ce terme vise à qualifier de façon fonctionnelle la tâche en question, ainsi qu'à englober les différents problèmes qu'elle doit traiter, y compris l'ambiguïté. Une systématisation méthodologique correspondant à cet objectif d'identification existe en dehors du paradigme du Web Sémantique, dans le cadre autonome de la tâche de Population de Bases de Connaissances. Cette tâche ainsi que les définitions et méthodes qu'elle propose font l'objet de la section suivante de ce chapitre.

3 Approche systématique de l'identification d'entités

Inscrite dans le paradigme du Web Sémantique et l'objectif d'enrichissement de contenus textuels, l'AS constitue un moyen mise en correspondance entre texte et modèle. La recherche de cette relation entre niveau linguistique et représentation de l'information place l'AS dans une dynamique de renouvellement de l'Extraction d'Information, dont elle se distingue par une formalisation et une standardisation des modèles adoptés. Pour les entités en particulier, l'accès aux connaissances ainsi permis vise des descriptions et faits pouvant être rassemblés et exploités par des traitements automatisés. Les entités constituent en effet dès les travaux initiaux en Extraction d'Information (cf. chapitre 2, section 3) un point d'attention central, auquel le Web Sémantique, les LD et l'AS

apportent un niveau de concrétisation crucial par l'adoption de la représentation ontologique et l'instanciation d'individus, référencés et accessibles de façon systématique. On a néanmoins pu observer que, si l'AS n'ignore pas toujours la prise en charge de la non-univocité existant entre niveau textuel et modélisation, le cœur de ses méthodes et de son fonctionnement cherche avant tout à répondre aux enjeux tenant à la mise en œuvre du Web Sémantique, au développement des LD et aux pratiques de formalisation associées. Il fournit néanmoins au processus d'identification d'entités un composant fonctionnel essentiel en donnant lieu à une vaste production et mise à disposition de données formalisées : à partir de ces données, la sémantique référentielle nécessaire à un traitement des entités en tant qu'individus, c'est-à-dire leur identification automatique, peut être obtenue.

3.1 La Population de Bases de Connaissances et le Liage d'Entités

3.1.1 Population de Bases de Connaissances

L'Extraction d'Information ne constitue pas qu'un prédécesseur aux pratiques récentes du Web Sémantique : elle donne également lieu à un renouvellement interne mené de façon parallèle, dont l'aspect principal consiste en une modélisation de l'information sous forme de *bases de connaissances*. En tant que structure de représentation de l'information, les bases de connaissances (BC) jouent le rôle des formulaires employés dans le cadre des diverses approches historiques d'Extraction d'Information, mais s'apparentent aux structures mises en avant notamment dans le Web Sémantique, c'est-à-dire aux modèles fondés sur une conceptualisation formelle et explicite tels que les ontologies. Les BC constituent une généralisation de ces modèles formels, en tant qu'artefact permettant le regroupement de descriptions, faits et règles propres à un domaine. L'association d'une ontologie en tant que support d'une conceptualisation et d'un ensemble d'individus pouvant être décrits par cette ontologie est ainsi usuellement désigné comme une BC, sans qu'une BC doive nécessairement faire l'objet d'une telle formalisation de façon explicite. Il s'agit ainsi d'un mode de représentation générique dans laquelle un ensemble d'entrées peuvent être enregistrées et associées à des informations dont la structuration est systématisée. Chaque élément d'une BC, de façon similaire à celui d'une ontologie, peut être qualifié de *nœud* .

Dans la perspective d'un traitement de l'information et de sa représentation dans des BC, les campagnes ACE [Dod+04] et TREC [30], consacrées notamment à la Reconnaissance d'Entités Nommées à partir de contenus textuels pour la première (cf. chapitre 2, section 3), et aux systèmes de Question-Réponse pour la seconde, connaissent un renouvellement de leus problématiques dans le cadre de la campagne TAC depuis 2009. Organisée par l'agence américaine NIST [31] autour des recherches en TAL, TAC (*Text Analysis Conference*) présente en effet une tâche consacrée à la Population de Bases de Connaissances (*Knowledge Base Population*, abrégé en anglais en KBP, ci-après PBC), dans laquelle il s'agit d'adapter l'Extraction d'Information classique à la forme de structuration des BC, plus sophistiquée, stable et persistante que les formulaires qui l'ont précédée.

La voie ouverte par la PBC est particulièrement pertinente en regard du problème de l'identification d'entités dans notre cadre de travail : la tâche concerne en effet une BC principalement constituée d'entités et sa population à partir de documents textuels. Les types d'entités considérés correspondent à la restriction usuelle hérité du consensus en Extraction d'Information — personnes, organisations et entités géopolitiques, . Elle atteste ainsi de façon comparable au Web Sémantique de la place essentielle des données textuelles dans le processus d'acquisition de connaissances. La population visée s'entend en termes d'augmentation des connaissances associées aux entités, sous forme d'attributs définis dont il s'agit de donner la valeur en fonction

30. http://trec.nist.gov/
31. http://www.nist.gov/

des informations repérées dans les documents fournis. Cette augmentation peut par ailleurs s'entendre au niveau des entités elles-mêmes, par la découverte de nouvelles entités à partir de ces documents, susceptibles de venir enrichir la population existante.

TAC définit pour la PBC deux sous-tâches fondamentales à réaliser dans cette perspective : *Entity Linking* ou Liage d'entités et *Slot Filling* ou « remplissage de champs ». La première consiste, à partir d'une mention textuelle d'entité au sein d'un document, à identifier parmi les entrées de la BC l'entité à laquelle elle réfère. La seconde sous-tâche procède ensuite à l'extraction d'informations concernant cette entité dans le contexte d'occurrence de la mention, ces informations devant correspondre aux attributs prédéfinis pour chaque entité de la BC — date de naissance pour les personnes ou année de création pour une organisation, par exemple. Il importe donc que la première sous-tâche de Liage identifie de façon univoque l'entrée de la BC concernée par une mention afin que les informations collectées dans son contexte d'occurrence puissent être agrégées au niveau de l'entrée adéquate.

L'identification d'entités trouve ainsi dans le Liage d'Entités une formulation pertinente et utile : le Liage bénéficie dans le cadre de la PBC d'une spécification explicite et motivée des différents problèmes et cas à traiter quant au phénomène dénotationnel, et ce relativement à des ressources d'entités préalablement constituées. On peut observer un développement parallèle caractérisant l'AS et la PBC, avec une distribution des points d'intérêt centraux : il s'agit en AS de donner forme à l'objectif de formalisation des connaissances pour le Web Sémantique, tandis que la PBC donne lieu à une orientation des recherches vers des techniques spécifiques au passage entre niveau textuel et niveau formel, autour des entités en particulier.

Les trois éditions de PBC dans le cadre de TAC (2009, 2010 et 2011) ont donné lieu à des synthèses descriptives, dans lesquelles McNamee et Dang [MD09] ainsi que Ji et al. [Ji+10] et Ji et al. [JGD11] soulignent les enjeux, solutions et problèmes restant à traiter dans cette tâche. L'organisation de cette conférence permet en outre de susciter un nombre important de participations sous forme de différents systèmes, porteurs d'une variété méthodologique et d'innovations utiles pour une approche conséquente de la tâche. Enfin, les différentes éditions de PBC apportent un élément essentiel manquant à l'AS en définissant un cadre et des métriques d'évaluation dédiées à la tâche. Le Liage dans le cadre de la PBC de TAC demeure cependant partiel quant à la prise en charge complète du problème dénotationnel dans des configurations plus réalistes que celles d'une campagne d'évaluation ; ces métriques sont donc limitées au cas spécifique traité par TAC, comme cela sera discuté plus loin.

3.1.2 Le Liage d'entités : problème visé

À la différence de l'approche réservée aux entités dans des cadres d'Extraction d'Information tels que ACE et TREC, TAC vise avec la PBC à rassembler des informations pertinentes relativement à un ensemble d'entités pré-établi. La dimension référentielle des entités y est ainsi explicitement modélisée, par opposition aux tâches de résolution d'anaphore et de coréférence (cf. chapitre 2, section 3.3) où le référent demeure implicite et interne à l'ensemble de documents traités. Ces entités sont identifiées en tant qu'entrées d'une BC, qui définit pour chacune d'elles un ensemble de champs informatifs normalisés. La structure d'une BC dépasse en effet les formulaires traditionnels de l'Extraction d'Information en maintenant les cibles d'extraction en un tout cohérent et persistant, et en liant de façon systématique les sources d'extraction à ces cibles. Il s'agit notamment de prendre en charge les phénomènes de redondance, de complémentarité et de conflit pouvant toucher les informations collectées. Chaque conduite d'un processus d'Extraction d'Information sur un nouveau corpus documentaire peut en effet mener au repérage d'informations sur une entité dont il est utile de déterminer si la BC en dispose déjà, c'est-à-dire si l'attribut correspondant présente déjà une valeur, si cette valeur est identique à la nouvelle proposition ou si une contradiction en émerge. Il peut par exemple s'agir de la date de naissance d'une personne,

à ajouter si elle est non spécifiée, mais à indiquer comme un attribut en conflit si une valeur différente est déjà indiquée ; l'identité de la personne occupant un poste dans une organisation peut en revanche changer au cours du temps, et l'attribut correspondant peut donc voir sa valeur mise à jour au fil des traitements. Ces aspects de maintenance de l'information sont directement liés à l'intérêt de la PBC pour une Extraction d'Information à partir de larges corpus textuels et à la variété informative qu'ils véhiculent.

La sous-tâche de Liage d'Entités (ci-après *Liage*) se définit alors comme l'ancrage des mentions observées en corpus dans un nœud de la BC, avant que toute information relative à l'entité dénotée puisse y être associée. La correction des informations s'entend ainsi en PBC au niveau de la BC elle-même, par contraste avec une erreur en Extraction d'Information traditionnelle qui peut n'affecter qu'un formulaire. La définition du Liage en PBC vise à une prise en charge totale de ce problème d'ancrage en tenant compte du caractère nécessairement non exhaustif de la couverture fournie par la BC : le cas d'impossibilité de Liage est ainsi prévu et modélisé dans la tâche. Comme cela a été mentionné précédemment au sujet des ressources employées en AS (section 2.3), toute ressource présente potentiellement des lacunes quant aux éléments qu'il s'agit de mettre en relation avec des contenus textuels. Dans une BC d'entités, une cible peut être manquante en raison d'un processus de collecte incomplet ou erroné d'une part, ou de l'émergence d'une nouvelle entité dans l'actualité ou un domaine particulier d'autre part. Le Liage doit ainsi tenir compte d'une telle possibilité et un système idoine devra être en mesure de retourner, pour une mention donnée, une réponse vide plutôt qu'un ancrage sur un nœud quelconque de la BC, nécessairement erroné. Cette réponse vide modélise la notion d'entité *inconnue* relativement à la BC et est identifiée en PBC par le terme « NIL ». L'identification des cas NIL permet par ailleurs de fournir des candidats pour une augmentation de la BC en termes d'entités et non plus seulement d'attributs d'entités préexistantes. Bien que cet aspect de la population ne soit pas concrètement en jeu dans la tâche de PBC telle qu'envisagée par TAC (2009 et 2010), cette identification permet néanmoins une délimitation de la couverture de la BC à l'égard du corpus documentaire traité. L'édition de TAC de 2011 oriente le traitement des cas de NIL vers un statut d'entité davantage spécifié, en ajoutant à leur reconnaissance une tâche de *clustering* : les différentes mentions non liées doivent faire l'objet d'un partitionnement, dans lequel chaque partition ou *cluster* doit représenter une référence d'entité et non des chaînes non liées.

De façon générale, la spécification du Liage dans la PBC répond à un impératif de cohérence de l'information concernant les entités, dans la perspective de son exploitation ultérieure. En effet, une agrégation d'informations erronées sur une entité, due à un Liage de mention vers une entité incorrecte — erreur entre plusieurs entrées de la BC ou entre la BC et NIL —, conduit non seulement à une BC inexacte, mais également à la génération d'un bruit se propageant à tout traitement aval. Une indexation par entités pour un système de Recherche d'Information, par exemple, retournerait un ensemble de documents bruités dans le cas de mentions liées à une entité E_1, alors qu'elles réfèrent en réalité à une entité E_2 ou à une entité NIL.

La tâche de PBC et celle du Liage en particulier sont envisagées dans le cadre de TAC pour le traitement de documents en anglais et la BC elle-même provient de ressources en anglais. L'édition de 2011 intègre un composant multilingue avec une seconde sous-tâche de Liage sur des documents en chinois, à partir desquels les mentions d'entités à lier doivent l'être en direction de la BC construite en anglais. Cette extension s'intéresse ainsi au problème de la dénotation interlingue, les entités elles-mêmes ne relevant pas d'une langue en particulier.

3.1.3 Enjeux du Liage

Le problème de l'établissement d'une relation systématique entre texte et représentation est formulé dans le Liage en termes de mentions textuelles d'entités et d'entrées d'une BC, celles-ci

donnant une représentation des entités. Comme en AS et à la différence de l'Extraction d'Information, une telle relation est de nature référentielle et vise des objets qualifiables d'individus, en lieu et place d'un modèle uniquement typologique. La notion d'identification peut être attachée au Liage en raison de la nature de cette relation.

La tâche de Liage intègre de façon systématique les raisons possibles de la non-univocité existant entre mentions (expressions linguistiques) et entités (individus extra-linguistiques) [MD09 ; Ji+10 ; JGD11] :

Ambiguïté Le phénomène dénotationnel à l'œuvre entre mentions et entités est touché par une ambiguïté pouvant avoir deux origines principales :

Synonymie Une même entité peut être désignée par plusieurs expressions linguistiques. Celles-ci peuvent être de différentes natures : noms propres, descriptions définies ou pronoms, qui constituent un premier degré de variation touchant la dénotation. Au niveau des seuls noms propres, auxquels le Liage s'intéresse exclusivement, la variation est également à considérer.
Elle se réalise d'une part au niveau surfacique, comme dans les exemples :

(9) *Hillary Clinton, H. Clinton, H. R. Clinton, Clinton*

(10) *Organisation des Nations Unies, ONU*

où l'abbréviation ou l'acronymie interviennent. Les variations surfaciques sont également pour une large part dues aux phénomènes de translitération et de traduction, lorsque des noms d'entités étrangères sont adaptés à la langue de mention. On trouve ainsi pour une même entité, l'ancien chef de l'État lybien, une centaine de variantes lexicales différentes selon les règles de translitération adoptées, dont :

(11) *Kadhafi, Khadafi, Gaddafi, Kadhaffi, Mu'Ammar El Qathafi, Moammar Qudhafi,* etc..

Certains noms de personnes ou de lieux font l'objet de traductions, de façon plus ou moins systématique : on a en français

(12) *Londres, Moscou*

et non *London* ou *Moskva*, mais

(13) *Istamboul* ou *Istanbul*

ainsi que

(14) *Fiodor, Fédor, Fedor* ou *Théodore Dostoïevski* ou *Dostoiewsky*

où se combinent variation de translitération et traduction [32].
La variation se réalise d'autre part à un niveau qui peut être qualifié d'encyclopédique, dans les cas de surnoms, pseudonymes, changements au cours du temps comme dans ces exemples :

(15) *Ali Hassan al-Majid, Ali le Chimique*

(16) *Paris, la Ville Lumière*

(17) *Prince, Love Symbol, TAFKAP, The artist formerly known as Prince*

(18) *Kate Middleton, Duchesse de Cambridge*

32. Les variations surfaciques peuvent également provenir d'erreurs orthographiques, dont il n'est pas spécifiquement question dans le cadre de TAC-KBP mais qui constituent également un problème dans l'établissement de liens entre mentions et entités. On peut également évoquer le cas des variations dans l'ordre prénom-nom selon les langues, ou des modes de dénomination particuliers concernant notamment les membres de familles royales et princières (*Albert II*, par exemple).

Polysémie Une même expression linguistique peut référer à plusieurs entités et est alors polysémique, de façon incidentelle en cas d'homonymie entre plusieurs entités :

(19) En 1720, la peste frappe *Orange* et y fait 550 victimes.

(20) L'action d'*Orange* perd aujourd'hui 4 points.

Si ces entités entretiennent une relation, la polysémie peut alors relever non d'un caractère incidentel mais d'un phénomène de métonymie, particulièrement saillant entre lieux et organisations. Le nom d'un lieu peut en effet régulièrement être employé pour dénoter une organisation qui y est localisée, qui la représente, etc.

(21) *Barcelone* remporte pour la quatrième fois la Ligue des Champions en 2011.

(22) Parmi les autres marchés européens, *Francfort* a perdu 0,68% et *Londres* 0,03%. [33]

Couverture La relation entre mention et entité par rapport à une BC donnée n'est pas systématique, en raison de l'incomplétude pouvant caractériser cette BC, comme évoqué précédemment.

3.2 Approche générale du Liage

3.2.1 Alignement de mentions et d'entités

La correspondance à établir entre mention et entité en PBC est formulée comme une tâche d'*alignement*, par lequel une mention dans un document issu d'un corpus doit être liée à une entité ; celle-ci est identifiée parmi l'ensemble constitué de l'union des entrées de la BC et de l'entité spéciale NIL. Pour résoudre ce problème d'alignement, le Liage considère l'entité dénotée par une mention comme le *sens* de cette mention. Une mention peut donc avoir plusieurs sens, parmi lesquels peut être déterminé un sens par défaut ; il peut s'agir de l'entité la plus souvent dénotée par une mention donnée, que cette fréquence de dénotation soit définie dans un corpus spéficique ou relativement à la notoriété d'une entité dans l'espace des connaissances générales. La mention *Paris* dénote et évoque par exemple plus fréquemment la capitale française que l'une des villes homonymes situées aux États-Unis. L'alignement systématique d'une mention sur son sens par défaut est cependant une méthode manifestement insuffisante pour la prise en compte des autres dénotations possibles, qui constituent un des enjeux de la tâche de Liage.

L'approche générale du Liage consiste ainsi à supposer que le sens d'une mention, sous la forme d'une entité, peut être retrouvé par application de l'hypothèse distributionnelle de Harris [Har54], selon laquelle des contextes de mentions et d'entités similaires indiquent un même sens. La tâche de Liage intègre donc à la BC un ensemble informatif pour chaque entité, lui tenant lieu de contexte ; pour les mentions, ce contexte est constitué par le document d'occurrence lui-même. La recherche de l'alignement est ainsi vue comme celle de la maximisation d'une proximité sémantique au travers des entités de la BC pour une mention donnée, fondée sur leurs contextes respectifs. Autrement dit, les entités de la BC peuvent être ordonnées selon leur degré de proximité avec la mention considérée. L'entité spéciale NIL doit être ajoutée à l'ensemble des entités de la BC afin que l'alignement ait un résultat dans tous les cas.

L'alignement peut être formalisé de la façon suivante :
Soient

- $m \in M$ une mention à lier dans un document d issu d'un corpus D
- e_m l'entité liée à m par le système
- E l'ensemble des entités constituant des entrées de la BC

33. Exemple extrait du site Web `http://www.la-croix.com` daté du 23 janvier 2012.

- e_{out} l'entité spéciale NIL

on a

- $E = \{e_1, ..., e_n\}$ t.q. $n = |E|$
- $E_{ext} = E \cup \{e_{out}\}$
- $e_m \in E_{ext}$

On définit la fonction d'alignement f :

$$f : M \mapsto E_{ext}$$

comme

$$f(m) = \underset{e \in E_{ext}}{\operatorname{argmax}}\ g(m, e) = e_m$$

où g est une fonction de quantification de la proximité entre une mention m et une entité e, représentées par leurs sens respectifs, eux-mêmes dérivés d'une représentation des contextes correspondants. Les modalités de cette dérivation constituent le focus principal des variations méthodologiques auxquelles la tâche de Liage donne lieu. Le traitement du cas NIL relève également de cette approche en termes de proximité sémantique, en tant qu'il peut être déterminé par le degré ou l'absence de cette proximité. La fonction de quantification g peut par ailleurs être vue comme une fonction de score des entités $e \in E$, dont la fonction f utilise le résultat afin de déterminer e_m.

On a ainsi, pour une mention m à lier :

- pour tout $e \in E$, un score s_m de e tel que $s_m(e) = g(m, e)$
- d'où $f(m) = \operatorname{argmax}_{e \in E_{ext}}\ s_m(e)$

3.2.2 Fonctionnement de la tâche

La tâche de PBC est configurée dans le cadre de TAC selon une orientation propre à l'évaluation et non à la réalité d'une application concrète. Le Liage est à réaliser avec les paramètres suivants :

- Un ensemble de **requêtes** est défini. Chaque requête consiste en un document muni d'un identifiant et d'une mention d'entité dans ce document ; il n'y a donc qu'une mention à aligner par document. Le corpus fourni aux participants, élaboré par le LDC [34] [Sim+10] à partir des données utilisées par ACE, est constitué d'articles journalistiques (environ 1,3 million), de documents issus du Web (environ 500 000) et de quelques centaines transcriptions de documents audio [35]. Ces documents sont datés de 2007 et 2008. La sélection des mentions pour les requêtes d'évaluation correspond à des critères de variation (mentions d'entités présentant un nombre relativement élevé de variantes), de « confusabilité » [Sim+10] (mentions pouvant référer à un nombre relativement élevé d'entités) et de couverture (nombre relativement élevé de mentions référant à des entités absentes de la BC). L'édition de 2009 présente 3904 requêtes (372 noms de personnes, 1 697 noms d'organisations et 160 noms d'entités géopolitiques) ; les éditions de 2010 et 2011 comptent 2250 requêtes (750 noms de personnes, 750 noms d'organisations et 750 noms d'entités géopolitiques).

34. `http://www.ldc.upenn.edu/`

35. Ces données quantitatives sont valables pour les éditions de TAC-KBP de 2010 et 2011. L'édition de 2009 présente environ 1,3 million de documents en majorité journalistiques.

- Une BC également élaborée par le LDC est constituée à partir de l'édition en anglaise de Wikipedia datée d'octobre 2008. Les entités formant les entrées de la BC correspondent aux articles de Wikipedia disposant d'une infobox (cf. section 1.3) correctement formée et dont le type convient pour la tâche [Sim+10]. Cette sélection conduit à la représentation d'environ 818 000 entités, pour lesquelles la BC renseigne :
  - un type parmi PERSON, ORGANIZATION et GPE (pour *geopolitical entity*, entité géopolitique), dérivé du type d'infobox de l'article Wikipedia correspondant ;
  - le titre de l'article correspondant ;
  - l'ensemble des faits dérivés de l'infobox correspondante (ensemble d'attributs et de valeurs associées) ;
  - le texte complet de l'article correspondant ;
  - un identifiant unique interne à la BC.
- Pour chaque requête, les systèmes participant doivent retourner l'identifiant de l'entité liée à la mention considérée, ou NIL.
- L'édition de 2011 introduit une tâche de clustering des réponses NIL, chaque cluster devant représenter une entité unique.

3.2.3 Évaluation

Les éditions de 2009 et 2010 utilisent la métrique de l'exactitude (*accuracy*), calculée à partir du nombre de mentions correctement alignées divisé par le nombre total de requêtes. Cette métrique est désignée par le terme *micro-averaged accuracy* [MD09 ; Dre+10], et donne à chaque requête un poids égal. Une seconde mesure d'exactitude est également calculée après un regroupement des mentions par entité, et correspond au nombre d'entités correctement liées à leurs mentions sur le nombre total d'entités cibles des alignements. Elle est désignée par le terme *macro-averaged accuracy*.

L'édition de 2011 introduit la mesure B-Cubed+ ou B^3+, version modifiée de B-Cubed [BB98] destinée à l'évaluation du clustering dans la tâche de résolution de coréférence. Cette nouvelle métrique met ainsi en relation le problème de l'alignement avec celui de la coréférence : les mentions à lier ne sont plus considérées indépendamment les unes des autres mais en tant qu'elles forment des clusters correspondant à des entités uniques.

Ces métriques se distinguent de la précision et du rappel adoptés en Reconnaissance d'Entités Nommées, dans la mesure où les mentions d'entités sont fournies à la tâche de Liage et ne sont donc pas concernées par un processus de reconnaissance au sein des données textuelles. On pourra observer dans la suite que certains systèmes accomplissent néanmoins une étape de reconnaissance des mentions sur l'ensemble du document afin d'augmenter la requête, qui n'est constituée que d'une mention par document d'après la définition de la tâche ; ces systèmes s'appuient sur ces mentions supplémentaires comme éléments de contextualisation pour l'alignement de la requête.

L'approche générale définie pour le Liage donne lieu à une décomposition méthodologique à partir de laquelle diverses propositions de systèmes ont été faites, dans le cadre de la tâche de PBC de TAC mais également dans des présentations de travaux à l'occasion des principales conférences internationales dédiées au TAL, telles que ceux de Dredze et al. [Dre+10]. On peut citer le cas particulier de Mendes et al. [Men+11b], dont le système Spotlight, présenté précédemment dans le cadre de l'AS, a été soumis par ses auteurs à l'évaluation du Liage proposée par TAC sans adaptation spécifique, à l'exception d'une mise en correspondance entre les instances de DBpedia

et des entrées de la BC. Ce passage de l'AS au Liage témoigne d'une parenté forte entre ces deux cadres de traitement des entités.

3.3 Méthodologie pour le Liage

3.3.1 Décomposition de la tâche de Liage

La typologie méthodologique du Liage fait apparaître un certain nombre de composants relativement communs à tous les systèmes présentés ; ces composants correspondent au fonctionnement de la tâche tel que prescrit par les organisateurs, ainsi qu'au problème central d'alignement entre mentions et entités. Les différents systèmes varient principalement selon les modalités de représentation sémantique de la proximité entre mentions et entités. Ces variations relèvent notamment du caractère purement lexical de cette représentation, ou de l'intégration de facteurs d'ordre structurel, thématiques ou de domaine.

Une décomposition systématique du Liage apparaît dans la synthèse de la tâche proposée en 2010, à l'issue de la deuxième édition de PBC [Ji+10]. D'après les propositions méthodologiques présentées au cours des trois éditions de TAC ainsi que dans les travaux portant sur le Liage publiés par ailleurs, il est possible d'établir une décomposition minimale comptant deux sous-tâches essentielles, ainsi qu'une extension jusqu'à cinq sous-tâches.

Composants minimaux

1. **Génération de candidats** L'espace de recherche pour l'alignement des mentions avec la BC, qui consiste en principe en l'ensemble des entités membres de cette BC, est réduit à un sous-ensemble de taille plus manipulable. Celui-ci comprend les entités de la BC dont il peut être établi qu'elles constituent des **candidats** valides pour l'alignement d'une mention donnée. Cette réduction implique l'introduction d'un critère permettant de sélectionner les candidats, ce critère pouvant être considéré comme un élément de connaissance partiel sur la relation entre mentions textuelles et entités. Dans sa formulation générale et au niveau de cette réduction en particulier, le Liage s'apparente à la tâche de désambiguïsation lexicale [IV98], où il s'agit de déterminer le sens d'un mot dans un contexte d'usage donné. Il est supposé qu'un mot comprend un nombre fini de sens discrets, disponibles sous forme de références dans une ressource telle qu'un dictionnaire ou un thesaurus. Le processus de désambiguïsation consiste alors à associer le mot considéré à l'un de ses sens, relativement au contexte courant. En transposant ce schéma fonctionnel au Liage, on peut procéder à une réduction de l'espace des entités effectivement candidates à l'alignement d'une mention, à condition de disposer d'un moyen de sélection de ces entités traduisant une relation de correspondance sémantique.

La majorité des systèmes de Liage procèdent à la génération de candidats en s'appuyant sur une connaissance *a priori* des différentes variantes lexicales pouvant dénoter chaque entité de la BC. Ces variantes sont alors collectées au préalable à partir de Wikipedia, selon une méthode identique à celle présentée pour la constitution des labels d'entités dans DBpedia, Aleda ou NLGbAse (section 1.3) : le titre de l'article, les liens de redirections, les pages de désambiguïsation ainsi que, dans certains cas, les ancres textuelles de wikilinks correspondant à une même entité sont rassemblées sous forme de dictionnaire de variantes, dont on peut obtenir un index inversé. Pour une mention donnée, un ensemble de candidats possibles est ainsi automatiquement accessible.

La réduction de l'espace des cibles pour l'alignement des mentions répond à un objectif d'efficacité de calcul, puisque l'intégralité de la BC, qui compte plus de 800 000 entrées, se prêterait difficilement à une recherche intégrale. Il faut cependant souligner que cette étape influe potentiellement sur l'accomplissement global de la tâche : elle doit en effet garantir la présence de l'entité effectivement dénotée parmi les candidats. Le processus de sélection doit donc réduire

l'espace de recherche de façon notable tout en évitant le silence. Les systèmes de Bunescu et Pasca [BP06], Cucerzan [Cuc07], Mendes et al. [Men+11b], Zhang et al. [Zha+10], Han et Sun [HS11] ou Ploch [Plo11] utilisent cette méthode. Ji et al. [JGD11] rapportent un taux rappel supérieur à 95% pour la majorité des systèmes.

Un cas notable d'inefficacité de l'établissement *a priori* des correspondances entre variantes et entités serait celui où une entité serait dénotée à l'aide d'une mention nouvelle pour cette entité ou manquée lors de la collecte préalable. Cette mention non associée à l'entité considérée par le processus d'indexation en amont ne pourrait donner lieu à la génération du candidat approprié, qui ne serait donc pas considéré pour son alignement.

L'étape de génération peut par ailleurs donner lieu à un pré-ordonnancement, notamment sur la base d'une probabilité *a priori*. Cette probabilité peut correspondre à un sens par défaut attribué à une mention. Il peut être modélisé par un facteur de « popularité », déduit de l'importance de l'entité dans Wikipedia, notamment à partir de la taille de l'article correspondant [Men+11a ; Hof+11]. Cette probabilité *a priori* peut également être dérivée du nombre d'associations entre une mention donnée et chaque entité qu'elle dénote dans un corpus : dans Wikipedia, les dénotations d'entités par une mention donnée sont identifiables par les wikilinks ; l'entité la plus souvent dénotée par ce biais peut alors être considérée comme le sens par défaut de la mention, comme chez Ratinov et al. [Rat+11].

2. **Ordonnancement des candidats** À partir d'une requête (mention et document), l'ensemble des entités candidates doit être ordonné afin d'obtenir au premier rang l'entité adéquate. Comme évoqué précédemment, l'ordonnancement des candidats est généralement vu comme fonction d'une proximité sémantique quantifiée pour chaque candidat en regard de la mention. Cet ordonnancement est obtenu par l'application de la fonction $g(m, e)$ ou fonction de score $s_m(e)$ à l'ensemble des candidats, dont le candidat obtenant le score maximal est retourné par f (cf. section 3.2.1).

La définition de g se fonde sur une représentation des mentions et des entités dérivée de leurs contextes respectifs. Il s'agit pour les mentions des documents dans lesquelles elles apparaissent, et pour les entités candidates des éléments rassemblés pour chacune d'elles dans la BC, principalement le contenu textuel de l'article Wikipedia leur correspondant. Le cas NIL devant également être une réponse possible à la question de l'alignement d'une mention, un candidat spécial représentant une entité absente de la BC est intégré au processus d'ordonnancement. Il peut ou non faire partie de l'ensemble des candidats généré à l'étape 1 et être manipulé par la méthode d'ordonnancement de façon plus ou moins directe.

Décomposition élargie

1.a Expansion de la requête Avant l'étape de génération de candidats pour une mention donnée, la requête peut faire l'objet d'une **expansion**, c'est-à-dire d'un enrichissement permettant de ne pas limiter les possibilités de génération de candidats à la seule chaîne de caractères de la mention. Ainsi, les mentions consistant en des acronymes, tels que *FMI*, donnent lieu chez Zhang et al. [Zha+11] à une recherche à l'échelle du document permettant d'associer les chaînes étendues, telles que *Fonds monétaire international*, à la requête.

Une requête peut également être enrichie par d'autres mentions de formes différentes, par les procédés de normalisation voire de résolution de coréférence appliqués à l'ensemble des mentions d'un document. Gottipati et Jiang [GJ11] étendent ainsi la requête, pour une mention m de type PERSON ou ORGANIZATION, par la localisation des mentions dont m forme une sous-chaîne (cas des noms de personne apparaissant avec le nom de famille seul et avec prénom et nom de famille, par exemple) ; les requêtes dont la mention est de type LOCATION sont augmentées de toutes les autres

mentions également de type LOCATION (cas de noms de villes et de pays, les seconds étant un élément de contextualisation des premiers, par exemple). Gottipati et Jiang enrichissent également la requête des titres d'articles Wikipedia pour lesquels la mention constitue un lien de redirection ou identiques à la mention. L'expansion de la requête est obtenu par résolution de coréférence sur l'ensemble des mentions du document chez Taylor Cassidy et al. [TC+10]. On peut observer que, seule une mention dans le document étant fournie aux participants en tant que requête, les autres mentions utilisées dans ce procédé d'expansion doivent être obtenues par ailleurs, notamment par l'utilisation d'un système de Reconnaissance d'Entités Nommées, par exemple Stanford NER [FGM05] pour Taylor Cassidy et al.

1.b Génération de candidats Lors de l'étape de génération de candidats (cf. composant minimal 1 *supra*), une requête étendue augmente le nombre d'entités pouvant correspondre à l'entité dénotée par une restriction des sens de la mention. On peut observer que l'expansion de requête dans cet objectif, notamment à partir de mentions coréférentes au sein d'un document, part de la supposition qu'un même terme employé à plusieurs reprises dans un tel espace informatif véhicule nécessairement un sens unique [GCY92].

2.a Ordonnancement des candidats L'alignement est vu comme un problème d'ordonnancement (cf. composant minimal 2 *supra*). Une proximité sémantique est mesurée pour chaque mention et chacun de ses candidats, à partir de leurs contextes respectifs. Le candidat permettant de maximiser cette proximité placé au premier rang de l'ordonnancement ainsi obtenu est retourné. Il est important d'observer que cette formulation se distingue de la configuration classique d'un problème d'ordonnancement ; il s'agit typiquement de la Recherche d'Information, qui retourne n documents pour une requête donnée, suivant un ordre décroissant de pertinence. Pour le Liage, le seul résultat pertinent à l'issue du classement par ordre de proximité est la valeur placée au premier rang. En effet, la notion d'ordre n'est plus discriminante à partir du deuxième rang, puisqu'un seul candidat peut être retenu comme réponse exacte, tous les autres étant considérés comme invalides pour cette réponse — aucun candidat non aligné avec la mention n'est une réponse plus ou moins exacte. Le problème de l'alignement demande ainsi une réponse discrète — une seule entité — tandis que le moyen d'obtention de cette réponse est fondé sur une distribution de valeurs continues.

Les différentes approches méthodologiques proposées autour de l'ordonnancement des candidats constituent autant de définitions de la fonction f introduite précédemment (section 3.2.1), et plus particulièrement de la fonction g utilisée dans f. Elles sont présentées ci-après (section 3.3.2).

2.b Intégration du cas NIL La réponse à la requête peut être une entité issue de la BC, mais également l'entité spéciale NIL représentant une entité absente de la BC. Ce cas est pris en compte de façon directe dans l'ordonnancement, par adjonction de l'entité spéciale à l'ensemble des candidats, ou indirecte, par décision au vu de la réponse retournée par l'ordonnancement. Les différentes approches pour l'intégration du cas NIL sont également présentées ci-après dans le cadre des méthodes d'ordonnancement.

3 Clustering NIL Lors de la dernière édition en date de TAC, la tâche de Liage intègre, en plus de la possibilité de réponse NIL pour une requête donnée, un regroupement des réponses NIL sous forme de clusters représentant des entités uniques. Les requêtes sans correspondance dans la BC sont ainsi également alignées avec une représentation d'entité, même si celle-ci n'est pas identifiée et décrite formellement comme les entrées de la BC, de façon similaire à la tâche de résolution de coréférence. Sans ce clustering, un tel alignement n'est fait que sur une seule entité — NIL, qui représente toute entité hors BC et ne permet donc pas de distinguer les mentions les unes des autres en termes de sens. Pour des traitements ultérieurs, et notamment une augmentation de la

BC, chaque cluster ainsi formé peut se présenter comme une nouvelle entité possible, munie d'une ensemble d'informations, notamment contextuelles, associées à chacune des mentions regroupées en cluster et pouvant assister le processus de création d'une nouvelle entrée. Plusieurs systèmes participant à TAC effectuent ce clustering de façon simple, par correspondance de chaînes : les chaînes de mentions identiques et alignées sur NIL sont regroupées en cluster, par exemple par Taylor Cassidy et al. [TC+10]. Des méthodes de regroupement par paires, hiérarchiques ou par graphes sont rapportées dans [JGD11].

3.3.2 Méthodes d'ordonnancement pour l'alignement

Les différentes propositions méthodologiques formulées autour du problème de l'ordonnancement des candidats pour l'alignement d'une requête de Liage peuvent être étudiées selon la définition donnée à la fonction f, reproduite ici :

$$f(m) = \underset{e \in E_{ext}}{\operatorname{argmax}}\ g(m, e) = e_m$$

Elle se distinguent selon les deux types d'approche suivants, comme le proposent notamment Ji et al. [Ji+10] et McNamee et al. [McN+10] :

1. **Ordonnancement non supervisé** Cette première approche, déjà adoptée par les travaux précurseurs à la tâche de Liage de Bunescu et Pasca [BP06] et Cucerzan [Cuc07], consiste à définir la fonction d'ordonnancement f à l'aide d'une fonction de similarité g calculée selon un modèle vectoriel standard :

- La fonction g prend en arguments les représentations contextuelles respectives de m, la mention et de e, le candidat courant, sous forme de modèles vectoriels. Le document-requête noté d et l'article Wikipedia correspondant à e, noté $e.art$ sont ainsi représentés par deux vecteurs v_d et $v_{e.art}$ de dimension égale à la taille du vocabulaire — celui du corpus Wikipedia, réduit au sous-corpus formé par les $e.art$ de chaque e chez Mendes et al. [Men+11b] et Ratinov et al. [Rat+11].
- Chaque élément de v_d et $v_{e.art}$ est un poids associé au mot d'indice i étant donné d et $e.art$. Ce poids correspond à la mesure TFIDF [BP06 ; Cuc07], modifiée en TFICF chez [Men+11a ; Rat+11] (cf. section 2.2) pour une prise en compte du pouvoir discriminant d'un mot donné relativement à un candidat particulier.
- La fonction g est définie comme une mesure de similarité cosinus chez Mendes et al. [Men+11a] ; elle peut être combinée chez Bunescu et Pasca [BP06] à l'apprentissage d'une corrélation entre mots et catégories d'articles Wikipedia, donnant alors lieu à une fonction de score linéaire dont les paramètres sont dérivés de cette corrélation ainsi que de la similarité cosinus.

L'ordonnancement non supervisé est ainsi fondé sur l'usage des contextes de mentions et entités sans étiquetage, à l'exception de la configuration avec fonction linéaire de score de Bunescu et Pasca [BP06], et une fonction de similarité retournant de façon directe un score pour chaque candidat ou intégrée en tant que paramètre d'une fonction de score.

Dans cette approche, le candidat spécial NIL peut donner lieu à différents traitements. Chez Bunescu et Pasca [BP06], dans la configuration vectorielle standard, un seuil minimal de similarité est défini : si aucun candidat n'obtient un score supérieur à ce seuil, NIL est retourné. Dans la configuration étendue aux catégories, la fonction de score intègre un paramètre supplémentaire correspondant à ce seuil, dont le poids est appris avec ceux des autres paramètres utilisés sur le corpus d'entraînement fourni pour l'évaluation.

La mesure de similarité employée dans l'approche non supervisée peut intégrer des éléments non uniquement lexicaux et ainsi distinguer le contexte d'un simple sac de mots. Ces éléments, qualifiés de sémantiques notamment par Han et Zhao [HZ10], relèvent de la similarité thématique existant entre les contextes considérés. La distribution des mentions d'entités au sein de chacun d'eux peut également jouer le rôle de contexte : les co-occurrences d'entités sont déterminées à partir du corpus Wikipedia ; dans un document-requête, les différentes mentions, qu'il faut alors repérer à l'aide d'un système de Reconnaissance d'Entités Nommées, peuvent refléter de façon plus ou moins similaire ces co-occurrences par rapport au candidat courant. Ce type d'éléments contextuels est intégré par Han et Zhao dans une mesure de similarité sophistiquée utilisée pour ordonner les candidats.

2. **Apprentissage supervisé** En termes de classification classique, chaque paire (m, e) peut se voir assigner par un classifieur un label parmi $\{1, 0\}$, selon que e est dénotée par m ou non. Ainsi formulé, l'apprentissage supervisé se présente ainsi :

- Soient une mention m et un ensemble de candidats $C = \{c_1, ...c_n\}$ et $C \subseteq E_{ext}$ où n est le nombre de candidats générés par la requête parmi les entrées de la BC ; on peut avoir NIL $\in C$ ou NIL $\notin C$.
- Pour une paire (m, c_i), on définit un vecteur

  $$v_i = \phi(m, c_i) \in \mathbb{R}^d$$

  où d est le nombre de traits considérés, avec

  $$\phi(m, c_i) = \Big(\phi_1(m, c_i), \phi_2(m, c_i), ..., \phi_d(m, c_i)\Big)$$

- Pour une mention m, on calcule ainsi la séquence $V = [v_1, ..., v_n]$ de vecteurs de traits, avec un vecteur par candidat c.
- On cherche à apprendre une fonction h telle que

  $$h : M \times E_{ext} \mapsto \{0, 1\} \text{ et } \quad h(m, c) = \begin{cases} 1 & \text{si } c = e_m \\ 0 & \text{sinon} \end{cases}$$

  à l'aide de ϕ telle que

  $$h(m, c) = h_\phi\Big(\phi(m, c)\Big)$$

- On génère les exemples d'entraînement pour h' à partir des éléments de V associés à une classe dans $\{0, 1\}$, indiquée par les données de référence de la tâche :

  $$h'(v_i) = \begin{cases} 1 & \text{si } c = e_m \\ 0 & \text{sinon} \end{cases}$$

- On a donc pour chaque mention un ensemble d'exemples dont un seul est étiqueté avec la classe 1, la classe 0 étant attribuée à tous les autres.

Lors de la prédiction, plusieurs paires pour une même mention m peuvent alors recevoir la classe positive, ce qui contredit l'unicité inhérente de la réponse à apporter au problème de l'alignement, comme le soulignent notamment Zheng et al. [Zhe+10]. Le problème de l'alignement en termes d'ordonnancement peut alors être pris en charge par des méthodes d'apprentissage supervisé adaptées à l'ordonnancement. Ces méthodes, regroupées sous le terme *learning to rank*, font l'objet d'une présentation extensive par Li [Lil1].

Avec une intégration explicite de l'ordonnancement dans l'apprentissage supervisé, le Liage se présente ainsi comme une tâche dont les objets sont, pour chaque cas sujet à une prédiction, une mention et un ensemble de candidats. Les paires formées par la mention et chacun de ces candidats sont manipulés sous la forme d'une liste dont il s'agit de retourner l'élément placé au premier rang à la suite de l'ordonnancement de cette liste, comme l'exprime la fonction

$$f(m) = \underset{e \in E_{ext}}{\operatorname{argmax}}\ g(m, e) = e_m$$

introduite précédemment. Trois types de méthodes sont envisageables pour l'adaptation de la classification à l'ordonnancement :

Point-à-point (pointwise) Pour chaque instance d'un problème donné, un classifieur stochastique binaire retourne un nombre réel dont la classe prédite, positive ou négative, peut être dérivée. Ce nombre réel consiste ainsi en un score d'appartenance à une classe assigné à l'instance. L'apprentissage de l'ordonnancement selon la méthode point à point considère ce score indépendamment de la notion de classe et l'emploie pour définir un ordre sur un ensemble d'instances. Dans le cas du Liage et de paires (m, c) pour une même mention, les candidats sont alors ordonnés en fonction de ce score. Un exemple atomique fourni à l'algorithme d'apprentissage consiste dans ce cas en un vecteur de traits représentant une paire (m, c). Il y a donc autant d'exemples d'apprentissage par mention que de candidats à l'alignement, un seul de ces exemples étant étiquetés avec le label de classe positive. Lors de la prédiction, chaque paire reçoit un score attribué par le classifieur et l'alignement est réalisé par regroupement et ordonnancement des paires pour une même mention m.

Par paires (pairwise) L'ordonnancement par paires considère pour une mention m et ses candidats c tout ou partie des paires $\big((m, c_1), (m, c_2)\big)$, impliquant deux candidats distincts. Pour chacune de ces paires, un classifieur prédit une classe, positive si (m, c_1) est jugé plus probable que (m, c_2), 0 sinon. À partir de ces décisions locales, on construit un ordre total sur l'ensemble des (m, c).

Par liste (listwise) Cette méthode se distingue davantage de la classification et manipule directement les exemples d'entraînement sous forme de liste ordonnée. Dans une tâche de Rercherche d'Information, cet ordonnancement correspond au degré de pertinence d'un document parmi un ensemble de documents retournés pour une requête donnée. Dans le cas du Liage, un exemple atomique est constitué de l'ensemble des paires (m, c) pour une mention donnée. L'apprentissage ainsi formulé se distingue cependant de la configuration d'ordonnancement plus usuelle de la Recherche d'information : la paire (m, c) présentant le candidat correct reçoit le label de rang 1, les autres le label de rang 2. L'ordre n'est en effet pas défini sur les candidats non retenus au rang 1, qui sont tous vus comme également incorrects pour l'alignement, comme le soulignent Zheng et al. [Zhe+10]. Plusieurs algorithmes d'apprentissage tels que ListNet[Cao+07] ou SVMRank[Joa06], sont disponibles pour l'ordonnancement par liste et sont notamment utilisés dans la tâche de Liage par Li et al. [Li+09], Dredze et al. [Dre+10] ou Zheng et al. [Zhe+10].

L'ordonnancement par apprentissage supervisé présente ainsi, pour une requête donnée, un ensemble de paires, chacune correspondant à la mention de la requête et à un de ses candidats. On a donc autant de paires (m, e) que de candidats générés pour une requête donnée. Une paire est modélisée sous la forme d'un vecteur de traits dont chacun représente un élément d'information sur la proximité sémantique entre mention et candidat.

Si l'apprentissage de l'ordonnancement est largement adopté par les participants de TAC-KBP et paraît en effet plus approprié au Liage que la classification, on peut constater que cette

représentation du problème demeure partielle. Comme on l'observe avec l'apprentissage par liste (*listwise*), la notion d'ordre sur l'ensemble des candidats d'une même mention ne reflète pas en termes exacts la solution recherchée. Il s'agit en effet de retourner *un* meilleur candidat, l'ordre à établir concernant donc ce candidat *versus* tous les autres. Il serait par ailleurs malaisé de déterminer les conditions de constitution manuelle des exemples d'entraînement selon un ordre sur l'ensemble des candidats — ce qui conduit, dans la configuration par liste chez Zheng et al. [Zhe+10], à attribuer le rang 2 à tous les candidats sauf le candidat correct pour une mention. On rejoint alors une vue du problème où une solution unique est recherchée.

Les traits de ϕ sont dérivés différemment selon les approches adoptées, en fonction du type d'information collectées à partir des contextes de mentions et d'entités. Il peut s'agir d'informations lexicales (contexte vu comme un sac de mots), thématiques (prise en compte des catégories assignées aux documents, inférence de thèmes à partir du lexique ou *topic model*), ou concernant la distribution des entités elles-mêmes et leurs co-occurrences. De nombreux traits sont généralement utilisés et comprennent dans la plupart des cas des mesures de similarité comparables à celles utilisées de façon plus directe dans les approches non supervisées. Il peut s'agir de la similarité cosinus et donc lexicale calculée à partir du modèle vectoriel évoqué précédemment, de la similarité thématique calculée à partir des catégories de documents ou de thèmes inférés. La similarité surfacique entre mention et titre de l'article correspondant à l'entité peut également être prise en compte, ainsi que le type associé aux entités de la BC, qui peut être comparé à un type assigné à une mention par un module de Reconnaissance d'Entités nommées. Une présentation synthétique des types de traits utilisés par les approches supervisées est donnée dans la table 3.10. Les systèmes de Liage reposant sur un apprentissage supervisé à l'aide de traits relevant de la similarité contextuelle lexicale et du type sont notamment ceux de Dredze et al. [Dre+10] ou Li et al. [Li+09]. L'intégration d'informations sémantiques, c'est-à-dire non limitées à la surface lexicale des contextes de mentions et d'entités, est notamment proposée par Ploch [Plo11] : les relations entre entités sont intégrées à ce modèle par le biais de leurs co-occurrences observées dans Wikipedia à l'aide des wikilinks. Le système de Pilz et Paaß [PP09] s'appuie sur une dérivation thématique à partir des contextes lexicaux, celui de Kozareva et Ravi [KR11] sur la modélisation thématique par allocation de Dirichlet latente.

La popularité des candidats, telle qu'évoquée au niveau du composant minimal 1, est employée comme trait par certains systèmes, notamment Dredze et al. [Dre+10]. Si cette caractéristique permet à elle seule de déterminer un alignement correct dans de nombreux cas — 71% d'après l'étude de Ji et al. [Ji+10] —, elle constitue une information statique, ne tenant pas compte du contexte et dont la valeur dépend du corpus traité. Son introduction comme paramètre de la fonction de score permet de l'intégrer parmi d'autres critères de décision.

Le traitement du cas NIL dans les approches supervisées peut être intégré à la fonction de score dérivée du classifieur. Le cas NIL est alors considéré comme un candidat parmi les autres, avec les mêmes traits ou par l'addition de traits spécifiques, comme chez Dredze et al. [Dre+10] ou Han et Sun [HS11]. Alternativement, un classifieur spécial peut être entraîné afin de valider, après l'ordonnancement, le candidat retourné au premier rang, comme chez Li et al. [Li+09] et Zheng et al. [Zhe+10].

3. **Graphes** Plusieurs systèmes intègrent des contraintes de cohérence globale à l'alignement. Il s'agit d'approches où sont considérées toutes les mentions présentes dans un document, et non la mention unique de la requête. L'hypothèse formulée est celle d'une cohérence au niveau des entités co-occurrentes dans un même document ; il est attendu que l'alignement d'un ensemble de mentions doit résulter en un ensemble d'entités entretenant des relations sémantiques établies par ailleurs. Ces relations peuvent être représentées par des graphes dont les nœuds sont les entités

et les arcs des liens entrants et sortants dans Wikipedia. Dans des approches telles que celle de Han et al. [HSZ11], les possibilités d'alignement sont considérées simultanément et globalement sur l'ensemble des mentions ; elles sont pondérées [Rat+11] ou contraintes [Hof+11] afin de favoriser l'ensemble d'alignement le plus probable relativement aux relations entre entités, généralement à l'aide d'algorithmes d'optimisation complexes.

Catégorie de trait	**Type**	**Description**
Relation (m, c)	Similarité surfacique	correspondance exacte, acronymie, abréviation, distance d'édition, inclusion, composants de noms
	Fréquence d'emploi	Fréquence de m comme dénotation de c dans les wikilinks de Wikipedia
	Type	types (m, c) identiques
Similarité de documents	Lexicale	modèle vectoriel et similarité entre d et art
	Thématique	dérivation thématique à partir du lexique, catégories de d et de art
Similarité étendue	Relations	entités co-occurrentes dans Wikipedia vs. mentions co-occurrentes dans d
Popularité (c)	Web	premier article Wikipedia pour une requête avec m
	Wikipedia	nombre d'occurrences dans Wikipedia, taille de l'article

TABLE 3.10 : Présentation synthétique des types de traits utilisés dans les approches supervisées d'ordonnancement pour le Liage (adapté de [JGD11]). m : mention ; c : candidat ; d : document-requête ; art : article Wikipedia pour un candidat.

Le Liage dans le cadre de la PBC a fait depuis 2009 l'objet d'un grand nombre de propositions de systèmes et de méthodes, principalement caractérisées par l'introduction régulière de nouveaux paramètres au niveau de la fonction de score adoptée. L'accomplissement typique de la tâche en plusieurs étapes, dont la génération de candidats précédant leur ordonnancement, est attesté dans tous les travaux à son sujet. Les systèmes reposant sur un apprentissage supervisés sont majoritaires et intègrent souvent des informations non uniquement lexicales dans le calcul de la fonction de score.

La synthèse proposée à l'issue de l'édition de PBC à TAC en 2011 observe que plusieurs systèmes — dont ceux de Chen et Ji [CJ11] ou Cucerzan [Cuc11] — procèdent à la reconnaissance de l'ensemble des mentions d'entités des documents donnés en requête, afin d'en réaliser un Liage collectif vers la BC, tenant ainsi compte du facteur de co-occurrence pouvant affecter l'alignement de chaque mention. Une telle approche dépasse *de facto* la configuration d'évaluation définie par la tâche de Liage dans le cadre de TAC, qui ne prévoit qu'une seule mention par document pour la constitution des requêtes d'alignement sur la BC. Cette configuration est cependant éloignée de la réalité d'un système de Liage intervenant dans un contexte applicatif concret, où toute mention du document est potentiellement à lier. Un repérage des mentions généralisé au document est par ailleurs comparable à la situation de l'AS, également concernée par l'identification des entités.

La configuration du Liage dans TAC opère ainsi une séparation nette entre l'alignement de mentions et la procédure par laquelle elles sont obtenues. Comme en AS, une mention ne peut

donner lieu à un alignement sur une BC que si elle est au préalable repérée comme telle : cette nécessité n'est pas abordée par les travaux relatifs à la PBC puisque celle-ci considère les mentions comme données. Or, une configuration réaliste où ces mentions sont obtenues à l'aide d'un module automatique de Reconnaissance d'Entités Nommées ne peut garantir une correction totale de ses résultats. En cas de faux positif retourné par la reconnaissance, le Liage effectué donne ainsi lieu à un alignement factice — de façon comparable à une non prise en compte du cas spécial NIL, comme évoqué précédemment (section 3.1.2). Le Liage tel que défini par TAC ne tient pas compte de cette possible propagation d'erreurs, sur laquelle un système dépassant la configuration d'évaluation devrait porter son attention. Dans la perspective de traitements complets, partant de contenus textuels bruts et visant un enrichissement en métadonnées autour de l'identification des entités, le Liage se présente néanmoins comme une étape d'analyse bien définie et indispensable. Il peut ainsi intervenir dans de tels traitements, notamment en association avec des modules d'analyse spécialisés dans la Reconnaissance d'Entités nommées.

Chapitre 4

Expression de besoins : enrichissement de contenus textuels pour l'AFP

En tant qu'acteur majeur de la diffusion de l'information, l'AFP est concernée par des besoins particuliers de structuration et d'accès à l'information, dont l'enrichissement de contenus textuels forme un objectif central. La production de l'AFP, en majorité textuelle — les données multimedia présentant eux-mêmes des composantes sous forme de texte —, est en effet caractérisée par une quantité massive et quotidiennement augmentée d'environ 5 000 dépêches, 3 000 documents photographiques légendés, 150 documents video munis de transcriptions et d'une centaine de documents infographiques. Ce large périmètre de données constitue le centre des besoins formulés par l'AFP en termes de gestion de l'information, de systématisation des traitements et de cohérence globale. L'Annotation Sémantique présentée précédemment constitue la réalisation concrète de tels objectifs en tant que moyen d'obtention des métadonnées d'enrichissement et d'apport d'une sémantique formellement définie, en particulier pour les entités qui forment l'ensemble référentiel visé.

Le présent travail consiste ainsi, après avoir établi les modalités de traitement de l'information et la nature des ressources sémantiques définies par le paradigme du Web Sémantique, à proposer une méthodologie de principe ainsi qu'un système adaptés à l'identification des entités pour la production de métadonnées. Il semble utile de donner au préalable un cadre concret à l'intégration d'objectifs, de pratiques et de technologies du Web Sémantique dans des contextes applicatifs similaires, au travers notamment de deux cas d'utilisation : l'orientation de la gestion de contenus de la BBC constitue une application concrète de technologies du Web Sémantique au processus rédactionnel journalistique, menant à une mise en relation dynamique des contenus produits à l'intention des utilisateurs publics, notamment autour de DBpedia ; les activités historiques et récentes du *New York Times* illustrent les modalités et résultats d'un effort d'indexation des contenus et d'intégration aux Linked Data, élément essentiel du Web Sémantique. Une présentation de l'AFP et des éléments de traitement existants pourra ensuite être proposée afin de spécifier le cadre dans lequel une méthodologie de systématisation de ces traitements peut être envisagée. Les objectifs de l'AFP quant à l'enrichissement de contenus pourront ensuite être mis en perpective autour des points de renouvellement à considérer dans les traitements actuels ainsi que des différentes contraintes posées par l'environnement particulier de l'Agence.

1 Cas d'utilisation dans la presse numérique

Le paradigme de renouvellement de la production documentaire proposé par le Web Sémantique et dans lequel s'inscrit le besoin d'enrichissement formulé par l'AFP constitue d'ores et déjà le cadre de développement de projets à caractère industriel dans le domaine de la presse. L'orien-

tation numérique de la production documentaire en général et de la diffusion de l'information en particulier place en effet les acteurs de la presse au premier rang quant à l'adoption de ce paradigme.

Il est important de noter que l'adoption de ces technologies qualifiées de sémantiques par des organisations telles que celles présentées ici s'inscrit dans une continuité d'un traitement des données visant à leur structuration et à leur distribution dans la chaîne de production. Les méthodes existantes correspondent historiquement à des pratiques de *data curation*, autrement dit de conservation et d'édition des données : il s'agit principalement de vocabulaires contrôlés et de processus d'indexation mis en œuvre dans l'objectif d'une gestion de la production centralisée et cohérente. Ce paradigme classique est cependant caractérisé par un phénomène de dispersion et un manque de formalisation : une organisation de taille importante connaît en effet presque systématiquement un éclatement de ses processus de production, où chaque service, département ou direction, prenant en charge une tâche et des données particulières, adopte des schémas de structuration et des vocabulaires contrôlés différents, empêchant un partage de l'information à l'échelle de l'organisation et *a fortiori* externe. Le renouvellement proposé par l'architecture du Web et ses technologies sémantiques trouve précisément dans ce besoin d'unification son point d'influence, en permettant une intégration de l'information immédiate et définie pour l'ensemble d'une organisation donnée, favorisant ainsi le développement d'opportunités d'ordre à la fois fonctionnel et commercial.

L'orientation de ces cas d'utilisation relativement au Web Sémantique peut varier autour des objectifs visés : La BBC place au centre de ses développements la mise en relation et la publication dynamique de contenus sur la base de métadonnées sémantiques, notamment ancrées dans DBpedia, afin de proposer une plus grande richesse de navigation aux utilisateurs de son site Web. Pour le *New York Times*, les pratiques historiques d'indexation mise à disposition du public trouvent dans le réseau des Linked Data un mode de réalisation renouvelé.

La mise en œuvre de tels cas d'utilisation s'illustre notamment par le mode d'intégration des technologies dites sémantiques, c'est-à-dire relevant de pratiques définies par le Web Sémantique, aux outils fonctionnels de production et de gestion de l'information. Ces cas d'utilisation concrets, marqués par un stade de développement avancé, illustrent la façon dont se construisent et se présentent les réponses pratiques au renouvellement de la publication et de la gestion documentaire. Il s'agit de déterminer la place attribuée au modèle adopté, le degré d'automatisation des traitements envisagés ainsi que le type de techniques employées dans les systèmes de gestion de contenus ou CMS (*Content Management System*).

1.1 La BBC et DBpedia : mise en relation dynamique de contenus

La BBC présente par son site Web un exemple applicatif concret et notable du Web Sémantique et de ses technologies. La refonte de ce site, initiée à l'occasion de la couverture de la Coupe du monde de football de 2010 et présentée dans [MK12] et [Kob+09], est menée dans l'objectif principal d'améliorer l'exploration des contenus par les utilisateurs par des modalités de navigation enrichies. Il s'agit de mettre en relation les documents produits à travers les différents domaines traités par la BBC par l'usage de techniques et de ressources émanant du Web Sémantique. DBpedia est ainsi adopté comme vocabulaire contrôlé dans la perspective d'une formalisation sémantique de référence des données manipulées.

Le renouvellement du processus de maintenance et d'édition des contenus de la BBC concerne notamment les processus d'indexation à l'œuvre dans l'édition des contenus. Avant ce renouvellement, chaque domaine repose sur un index, rassemblant des données référentielles maintenues manuellement et en permanence. Plusieurs index peuvent présenter des métadonnées correspondant aux mêmes entités, sans que celles-ci ne présentent de référence commune. L'accès aux contenus est alors statique, défini par les seules métadonnées gérées par l'index de référence.

Il est ainsi impossible, par exemple, d'accéder à des articles mentionnant un acteur dans le domaine de la politique à partir de documents appartenant au domaine du cinéma. Cette organisation des ressources sans interconnexions constitue donc un frein majeur à une exploration des contenus cohérente et sophistiquée. On retrouve ici une caractéristique usuelle des données d'entreprise, abordées au chapitre 1 (section 3), où l'emploi d'un vocabulaire contrôlé pour la gestion des données se heurte à des pratiques de maintenance séparées et à l'absence de schéma de représentation unifié.

L'adoption d'un modèle de représentation unifié pour l'indexation des contenus constitue donc le cœur de l'approche proposée par la BBC pour son site Web. À partir d'un tel modèle, dont DBpedia se fait ici le pivot central, une indexation des contenus selon une sémantique définie peut être envisagée, permettant ainsi une mise en relation des documents à travers les différents domaines traités. Les points d'accès aux contenus sont alors vus comme multiples et ouvrant la navigation sur l'ensemble de la production sans limitation à un domaine particulier. L'ancrage des métadonnées employées pour l'indexation dans un modèle sémantique tel que DBpedia permet par ailleurs de prendre en compte les évolutions potentielles touchant les entités ainsi référencées : l'attribution à un document d'une métadonnée correspondant à une entité particulière n'est plus figée dans un état de maintenance de l'index mais au modèle de représentation définissant cette entité.

La mise en œuvre de cette modélisation unifiée ainsi que des processus d'indexation renouvelée qu'elle permet prend à la BBC la forme concrète suivante :

Modèle DBpedia est adopté comme pivot pour la représentation des données en raison de sa large couverture et de sa place centrale dans le réseau des Linked Data, ainsi que de la persistence des références disponibles due à sa disponibilité sur le Web, contrairement à des ressources propriétaires et fermées. Les données d'indexation existantes — les index préalablement utilisés par la BBC — sont intégrées au nouveau modèle par le biais d'ontologies de domaines (*Sport* par exemple) dont les instances sont mises en correspondance avec leurs équivalents dans DBpedia. Le modèle ainsi obtenu est donc associé aux Linked Data. Deux ontologies fonctionnelles sont associées à ce modèle (cf. figure 4.1) afin de représenter les informations d'association entre métadonnées (*Tagging Ontology*) et documents (*Asset Ontology*) d'une part, et entre métadonnées et domaines (*Domain Ontologies*) d'autre part.

Traitement des contenus Les contenus sont annotés à l'aide du système de Reconnaissance d'Entités Nommées de la chaîne d'Extraction d'Information GATE [Cun+11b]. Les mentions ainsi obtenues donennet lieu à la sélection des possibles instances correspondantes dans DBpedia, puis une phase de désambiguïsation s'appuyant sur le contexte d'occurrence retient l'une d'elle pour constituer une métadonnée du document traité. L'URI fournie par DBpedia pour l'instance est ajoutée à ce document au niveau de la mention d'entité identifiée. Le format rNews, élaboré par le consortium IPTC[1], permet l'intégration de métadonnées de contenus au format HTML et est à l'étude pour une intégration à la BBC.

Les journalistes appellent ce traitement sur les documents rédigés à partir du CMS mis à leur disposition (cf. figures 4.2 et 4.3), puis procèdent à une étape de validation manuelle des annotations proposées : chaque lien retourné par l'analyse peut être accepté, refusé ou corrigé avant que le document ne soit transmis pour diffusion. En cas d'ambiguïté entre plusieurs instances de DBpedia pour une même mention, les journalistes sont amenés à choisir l'instance adéquate. Les métadonnées ainsi ajoutées aux documents sont par ailleurs stockées au format RDF et reliées aux ontologies prévues à cet effet (cf. figure 4.1). Le traitement sémantique des contenus est ainsi largement automatisé et ne nécessite

1. `http://www.iptc.org`

plus de maintenance coûteuse et permanente des index de métadonnées, tout en assurant des fonctionnalités de contrôle et de validation par les journalistes, nécessaires à l'édition appropriée des contenus.

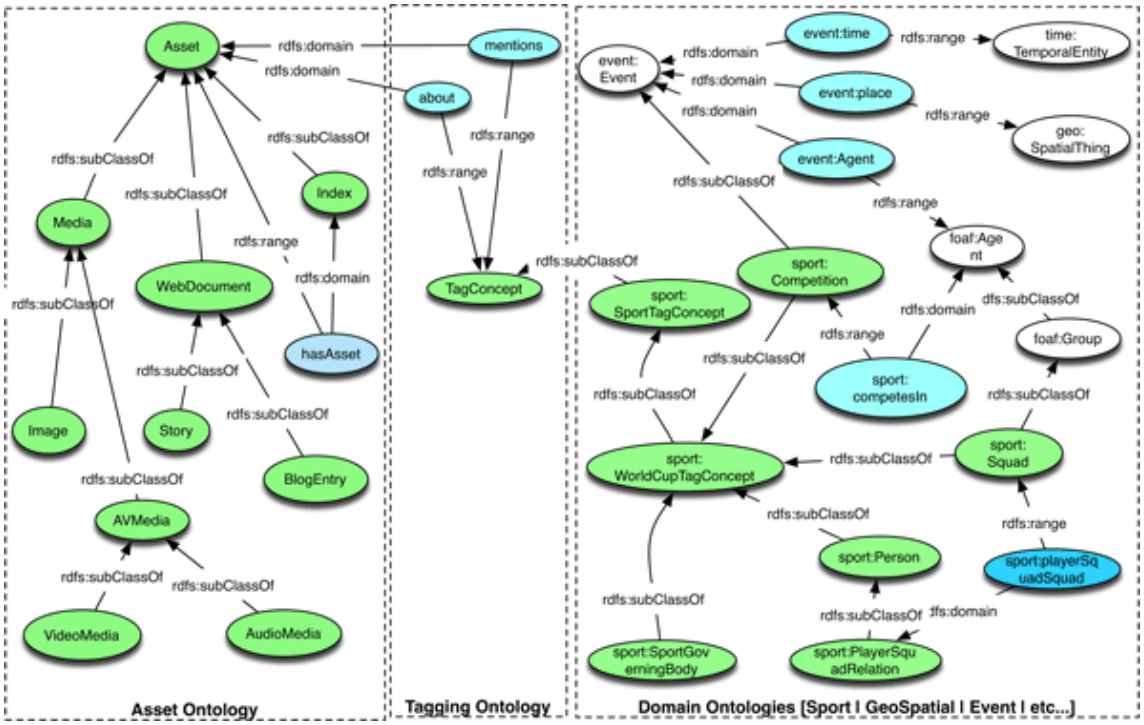


FIGURE 4.1 : Ontologies fonctionnelles et de domaine de la BBC (reproduit à partir de [MK12]).

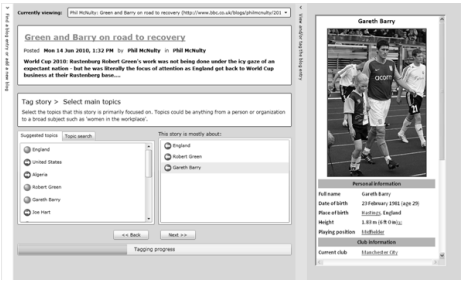


FIGURE 4.2 : CMS de la BBC : Indexation sur l'entité *Gareth Barry* (domaine *sport*) d'une dépêche concernant le footballeur (reproduit à partir de [MK12])..

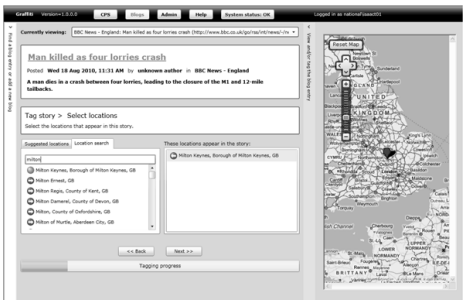


FIGURE 4.3 : CMS de la BBC : Annotation d'une dépêche avec le lieu *Milton Keynes* (reproduit à partir de [MK12]).

Le résultat du traitement sémantique mis en place pour le site Web de la BBC consiste en une évolution de la mise à disposition des contenus d'un état non relié à une interconnexion complète à l'aide d'un modèle unique et simple, reposant sur DBpedia. La publication de documents prend alors une orientation dynamique et non plus statique, en passant d'une gestion figée des métadonnées à un modèle sémantique clairement défini et accessible, dont les données peuvent changer au cours du temps sans que les liens entre domaines et documents n'en soient affectés.

Au-delà d'une navigation à travers les contenus à partir de métadonnées sémantiques, principalement les entités mentionnées dans les documents de la BBC, ce renouvellement des pratiques d'indexation et de publication permet le développement d'applications reposant sur le réemploi des documents, dépassant leur publication originale. En effet, de nouveaux contenus peuvent être dynamiquement créés à partir des documents enrichis en métadonnées, notamment par agrégation autour d'un ensemble de métadonnées spécifiées. La figure 4.4 illustre le résultat d'une telle agrégation dynamique d'information autour de l'entité *Chelsea FC*, telle qu'elle est présentée aux utilisateurs du site Web de la BBC.

Figure 4.4 : Page dynamique du domaine *sport* de la BBC pour l'équipe Chelsea FC : agrégation automatique de métadonnées, statistiques sportives actualisées et navigation dynamique dans le domaine *Sport* (reproduit à partir de [MK12]).

1.2 Le *New York Times* : pratique historique de l'indexation et intégration aux Linked Data

Le *New York Times* (NYT) présente une tradition d'édition et d'indexation de ses contenus âgée d'un siècle et le menant dans la période contemporaine à compter parmi les premiers ensembles de données intégrés au réseau des Linked Data. La place du NYT dans les développements de la presse numérique, illustrée par son rang de site Web de presse [2] le plus populaire aux États-Unis [Woo10], est ainsi le reflet d'une migration naturelle d'une pratique ancrée dans son histoire vers un renouvellement dans le cadre du Web Sémantique.

1.2.1 *Data Curation* au NYT : historique

Le processus de *data curation* au NYT débute dès 1913 par une volonté d'enrichissement de la production journalistique déjà perçue comme un facteur de valorisation face à ses concurrents. Cet enrichissement se traduit par la création du *New York Times Index*, sous la forme d'un catalogue recensant tires et résumés d'articles, publiés périodiquement et catégorisés par thèmes

2. `http://www.nytimes.com`

et noms propres. Les contenus du NYT se présentent dès lors non seulement comme un véhicule d'information classique mais également comme une source majeure d'archives historiques, ouvertes à la recherche et servant de référence dans divers débats d'idées [Woo10].

La création et la maintenance de ce catalogue sont mises en œuvre par un service dédié de 15 personnes, le « Index Department » (ID) [Woo10], qui inaugure un processus systématique de *data curation* sur l'ensemble de la production du NYT, notamment à partir d'un index jusqu'alors utilisé uniquement de façon interne depuis 1851. Le catalogue se développe alors autour d'un vocabulaire contrôlé couvrant un ensemble de thèmes et d'entités (noms de personnes, d'organisations, de lieux ainsi que d'œuvres telles que des livres, films, etc.), liés aux articles les mentionnant.

1.2.2 Processus d'indexation des contenus

Deux facteurs principaux déterminent les modalités d'évolution du processus d'indexation des contenus initié par le NYT au début du siècle : des changements onomastiques et terminologiques touchent nécessairement les entités et sujets de catégorisation au cours du temps, et *a fortiori* sur une telle période ; ils induisent également une quantité considérable de termes d'indexation, atteignant les centaines de milliers et donnant lieu à une complexité de traitement nécessitant un renouvellement de l'approche historique. D'autre part, l'importance croissante du Web comme medium de diffusion s'accommode difficilement des temps de traitement propres à l'ID, de l'ordre de plusieurs jours pour l'indexation des contenus, quand l'édition en ligne du journal s'inscrit dans des standards en temps réel.

La prise en charge des évolutions nécessaires du processus d'indexation se traduit au NYT par une chaîne de traitement à deux niveaux, partagée entre les journalistes eux-mêmes d'une part et l'ID d'autre part. Aux premiers incombe une tâche de *data curation*, autrement dit d'enrichissement des contenus par indexation, simultanément à la rédaction et à la publication en ligne, sur la base de méthodes semi-automatiques. L'ID accomplit quant à lui une *data curation* postérieure et à plus long terme des contenus, dans un cadre de plus grande expertise relativement aux problématiques de la catégorisation et de l'archivage. La première phase est ainsi caractérisée par l'instantanéité de l'accès aux informations additionnelles, tandis que la seconde fournit une édition et une structuration de ces données garantissant leur correction et leur fiabilité de façon pérenne.

Le premier niveau d'enrichissement des contenus, pris en charge par les journalistes et complété par deux responsables de taxonomie (*taxonomy managers*), suivi d'un traitement plus profond par l'ID, est illustré à la figure 4.5. À partir d'un article rédigé, un journaliste appelle un service Web constitué d'un système d'Extraction d'Information, SAS Teragram [3], dont le résultat est un ensemble de termes d'indexation potentiels pour l'article donné. Le système d'analyse est bâti sur un ensemble de règles d'ordre linguistique, créées par les responsables de taxonomie à partir d'un sous-ensemble du vocabulaire contrôlé mis à disposition par l'ID. Les termes suggérés en font donc partie, et sont soumis à une sélection par le journaliste, qui peut également procéder à l'insertion de nouveaux termes d'indexation si nécessaire. Après revue des responsables de taxonomie, l'article est publié en ligne et reçoit dans un second temps un traitement plus profond par l'ID : ajout de termes d'indexation supplémentaires, résumé, stockage.

1.2.3 Intégration aux Linked Data

En 2009, le NYT publie un sous-ensemble de 10 000 termes issus de son vocabulaire d'indexation sous forme de nœud des Linked Data [4]. Il s'agit plus précisément de noms de personnes, de lieux et d'organisation ainsi que de descripteurs (termes ou mots-clés), complétés par une interface de

3. http://teragram.com/
4. http://data.nytimes.com

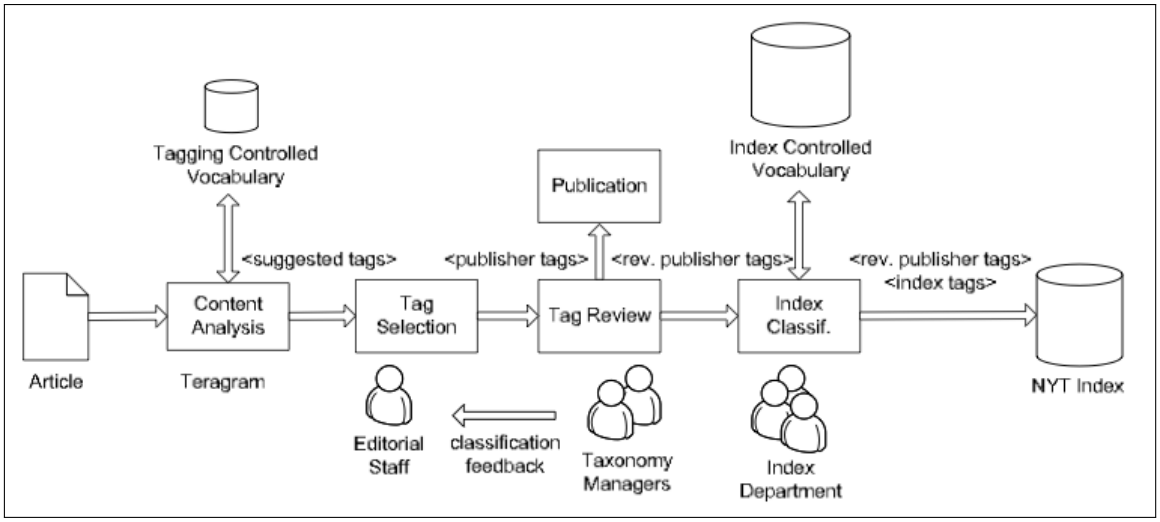


Figure 4.5 : Chaîne de traitement pour l'indexation des contenus du NYT (reproduit à partir de [Woo10]).

programmation (API, *Application Programming Interface*) permettant à toute personne ou organisation de développer des applications à partir de ces données librement disponibles. Cette initiative bénéficie au NYT en termes de volume de trafic sur son site Web, de valorisation de ses données par des applications tierces et de possibilités étendues de Recherche d'Information. La figure 4.6 illustre la mise à disposition des données d'indexation du NYT. La figure 4.7 présente un exemple d'application construite à partir de ces données d'indexation, qui permet une recherche d'information sur les anciens élèves d'université américaines et leur présence dans l'actualité couverte par le NYT.

2 Indexation et classification des contenus à l'AFP : état des lieux

2.1 Organisation générale du flux d'information

L'Agence France-Presse (AFP) est l'une des principales agences de presse internationales et généralistes, aux côtés des agences américaines Associated Press et Reuters. Sa couverture quotidienne de l'actualité dans le monde sous forme de flux est assurée par 2 260 journalistes répartis dans 150 pays à travers 200 bureaux, dont cinq bureaux principaux correspondant à la régionalisation géographique du réseau de l'AFP : Paris, Nicosie, Washington, Montevideo, Hong Kong. La production compte chaque jour environ 5 000 dépêches, 2 500 photographies, 150 videos et une centaine d'infographies. Des contenus multimedia sont également publiés sur le Web. L'ensemble des domaines traités par l'AFP le sont en six langues de travail : français, anglais, allemand, espagnol, portugais, arabe. Des partenariats avec des acteurs de la presse dans le monde donnent par ailleurs lieu à une traduction de la production dans d'autres langues : chinois, russe...

Le système de production de l'AFP est caractérisé par le principe du flux d'information, régionalisé géographiquement, catégorisé thématiquement et caractérisé formellement, mis à la disposition d'un ensemble de clients par le biais d'un abonnement. Cette mise à disposition peut relever de deux modes d'accès : le mode *push* et le mode *pull*. Le premier livre directement les contenus de l'AFP à ses clients sous la forme d'un flux continu, selon un certain nombre de critères de filtrage de la production si ceux-ci ont été définis. L'accès *pull* est quant à lui défini par le client lui-même, qui peut interroger les contenus de l'AFP *via* un serveur et la

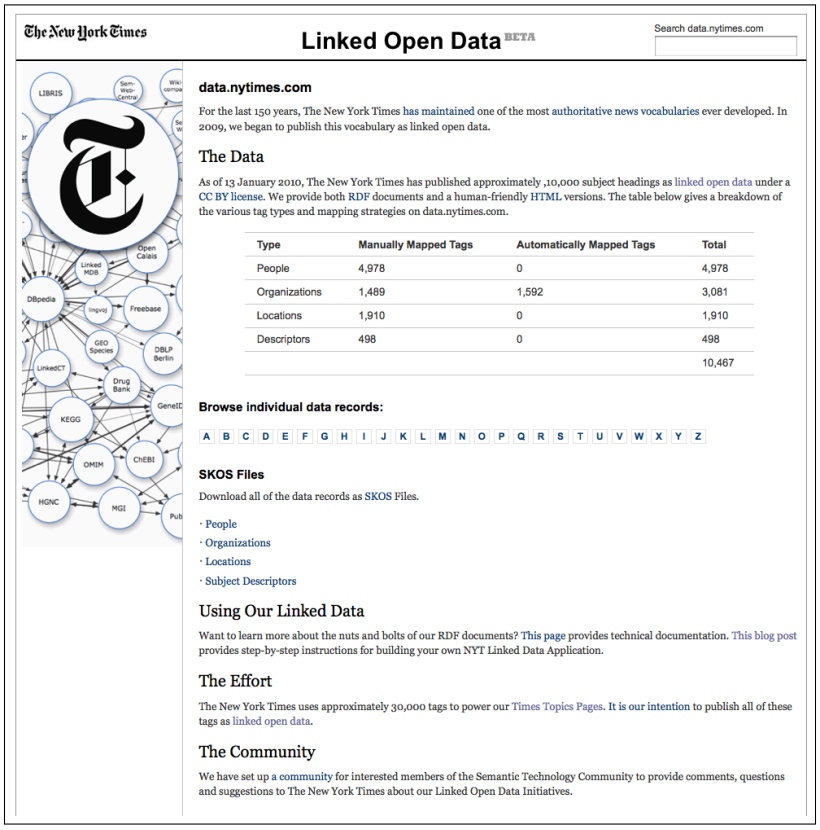


FIGURE 4.6 : Interface d'accès aux données d'indexation du NYT (`http://data.nytimes.com`).

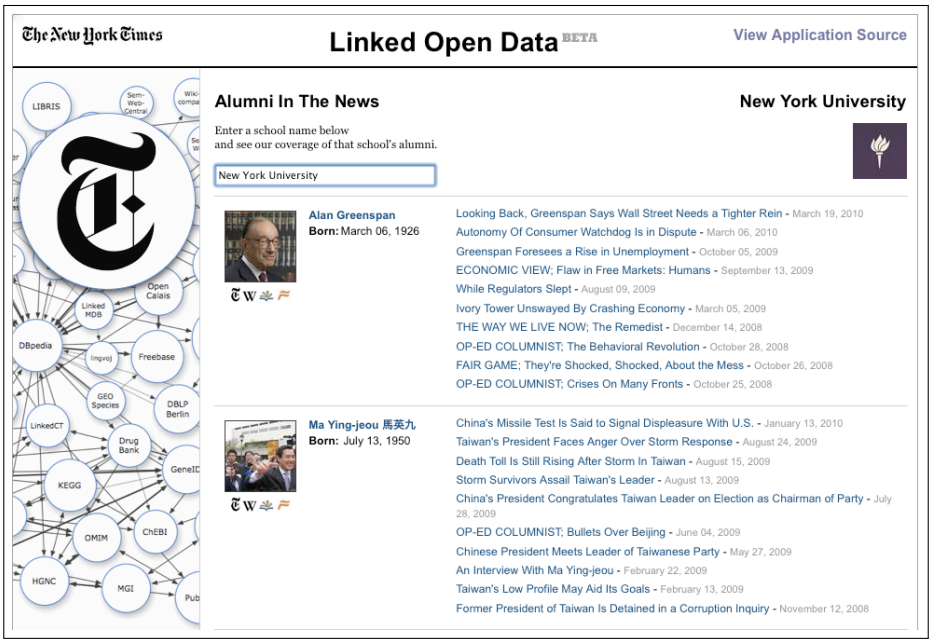


FIGURE 4.7 : Exemple d'application à partir des données d'indexation du NYT (`http://data.nytimes.com`).

définition de critères de sélection ; un ensemble de documents correspondant à la requête est alors retourné. Les clients de l'AFP sont principalement des organes de presse et medias français et internationaux ainsi que des organisations et institutions, notamment administratives.

Le flux d'information donne lieu à différents *fils*, général ou économique, par exemple. À partir d'une première information d'un journaliste de terrain, une dépêche est relue par l'un des bureaux ou *desks* pour vérification et transmission sur le fil approprié. Parallèlement aux fils existe un mode de publication sur le Web ou *journal Internet*, accessible aux clients de l'AFP. Ce journal propose depuis 1996 un ensemble d'informations sélectionnées et présentées dans un environnement graphique multimedia, également destiné à la consultation depuis les terminaux mobiles.Une banque d'images, *Image Forum*, regroupant les documents photographiques de l'AFP permet à ses clients d'accéder à ces ressources de façon comparable aux fils textuels d'information.

Dépêches La figure 4.8 présente l'architecture d'une dépêche issue de l'AFP. Chaque dépêche est caractérisée par une information à diffuser, sur un sujet précisé par le rédacteur et de façon générale accompagné d'une indication de source, garantissant la fiabilité de l'information. Les dépêches n'indiquant pas de source particulière correspondent aux informations concernant des événements de notoriété publique, par exemple la tenue d'élections nationales.

Un ensemble de *métadonnées* accompagne chaque dépêche, comme c'est également le cas pour les autres types de documents. Ces métadonnées correspondent à des attributs formels tels que les informations d'auteur, de date, de lieu de rédaction, ainsi qu'aux informations de catégorisation thématique. Il ne s'agit pas là de métadonnées ancrées dans le contenu du document, donnant une description formelle des éléments informatifs mentionnés. Ce second type de métadonnées constitue la cible des traitements proposés par le présent travail.

En termes de construction, une dépêche est centrée sur une information principale, formulée dans le titre puis dans le premier paragraphe du document, qualifié de *lead*. La source principale de l'information est également indiquée dans le lead. Les paragraphes suivants développent l'information par des éléments de détails et de contextualisation étroitement liés. On parle ainsi de forme pyramidale inversée pour décrire la structure informative d'une dépêche, allant de l'essentiel aux éléments de détails. Il doit ainsi être possible de tronquer une dépêche d'un ou plusieurs de ses derniers paragraphes sans retirer de cohérence à l'information véhiculée [Age10]. Dans les traitements automatiques effectués à partir des dépêches de l'AFP, et notamment en TAL, cette structure informative ainsi que les règles rédactionnelles suivies par les journalistes peuvent faire l'objet d'une exploitation. Quelques exemples de dépêches, suivant cette structure hiérarchisée de l'information, sont reproduits à l'annexe A (figures A.6 à A.16), dans le format d'échange dérivé du langage XML, NewsML, défini par le consortium IPTC [5] (cf. section 2.2). Les règles rédactionnelles intéressant particulièrement notre tâche d'enrichissement organisé autour des entités concernent les mentions de personnes dans les dépêches : lors de la première occurrence, le nom canonique ou complet doit être utilisé, par exemple *Barack Obama* ; les occurrences suivantes peuvent ne reprendre que le nom de famille, alors précédé du titre usuel, par exemple *M. Obama*, mais pas de l'initiale du prénom, par exemple *B. Obama*. La forme *M.* ne peut donc normalement pas être interprétée hors de son emploi pour *Monsieur* et *M. Obama* ne peut donc pas référer à l'épouse du président américain, dont le nom canonique est *Michelle Obama*.

Données textuelles La production de l'AFP est constituée de dépêches dont le contenu textuel représente la source de données principales dans le cadre de notre travail sur l'enrichissement à partir d'identification d'entités. Les autres types de documents produits par l'AFP peuvent cependant également être intégrés à des traitements visant des données textuelles dans la mesure

5. International Press Telecommunications Council (`http://www.iptc.org`)

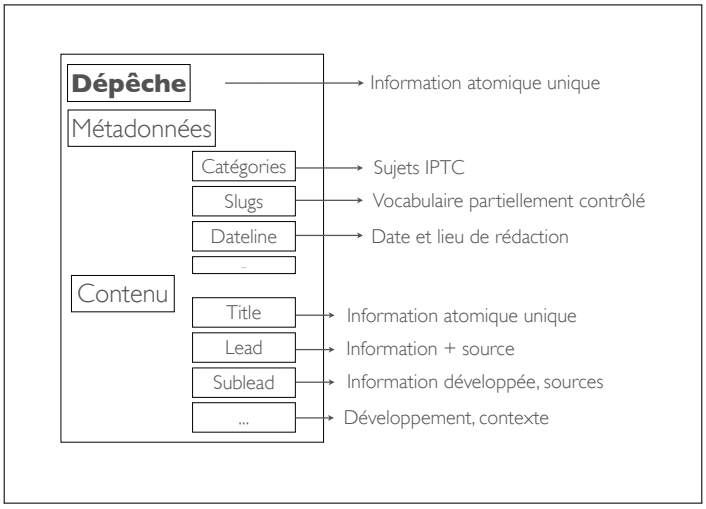


FIGURE 4.8 : Schéma structurel d'une dépêche AFP.

où ils présentent pour la plupart, en association avec leur contenu proprement non textuel — photographie, video, infographie —, un champ descriptif sous forme textuelle : légendes de documents photographiques, retranscriptions de video, texte inséré dans les infographies.

2.2 Classification thématique : la taxonomie de l'IPTC

La production de l'AFP est organisée selon une classification thématique, correspondant à une division de l'actualité générale en différents domaines. Cette structuration constitue un pivot central des modalités d'élaboration et de diffusion de l'information par l'AFP : les fils d'information ainsi que les critères de sélection qui lui sont applicables relèvent en majorité de la classification thématique adoptée.

Comme dans tout contexte d'organisation de données selon une caractérisation de l'information, un modèle définissant les catégories applicables doit être disponible. Dans le cas de l'AFP, la classification thématique de la production repose sur un vocabulaire contrôlé associé à une taxonomie élaborée par le consortium IPTC. Celui-ci regroupe les acteurs majeurs de la presse mondiale — agences telles que l'AFP, éditeurs de medias et industriels de la presse. Il exerce une fonction de développement et de maintenance de standards techniques destinés à structurer les modalités d'échange d'information entre ses membres. La taxonomie conçue à ce titre s'inscrit dans un ensemble de métadonnées proposées par l'IPTC sous le nom de NewsCodes, qui fournissent le vocabulaire de métadonnées nécessaires à l'encodage d'informations sur les documents produits. Ces informations concernent d'une part l'*administration* des documents — source, destination, date, statut de publication... — et d'autre part leur *description*, principalement en termes de catégorisation thématique à l'aide de la taxonomie évoquée. Celle-ci comprend trois niveaux hiérarchiques, un niveau inférieur présentant des termes d'ordre plus précis et détaillé que le niveau supérieur correspondant. L'usage de cette taxonomie par l'AFP est établi par correspondance entre le vocabulaire contrôlé propre à l'agence et celui de l'IPTC : le vocabulaire AFP est constitué de mots-clés, assignés par le journaliste à une dépêche lors de sa rédaction afin de la caractériser en termes de domaine. Chacun de ces mots-clés, appelé *slug*, est associé à un élément de la taxonomie IPTC, appelé *sujet*. Cette association de vocabulaire permet d'assigner automatiquement au document considéré une catégorie issue de la taxonomie IPTC et par transitivité de l'associer à l'un des sujets qu'elle définit au niveau supérieur. La figure 4.9 illustre la relation établie entre slugs AFP et sujets IPTC à travers les trois niveaux hiérarchiques existants, pour les sujets CULTURE

et Désastre. L'annexe A reproduit l'ensemble de cette double hiérarchie — sujets IPTC et slugs AFP (figures A.1 à A.5). Les dépêches figurant également dans cette annexe (figures A.6 à A.16) présentent les *slug lines* suivantes, où les termes en italique, en l'occurrence des noms d'entités, correspondent à des slugs hors vocabulaires :

Syrie-conflit (annexe A, A.6)
Social-emploi-syndicat-patronat-gouvernement (annexe A, A.10)
Somalie-France-otage-combats (annexe A, A.14)

Table des mots clefs (fr) à utiliser au [13 11 2009]

Sujet	Sujets et rubriques IPTC : nom des rubriques	Synonymes : nom des rubriques dans les systèmes AFP	Keywords : Mots de slugs
CLT	Arts, culture, et spectacles (01000000)	[Arts, culture et spectacles]	
CLT	Archéologie (01001000)	[Archéologie]	[Archéologie]
CLT	Architecture (01002000)	[Architecture]	[Architecture]
CLT	Tauromachie (01003000)	[Tauromachie]	[Tauromachie]
CLT	Festivals et commémorations (01004000)	[Carnaval] + [Festival] + [festivités]	[Carnaval] + [Exposition] + [Festival] + [commémoration] + [festivités]
CLT	Cinéma (01005000)	[Cinéma]	[Cinéma]
CLT	Festival de cinéma (01005001)	[cinéma festival]	[cinéma-festival]
CLT	Danse (01006000)	[Danse]	[Danse]
CLT	Mode (01007000)	[Mode]	[Mode]
CLT	Langue (01008000)	[langage]	[langage]
CLT	Bibliothèque et musée (01009000)	[musées]	[musée] + [musées]
CLT	Littérature (01010000)	[Littérature]	[Littérature] + [livre] + [livres]
CLT	Musique (01011000)	[Musique]	[Musique] + [opéra]
CLT	Peinture (01012000)	[Peinture]	[Peinture]
CLT	Photographie (01013000)	[Photo] + [Photographie]	[Photo] + [Photographie]
CLT	Radio (01014000)	[radio]	[radio]
CLT	Sculpture (01015000)	[Sculpture]	[Sculpture]
CLT	Télévision (01016000)	[Télévision]	[Télévision] + [audiovisuel] + [tv]
CLT	Théâtre (01017000)	[Théâtre]	[Théâtre]
CLT	Patrimoine (01018000)	[Patrimoine]	[Patrimoine]
CLT	Coutumes et traditions (01019000)	[tradition]	[tradition]
CLT	Arts (général) (01020000)	[Art] + [Arts]	[Art] + [Arts]
CLT	Divertissement (01021000)	[Spectacle] + [Spectacles]	[Spectacle] + [Spectacles]
CLT	Culture (général) (01022000)	[Culture]	[Culture] + [science-fiction]
CLT	Bande dessinée (01024000)	[BD]	[BD]
CLT	Dessin animé (01025000)	[cinéma animation]	[cinéma-animation]
CLT	Média (01026000)	[Média] + [Médias]	[Média] + [Médias]
CLT	Internet (01027000)	[Internet]	[Internet]
DIS	Désastres et accidents (03000000)	[Désastres et accidents]	
DIS	Sécheresse (03001000)	[Sécheresse]	[Sécheresse]
DIS	Tremblement de terre (03002000)	[séisme]	[séisme]
DIS	Famine (03003000)	[Famine]	[Famine]
DIS	Incendie (03004000)	[Incendie]	[Incendie] + [incendies]
DIS	Inondation (03005000)	[Inondation]	[Inondation] + [inondations] + [tsunami]
DIS	Désastre météorologique (03007000)	[Intempéries]	[Intempéries] + [canicule]
DIS	Accident nucléaire (03008000)		[accident-nucléaire]
DIS	Pollution (03009000)	[Pollution]	[Pollution]
DIS	Accident de la route (03010001)	[accident route]	[accident-route] + [route-accident]
DIS	Accident de chemin de fer (03010002)	[accident train]	[accident-train]
DIS	Accident dans l'air et l'espace (03010003)	[accident avion]	[accident-air] + [accident-aviation] + [accident-avion] + [aviation-accident]
DIS	Accidents maritimes (03010004)	[accident mer] + [naufrage]	[accident-mer] + [naufrage]
DIS	éruption volcanique (03011000)	[Volcan]	[Volcan]
DIS	Secours d'urgence (03012000)	[secours]	[secours]
DIS	Accident (général) (03013000)	[Accident] + [Accidents]	[Accident] + [Accidents]
DIS	Situation d'urgence (03014000)		[explosion]
DIS	Désastre (général) (03015000)	[Catastrophe] + [Catastrophes] + [Désastres]	[Catastrophe] + [Catastrophes] + [Désastres]
DIS	Catastrophe naturelle (03015001)	[Catastrophe naturelle]	[cyclone] + [ouragan] + [typhon]
DIS	Avalanche/Glissement de terrain (03015002)	[avalanche]	[avalanche]
DIS	Prévention et organisation des secours (03016000)	[Plans urgence] + [catastrophe plan]	

Figure 4.9 : Extrait de la table de correspondance entre sujets IPTC et slugs AFP (document AFP).

Catégorisation des dépêches Lors de la rédaction d'une dépêche, un ensemble de slugs lui est attribué. Cet ensemble contient des éléments du vocabulaire contrôlé de l'AFP mais peut également contenir tout terme jugé pertinent par le journaliste pour la caractérisation du contenu informatif, selon le mode de description général des mots-clés. Ces slugs hors liste contrôlée peuvent notamment correspondre à des entités nommées — *USA* ou *JO2012*. Au moins l'un des slugs assignés à une dépêche doit néanmoins être défini par la liste contrôlée, dans la mesure où la catégorisation selon la taxonomie IPTC est un élément formel obligatoire pour la transmission des dépêches et est automatiquement déclenchée par l'usage d'un slug défini dans la table de correspondance présentée plus haut. Ainsi, l'attribution du slug chômage, correspondant au sujet IPTC de même nom (code 00900900) induit une classification de la dépêche considérée dans la catégorie IPTC social (code 09000000). Il est important de noter que la correspondance établie entre slugs et sujets IPTC est univoque et ne prend pas en charge les phénomènes d'ambiguïté pouvant toucher les termes employés. Ainsi, le terme voile est prévu dans le vocabulaire des

slugs et la taxonomie IPTC (code 15050000), associé à la catégorie SPORT ; l'usage de ce terme pour caractériser une dépêche traitant de débats législatifs ou de faits divers en rapport avec l'interdiction de signes religieux ostentatoires dans les lieux publics devient alors impossible sans que la catégorie sport ne soit assignée à la dépêche, ce qui entre en contradiction avec son sujet réel. Dans les processus d'indexation, destinés à la diffusion et à la recherche d'information par thèmes dans la production de l'AFP, une telle catégorisation s'avère non seulement non pertinente mais erronée et potentiellement nuisible à sa qualité.

Ce dernier point relatif à l'adéquation entre la structuration de l'information adoptée et la caractérisation des contenus eux-mêmes illustre la problématique posée par un modèle tel que celui de la taxonomie employée par l'AFP. En effet, un modèle non restreint à une liste sous forme de vocabulaire, permettant de munir les termes définis d'une sémantique formelle et explicite, serait à même de prendre en charge l'ambiguïté des termes descripteurs. Dans un tel cas, des descripteurs différents renvoyant à leurs définitions respectives, établies dans le modèle et accessibles aux utilisateurs, seraient alors à même de distinguer des documents relatifs à des domaines distincts. Dans la situation actuelle, régie par l'emploi du vocabulaire de slugs et de la taxonomie IPTC, chaque terme descripteur implique une définition unique, non spécifiée explicitement et dont le bon usage est assuré par l'expertise, à un niveau implicite, des journalistes de l'agence.

Formellement, les informations de catégorisation sont ajoutées à la dépêche au format NewsML par le biais de balises dédiées, comme l'illustrent les exemples de dépêches reproduits à l'annexe A :

- Les slugs attribués à la dépêche sont regroupés dans la balise `NameLabel` ; des balises `Property` indiquent également les slugs sous forme de mots-clés, avec l'attribut `Keyword`.

- Les sujets IPTC correspondant aux slugs, lorsque cette correspondance existe, sont spécifiés dans les balises `SubjectCode`, `SubjectMatter` et `SubjectDetail` pour les premier, deuxième et troisième niveaux hiérarchiques, respectivement.

2.3 Ressources référentielles

L'AFP structure sa production selon un ensemble de ressources référentielles, regroupant les données communes à l'organisation dans sa globalité, mais également au sein de chaque silo et domaine de production. Le terme recouvre à la fois la notion de référence en termes de données considérées comme assez importantes et pertinentes pour être conservées sous une forme fixée, afin d'y référer lors des différentes procédures de production et de diffusion, mais également en tant que chaque élément de ces ressources constitue en principe l'indication d'un objet du monde porteur de cette importance. L'usage principal des ressources référentielles concerne les processus d'indexation des contenus, pouvant varier selon le silo de production — texte ou photo — et le domaine concerné — le fil SPORT par exemple dispose d'une base de données spécifique à ce domaine.

En dehors d'un certain nombre de ces ressources concernant le fonctionnement de l'agence, hors de notre sujet de travail — listes des personnels et localisations géographiques des bureaux, par exemple —, plusieurs *catalogues* sont disponibles dans le cadre de l'indexation des contenus et listent principalement des lieux, personnes, produits financiers, genres et catégories de dépêches. Les termes issus de ces catalogues, assignés aux documents produits, donnent lieu à des fonctionnalités de Recherche d'Information interne, notamment par le biais de la console de rédaction des agenciers ou *via* des serveurs d'archivage, ainsi qu'à destination des services de documentation. Ces catalogues sont caractérisés par une formalisation déjà évoquée lors de la présentation du vocabulaire des slugs et de la taxonomie IPTC (section 2.2) : il s'agit de listes

sous forme de vocabulaires contrôlés, chacune d'elles étant définie par un schéma de données distinct. La structuration générale des ressources référentielles de l'AFP ne correspond donc pas à un schéma unifié ; le projet de restructuration et de refonte rédactionnelle IRIS, actuellement en cours à l'AFP, propose d'apporter un ensemble de solutions à cette situation, notamment par la fusion et la restructuration des différents catalogues existants.

De par la catégorisation thématique des contenus selon la taxonomie IPTC, évoquée précédemment, les slugs et sujets IPTC constituent la première ressource référentielle utilisée par l'AFP. Comme cela a été illustré, l'ajout des slugs et codes IPTC aux métadonnées de documents permet une indexation ainsi qu'une récupération des documents selon cette classification aux différents stades de l'exploitation : archivage, consultation, documentation, diffusion.

Silo texte Le silo de production textuelle dispose par ailleurs de catalogues supplémentaires afin de caractériser les contenus :

Lieux Une liste de lieux est établie, regroupant l'ensemble des États du monde d'une part, et un sous-ensemble de villes et de zones géographiques notables d'autre part. Chaque élément de la liste correspond à un terme, accompagné des codes ISO-3166-2 et ISO-3166-3 [6] du pays concerné et de ses labels possibles dans les langues de travail de l'AFP. Les éléments de cette liste sont employés dans un champ spécial des documents (balise `Location`), indiquant leur lieu de rédaction. Ce champ ne constitue donc pas une indication relative au contenu mais à sa production, qui ne sont pas systématiquement synchronisés : une dépêche concernant la Syrie peut ainsi être rédigée depuis Paris en France, et le champ correspondant n'indiquera donc que ce lieu de rédaction.

Tickers Les *tickers* correspondent aux codes de produits financiers cotés en bourse, mentionnés dans les dépêches du flux FINANCE. Chaque élément de cette liste dispose d'un identifiant et de labels, ainsi que d'un label spécial définissant la forme à utiliser pour l'insertion d'un produit donné parmi les slugs d'une dépêche. Lors de la rédaction d'une dépêche du flux financier, le journaliste déclenche une recherche automatique des noms d'entreprise mentionnées, dont les résultats déclenchent à leur tour une association de chaque entreprise avec les codes de produits financiers correspondant. Ces *tickers* sont ajoutés en fin de document dans des balises XML dédiées, qui ne concernent que certains formats de diffusion de dépêches et ne sont pas conservées dans le format NewsML.

Genres Un certain nombre de dépêches peuvent appartenir à un genre spécifique. L'AFP fournit en effet, en amont et en aval des dépêches d'information typiques, des documents jouant des rôles variés dans la diffusion de l'information. Les genres correspondant à ces rôles sont notamment les suivants : *encadré, chronologie, biographie-portrait, interview, verbatim (reproduction de citations uniquement), revue de presse, analyse, synthèse, fiche technique, développement, réactions, reportage.*

Silo photographie L'AFP produit environ 3 000 documents photographiques par jour, maintenus par un service spécialisé et mis à disposition sur la base Image Forum. Celle-ci comprend des fonctionnalités de recherche portant à la fois sur les images elles-mêmes, par requêtes sur des spécifications formelles telles que la taille, la date ou le lieu de prise de vue, mais également sur les légendes accompagnant chaque photographie. Les règles rédactionnelles rendent systématique la mention des personnes figurant sur une photographie dans la légende, lorsque cela est pertinent — autrement dit, lorsque la photographie présente effectivement des personnes et si

6. `http://www.iso.org/iso/fr/home/standards/country_codes.htm`

leur identité est notable. Une liste de noms est constituée à partir de ces légendes et mentions et forme une base regroupant à ce jour quelques (1 300 000 labels dans plusieurs langues, correspondant à environ 290 000 identifiants distincts. Ces identifiants ne représentent cependant pas d'unicité référentielle et ne sont ancrés dans aucun schéma sémantique particulier ; les éventuels homonymes n'y présentent donc pas de critère particulier de discrimination. En pratique, en raison de l'accroissement de la production et de l'absence d'annotation par lots, les documentalistes photos n'annotent que 10% de la production environ.

Documentation Un service spécifique de l'AFP maintient un ensemble de ressources documentaires, sous la forme de descriptions biographiques, chronologies et synthèses. Elles concernent environ 40 000 biographies de personnalités ainsi que des synthèses descriptives des Etats du monde. Ces biographies et synthèses se présentent sour la forme de sélections de dépêches déjà diffusées, concernant les personnalités et pays en question. L'identification des entités dans ces ressources se fait par un nom. L'accès à ces informations est déterminé par les agents du service de documentation, qui les transmettent aux rédactions à leur demande ainsi que sur les fils de diffusion lorsque l'actualité requiert des informations utiles à la contextualisation des événements traités — fiche d'un pays où éclate un conflit, décès d'une personnalité... Les ressources de la documentation ne font en revanche pas l'objet d'une maintenance ou de mise à jour systématique.

3 Cas d'utilisation AFP

L'AFP occupe une place importante parmi les acteurs internationaux de la diffusion de l'information et est à ce titre concernée par des problématiques d'adaptation et de renouvellement pertinents de son mode de fonctionnement au cours du temps. Le besoin d'enrichissement des contenus s'inscrit à cet égard dans le contexte contemporain de la numérisation de l'information et de ses vecteurs, dont la dimension s'affirme comme d'autant plus essentielle avec l'avènement des pratiques de communication sociales et professionnelles liées au Web et à sa version sémantique en émergence. Présentées dans la première partie de ce mémoire, les techniques associées au renouvellement de la publication documentaire proposées par le Web Sémantique constituent pour l'AFP des cibles et opportunités de développement incontournables, en particulier pour l'enrichissement à l'aide de métadonnées. Ces dernières visent en premier lieu les entités, à l'image des applications usuelles de l'Extraction d'Information sur des données linguistiques depuis plusieurs décennies. Le Web Sémantique comme la Population de Bases de Connaissances proposent pour ces applications des ressources et méthodes dont les spécifications rejoignent de façon adéquate notre cas d'utilisation.

L'aperçu des ressources et pratiques d'indexation existantes à l'AFP proposé dans la section précédente témoigne de la nécessité d'introduire des traitements de données orientés vers les contenus eux-mêmes et non seulement leur formalisation technique. Il s'agit en effet d'intégrer la production ainsi que les modes de diffusion de l'AFP à l'espace d'échange et de communication défini dans le cadre du Web Sémantique, pour lequel les métadonnées venant enrichir les contenus jouent le rôle de points d'ancrage. Afin de remplir cette fonction, les métadonnées dont l'acquisition est nécessaire à partir des contenus traités doivent faire l'objet d'un ancrage formel dans un modèle sémantique défini et explicite. Un tel modèle, à déterminer de façon interne ou en association avec des ressources externes, se présente comme un facteur essentiel de cohérence globale au niveau de l'ensemble de la production de l'AFP en termes de structuration et de manipulation des données, notamment à des fins d'indexation et de Recherche d'Information dans le cadre de produits et d'applications spécifiques.

Avant une présentation détaillée, dans les chapitres 5 et 6, de l'approche proposée afin de mettre en œuvre cet enrichissement ainsi que du système conçu à cette fin, certains objectifs

applicatifs envisagés autour de cette tâche sont exposés. Un certain nombre de contraintes fonctionnelles dérivant de la chaîne de production existante à l'AFP doivent par ailleurs être prises en compte dans la méthodologie proposée, dont les aspects principaux sont esquissés ici.

3.1 Objectifs et applications

Le besoin d'enrichissement des contenus de l'AFP a déjà été évoqué à plusieurs reprises au cours de ce travail et conditionne la tâche à laquelle nous proposons une réponse méthodologique. Afin de cerner les objectifs et applications dont relèvent ce besoin et cette tâche, la formulation suivante reprend et fixe les éléments déjà énoncés quant à leur définition :

> [4.1] Les traitements envisagés doivent permettre un enrichissement des contenus textuels produits par l'AFP à l'aide de métadonnées ancrées dans un modèle sémantique et concernant les entités (personnes, lieux, organisations) mentionnées dans ces contenus. Les contenus visés sont en premier lieu les dépêches de l'agence, ainsi que les éléments textuels accompagnant les autres types de documents produits, notamment les légendes photographiques. Le modèle d'ancrage des métadonnées est utilisé par les traitements à effectuer ainsi que dans la phase d'exploitation des métadonnées. Il peut être conçu pour partie de façon interne à l'AFP et par utilisation de ressources externes existantes, pertinentes quant aux contenus de l'agence.

L'acquisition et l'adjonction de métadonnées aux contenus textuels vise à permettre à l'AFP de proposer une production enrichie pour la diffusion générale d'une part, et de concevoir un ensemble de produits et applications spécifiques à destination de ses clients et du public général d'autre part.

La diffusion de contenus enrichis participe d'une adaptation de la production de l'AFP à un mode renouvelé de publication et d'échanges de données, plus attractif du point de vue des utilisateurs, pour lesquels l'espace informatif livré par de tels contenus est à la fois augmenté et propice à diverses réutilisations. Celles-ci ne requièrent pas de définition *a priori* dès lors que les métadonnées associées aux contenus sont conformes aux standards élaborés dans le cadre du Web Sémantique et adoptés par les communautés intéressées à l'exploitation de données dans ce cadre technologique.

À partir de contenus enrichis en métadonnées et plus spécifiquement des URI identifiant les objets donnant lieu à ces métadonnées, les destinataires du flux de l'agence peuvent accéder à un ensemble d'informations sur ces objets — ici des entités. Les URI correspondantes sont en effet destinées à diriger l'utilisateur vers les ressources appropriées quant à la description de ces entités, qu'il s'agisse de données du Web — encyclopédies en ligne recensant les entités par page, par exemple — ou propres à l'AFP. Dans ce dernier cas, il est possible d'envisager que des ressources telles que celles de la documentation de l'agence puissent être formatées de façon à rendre accessibles les informations nécessaires par le mécanisme de référencement des URI et les protocoles de communication en réseau adéquats pour l'AFP et ses clients.

L'espace informatif ouvert par l'intégration de métadonnées sémantiques aux contenus est également envisagé dans la perspective de tâches internes à l'AFP, notamment la Recherche d'Information *via* des processus d'indexation reposant sur ces métadonnées. Le chapitre 7 illustrera notamment cet aspect du traitement des données permis par l'enrichissement des contenus.

Au-delà d'une fonctionnalité d'accès élargi à l'information, les métadonnées permettent de concevoir des traitements reposant sur les informations qu'elles véhiculent, en vue d'applications exploitant les contenus ainsi enrichis. Ces applications, notamment proposées par le service

Medialab de l'AFP, visent à la valorisation de la production de l'agence, parallèlement au processus de diffusion lui-même.

Les filtrages existants, consistant à sélectionner une partie de la production à destination de clients spécifiques, pour lesquels un ensemble de critères sont définis quant à leurs intérêts particuliers, reposent principalement sur les métadonnées de classification suivant les sujets IPTC, évoqués précédemment. Une sélection par mots-clés, sur les contenus plein-texte, est également possible. Par le biais de métadonnées, de tels filtrages peuvent être définis non seulement sur les informations de catégorisation thématique, mais également sur la présence d'entités particulières dans les contenus. Un envoi peut ainsi être paramétré pour un client souhaitant obtenir uniquement les informations concernant une entreprise et une zone géographique données.

Par extension, toute application issue de l'AFP reposant sur une sélection des contenus en termes de sujets et d'entités traités peut être développée à partir des métadonnées disponibles. Dans la cadre du projet Glocal [7] auquel a participé l'AFP, un accès aux contenus du *journal internet* (texte et photo) est proposé à partir de requêtes, construites autour de quatre questions fondamentales posées par la couverture médiatique d'événements : *quoi, qui, quand, où* (figure 4.10), identifiées par le concept journalistique nommé *5W*, de l'anglais *what, who, when, where, why* [8]. Les résultats présentés par cette application peuvent prendre différentes formes selon le choix de l'utilisateur, avec quatre orientations : basique (figure 4.11, cadre 1), géographique (figure 4.11, cadre 2), album d'images (figure 4.11, cadre 2) ou chronologique (figure 4.11, cadre 4). Des métadonnées de contenu identifiant de façon explicite et univoque les objets mentionnés sont indispensables à la mise en relation avec les champs de recherche spécifiés dans une telle application.

Les métadonnées concernant les entités peuvent également être associées à des traitements de contenus relatifs à l'extraction d'autres types d'information. Le service Medialab de l'AFP développe ainsi une plateforme de recherche et d'exploration de citations de personnalités dans les dépêches : pour un thème donné, correspondant à un mot-clé extrait du plein-texte, à un sujet IPTC ou à un slug AFP, l'utilisateur peut accéder à une base de citations associées à leurs auteurs, eux-mêmes identifiés de façon explicite et univoque (figure 4.12, avec le slug Élections2012). Cette plateforme fait l'objet d'une présentation détaillée dans le chapitre 7, consacré aux applications réalisées dans le cadre du présent travail.

Dans ces différents modes d'exploitation de la production, pour la diffusion et la conception d'application, la fonctionnalité d'identification des métadonnées concernant les entités est assurée par l'existence d'un modèle au sein duquel ces entités sont recensées et définies. La pertinence et la richesse de ce modèle sont notamment déterminées par la couverture qu'il fournit en termes d'entités ainsi que l'adéquation de son composant conceptuel aux contenus traités, mais également par les liens qu'il permet d'établir avec les ressources disponibles sur l'espace du Web et dans le cadre du Web Sémantique, notamment les ensembles de données du réseau des Linked Data. La valorisation de la diffusion et de l'exploitation est ainsi augmentée d'une dimension supplémentaire, facteur d'attractivité pour les utilisateurs des contenus et applications émanant de l'AFP.

3.2 Contraintes

La tâche d'enrichissement des contenus de l'AFP en métadonnées s'inscrit dans un environnement de travail et de production existant sur lequel repose l'efficacité du travail des agenciers. L'amélioration qu'il s'agit d'apporter en termes d'exploitation et de valorisation par cet enrichis-

7. Projet Glocal (`http://www.glocal-project.eu/`) achevé le 31 décembre 2013

8. Le 5W, réduit ici au 4W, où la question *why* (*pourquoi*) n'est pas considérée. La question *how* (*comment*) est parfois ajoutée au 5W. Ce concept est plus généralement présent dans les processus d'analyse d'événement, à visée pédagogique, d'investigation ou de recherche. On en trouve diverses formulations chez des auteurs anciens (Hermagoras de Temnos, Cicéron) ou chez Rudyard Kipling dans *Just so Stories*, 1902.

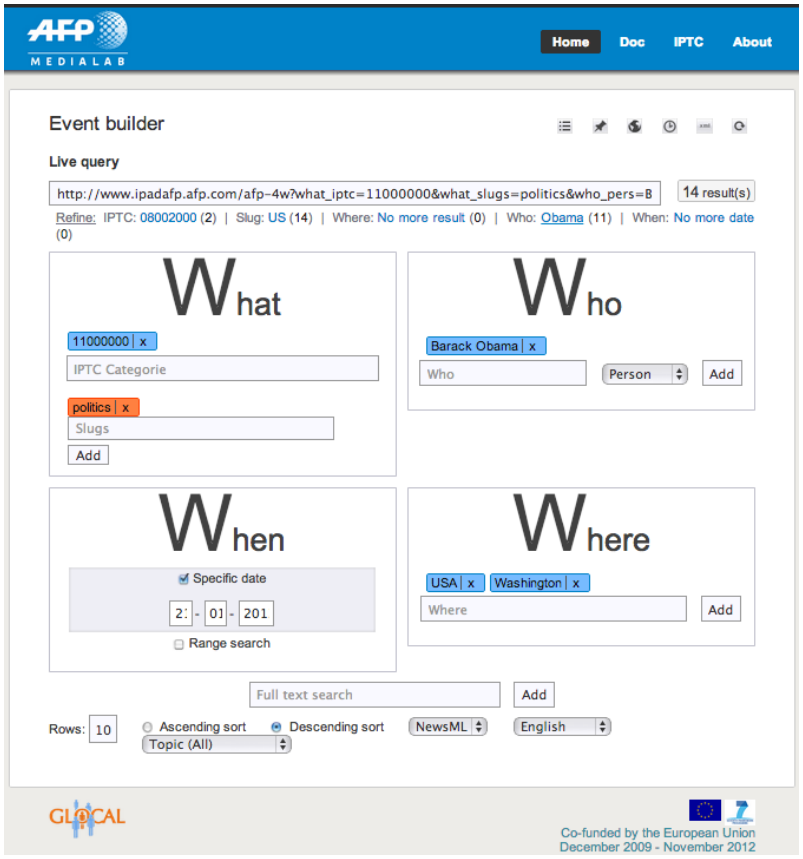


Figure 4.10 : Projet Glocal : formulaire de recherche.

sement doit donc être réalisée grâce à des outils et ressources intégrées à cet environnement, en cohérence avec ses caractéristiques essentielles.

Intégration aux CMS existants La chaîne de production de l'AFP repose sur l'usage d'une console de rédaction, apparentée aux CMS (*content management systems* évoqués précédemment. Les métadonnées d'ordre fonctionnel — date, langue, statut de publication... — ainsi que les informations de catégorisation thématique — slugs et sujets IPTC — sont ajoutées aux dépêches *via* cette console. Des CMS équivalents sont utilisés dans les autres silos de production, notamment pour la transmission de documents photographiques. Dans la perspective d'un enrichissement en métadonnées portant sur les contenus, les fonctionnalités correspondantes doivent faire l'objet d'un développement afin d'être intégrées à ces CMS et manipulées par les journalistes lors de la production. Dans un contexte d'enrichissement manuel, il s'agit de fonctionnalités de sélection des segments textuels donnant lieu à des métadonnées, ainsi que d'accès à la ressource correspondante, comprenant le modèle de définition et les éléments à même de constituer des métadonnées. L'enrichissement envisagé comme une tâche automatique ou semi-automatique doit quant à lui donner lieu à une fonctionnalité d'appel par le journaliste sur un contenu donné, ainsi qu'à un rendu visuel des résultats et de métadonnées proposées.

Intervention humaine L'intégration d'outils automatiques aux CMS, et plus précisément d'identification d'entités pour l'ajout de métadonnées aux contenus, doit permettre un degré d'intervention humaine en termes de contrôle, validation et correction. Ce degré doit être fixé en fonction

FIGURE 4.11 : Projet Glocal : présentations des résultats de recherche.

Figure 4.12 : Application AFP : recherche de citations dans les dépêches sur l'élection présidentielle française de 2012.

du temps de traitement jugé raisonnable quant à l'ensemble des manipulations à effectuer depuis l'appel du service automatique jusqu'à la validation finale de ses résultats. Ainsi, l'intervention humaine peut se réduire à une décision binaire consistant à accepter ou refuser les résultats présentés — validation ou suppression des métadonnées insérées — ou être étendue à des possibilités de correction et d'ajouts. Ces dernières allongent les temps de traitement et complexifient les modalités d'usage des outils, mais permettent un résultat de meilleure qualité. Leur ergonomie peut être améliorée par un accès aisé aux ressources à partir desquelles ajouts et corrections peuvent être définis, ainsi que par un contrôle automatique de la validité de ces opérations ; il s'agit notamment de garantir le respect des formats à employer ainsi que la restriction des insertions aux éléments préalablement définis dans les ressources adoptées.

Bruit et silence Comme c'est le cas dans nombre d'applications en Extraction d'Information, les résultats d'un système automatique sont largement évalués, dans un tel environnement de travail et pour des objectifs d'exploitation tels que ceux de l'AFP, en termes de bruit et de silence. Plus spécifiquement, un certain taux de silence peut être toléré, tandis que la précision des résultats fait l'objet d'un seuil minimal élevé : toute introduction de résultats incorrects est en effet directement visible et affecte concrètement la qualité des données. Autrement dit, le taux de précision des résultats détermine fortement la perception de la qualité des outils et ressources adoptés. Le taux

de rappel, loin d'être négligé dans leur évaluation, est quant à lui moins directement associé à des erreurs nuisant à la qualité générale, un seuil minimal jugé satisfaisant devant néanmoins être défini. La non reconnaissance d'une entité importante, notamment en termes de notoriété, tendrait cependant à dégrader fortement la perception de qualité chez l'utilisateur, même si un taux de rappel élevé est atteint lors de l'évaluation. Les outils et ressources d'identification d'entités mis en place afin de produire les métadonnées à ajouter aux documents produits par l'AFP doivent donc être paramétrés de façon à apporter la réponse la plus adéquate possible aux contraintes relatives au bruit et au silence ; leur configuration peut donc différer de celle qui donnerait les résultats optimaux lors de l'évaluation à l'aide d'autres métriques, employées dans les contextes de recherche et de développement préalables, telles que la mesure classique F_1.

Temps de traitement Le secteur d'activité de l'AFP étant largement concerné par les impératifs de temps réel de la production et de la diffusion , tout traitement associé à la production doivent tenir compte de cette contrainte de vitesse imposée. Le développement d'outils automatiques pour l'enrichissement des contenus y répond dans un premier temps en dégageant les rédacteurs de la tâche coûteuse, en temps mais également en énergie, de sélection et d'insertion relatives aux métadonnées. Ces outils doivent néanmoins accomplir cette tâche en temps quasi-réel afin d'être considérés comme efficaces et avantageux. En ce qui concerne l'intégration à la console de rédaction, le temps de traitement envisagé ne devrait pas dépasser quelques secondes, voire une seconde, par dépêche. Pour des applications spécifiques s'appuyant sur l'exploitation des métadonnées, des temps de traitements plus longs peuvent être envisagés, de l'ordre de quelques minutes ou quelques heures selon leur complexité et la quantité de données traitées, notamment si ces traitements impliquent d'autres formes d'Extraction d'Information telles que la détection de citations, évoquée précédemment. Les contenus ainsi traités ne sont en effet pas concernés par la diffusion en temps réel et font l'objet de processus journaliers par lots de données plutôt qu'au niveau de chaque document considéré individuellement.

Intégration des ressources existantes Les ressources référentielles présentées précédemment (sections 2.2 et 2.3) constituent une des cibles principales du renouvellement initié par l'AFP, à travers le projet de refonte du système de production et de diffusion actuellement mis en œuvre (projet IRIS), ainsi que, dans une mesure plus expérimentale, les méthodes d'enrichissement de contenus à l'aide de métadonnées abordées dans le présent travail. Ces ressources sont destinées, dans cette perspective, à être intégrées autant que possible à la nouvelle organisation des données et de leur manipulation. Une telle intégration implique des processus de migrations d'ensemble de données, définis dans plusieurs schémas distincts, au sein d'un schéma global et unifié. Dans cette optique, les ressources adoptées par les outils d'enrichissement et d'identification d'entités, notamment en termes de modèle, doivent présenter l'adéquation nécessaire quant aux spécifications de ce schéma global. Des procédures d'intégration doivent pouvoir être définies du modèle proposé vers le schéma unifié de l'AFP, notamment par une conformité avec les techniques et standardisations issues du Web Sémantique, précisément conçues dans un tel objectif. Les modalités de cette intégration reposent principalement sur les technologies de services distants ou Web services.

3.3 Méthodologie et spécifications

3.3.1 Intégration à la chaîne de production

Afin d'intégrer des fonctionnalités d'enrichissement des contenus en métadonnées à la chaîne de production de l'AFP en conformité avec les contraintes énoncées ci-dessus, les outils élaborés dans cet objectif sont envisagés sous la forme d'un module autonome. Dans les processus relatifs à la

rédaction et à la transmission de documents et plus spécifiquement de dépêches, un tel module peut être associé à la console dont disposent les journalistes par appel à un service accessible depuis le réseau de travail. L'interprétation des résultats retournés concerne quant à elle la console elle-même et doit donner lieu aux modifications adéquates. Au niveau des traitements non directement liés à la rédaction mais tournés vers l'élaboration d'applications exploitant les contenus et leurs métadonnées, le même module peut être accessible *via* les systèmes informatiques déployés à l'agence. L'installation des outils constituant ce module au niveau des systèmes d'information peut donc concerner l'ensemble des postes de travail de l'agence, sous la forme de programme directement exécutable ou de service distant par l'intermédiaire d'un logiciel client, intégré aux CMS rédactionnels. Enfin, ces outils doivent être adaptés aux formats de données utilisés par l'AFP, notamment le format NewsML défini par le consortium IPTC, tant en entrée qu'en sortie.

3.3.2 Identification d'entités et métadonnées

Les métadonnées visées par la tâche d'enrichissement concernent en premier lieu les entités mentionnées dans les contenus. Il s'agit donc de procéder à leur identification afin que les métadonnées dérivant des mentions d'entités soient porteuses de sens en termes d'interprétation et d'exploitation ultérieure. L'identité des entités mentionnées est issue d'un processus de mise en relation avec un modèle définissant un ensemble d'individus dont le type appartient à des classes conceptuelles telles que PERSONNE, LIEU ou ORGANISATION. L'ancrage de ces individus dans le modèle donné permet en effet de les identifier de façon unique et explicite, et cette caractéristique est transmise aux métadonnées insérées au niveau des mentions concernées au sein des contenus. L'identification des entités relativement à un modèle défini constitue ainsi la condition nécessaire au fonctionnement des métadonnées ainsi produites comme véhicule de sens à travers les contenus et leurs utilisations.

Les outils conçus pour la réalisation de cette tâche reposent donc sur la capacité à sélectionner les mentions textuelles susceptibles de constituer des métadonnées de document et à identifier les entités auxquelles elles réfèrent, étant donné un modèle et un ensemble d'instances préalablement constitués. Ces deux composants, destinés à automatiser la tâche d'enrichissement des contenus, correspondent aux méthodes d'Extraction d'Information et d'Annotation Sémantique présentées dans les chapitres 2 et 3. Au niveau plus spécifique de la tâche d'identification, les travaux relatifs à la Population de Bases de Connaissances et au Liage d'Entités évoqués au chapitre 3 proposent des orientations méthodologiques permettant la définition d'une approche complète. Il est utile d'observer que l'identification dans les contenus de l'AFP peut en partie s'appuyer sur les règles rédactionnelles relatives aux mentions d'entités suivies par les journalistes ; les personnes et organisations en particulier sont systématiquement mentionnées à l'aide d'un nom canonique et complet à la première occurrence dans un document, les occurrences suivantes pouvant consister en un nom de famille seul pour les personnes et un sigle ou acronyme pour les organisations. Il est par ailleurs établi que la mention d'une personne à l'aide d'un nom de famille est précédée d'un titre — *M.* ou *Mme* — et jamais d'un prénom abrégé par l'initiale. On trouvera ainsi la mention du président américain *Barack Obama* sous cette forme lors de la première occurrence, puis sous la forme *M. Obama* qui ne peut être interprétée comme référant à son épouse *Michelle Obama* en vertu des règles rédactionnelles.

Afin que le processus d'identification d'entités aboutisse à la production de métadonnées utilisables dans les traitements envisagés, un formatage des résultats fournis par les outils déployés doit être réalisé. Il s'agit notamment d'effectuer les conversions adéquates du format natif de ces outils, reposant sur le langage de balisage XML, vers des formats standardisés proposés par les communautés liées au Web Sémantique et à la publication documentaire numérique. La norme

RDFa [9] a par exemple obtenu le statut de recommandation du W3C en 2008 et 2012 pour sa version compatible avec le langage HTML. Elle permet l'insertion d'annotations relevant des Linked Data au sein de documents XML et de pages Web suivant le modèle RDF *via* un ensemble d'attributs de balise définis. La figure 4.13 illustre l'enrichissement d'un paragraphe de dépêche au format HTML, pour lequel l'attribut RDFa `vocab` définit le modèle de référence ; les entités forment des métadonnées avec des balises `<a>` et sont identifiées grâce à l'attribut RDFa `resource`, indiquant l'identité univoque de l'entité concernée dans le modèle. L'attribut RDFa `typeof` peut être ajouté afin d'indiquer la classe ontologique d'appartenance de l'entité, notamment à des fins de signalement aux moteurs de recherche. L'attribut RDFa `property` indique que la balise courante `<a>` constitue une URL liée à la ressource identifiée, spécifiée dans l'attribut HTML classique `href`. Le consortium IPTC propose depuis 2011 une extension spécialisée de RDFa,

```
<p vocab="http://afp.com/ontology/metadata/">
    Mme <a resource="#1000000000098348" typeof="Person" property="url"
    href="http://fr.wikipedia.org/wiki/Hillary_Rodham_Clinton">Clinton</a> a assuré que
    l'administration du président <a resource="#1000000000167398" typeof="Person" property="url"
    href="http://fr.wikipedia.org/wiki/Barack_Obama">Barack Obama</a> essayait de voir si des pays
    tiers pouvaient faire pression sur la junte militaire au pouvoir en <a resource="#2000000001327865"
    typeof="Country" property="url" href="http://www.geonames.org/1327865">Myanmar</a> afin d'obtenir
    la libération de la lauréate du prix Nobel de la paix, âgée de 63 ans.
</p>
```

FIGURE 4.13 : Exemple d'enrichissement de page HTML avec RDFa.

rNews [10], pour l'annotation sémantique de contenus journalistiques sur le Web. Cette extension intègre une référence au modèle conceptuel défini par l'IPTC. Un exemple d'annotation en rNews est donné à la figure 4.14.

```
<p>
    Mme <span about="http://afp.com/ontology/1000000000098348" typeof="rnews:Person">
    <a href="http://fr.wikipedia.org/wiki/Hillary_Rodham_Clinton" property="rnews:name"
    title="Hillary Rodham Clinton">Clinton</a></span> a assuré que l'administration du
    président <span about="http://afp.com/ontology/1000000000167398" typeof="rnews:Person">
    <a href="http://fr.wikipedia.org/wiki/Barack_Obama" property="rnews:name"
    title="Barack Obama">Barack Obama</a></span> essayait de voir si des pays tiers
    pouvaient faire pression sur la junte militaire au pouvoir en
    <span about="http://afp.com/ontology/2000000001327865" typeof="rnwes:Place">
    <a href="http://www.geonames.org/1327865" property="rnews:name" title="Myanmar">
    Myanmar</a></span> afin d'obtenir la libération de la lauréate du prix Nobel de
    la paix, âgée de 63 ans.
</p>
```

FIGURE 4.14 : Exemple d'enrichissement de page HTML avec rNews.

3.3.3 Modèle et ressources

La définition et la mise à disposition d'un modèle dédié au processus d'enrichissement est nécessaire afin de fournir un ancrage sémantique aux métadonnées qu'il s'agit d'obtenir à partir des contenus. Conformément aux différents requis de cette tâche ainsi qu'aux standards développés

9. http://www.w3.org/TR/xhtml-rdfa-primer/
10. http://dev.iptc.org/rNews

dans le cadre du Web Sémantique, en particulier au niveau de l'Annotation Sémantique, que nous nous proposons d'adopter, ce modèle est de nature ontologique au sens informatique du terme. Sa création ainsi que les modalités de sa population sont présentées en détail dans le chapitre 7 (section 1). Ses spécifications principales sont :

- La définition d'une taxonomie conceptuelle reflétant le domaine traité par la production de l'AFP et plus particulièrement celui des métadonnées ancrées dans ce modèle. Il s'agit en premier lieu de classes représentant les entités devant donner lieu à ces métadonnées, ainsi que des catégories employées pour la classification thématique des contenus.

- Une taxonomie conceptuelle simple, c'est-à-dire comptant relativement peu de classes et définissant un ensemble de propriétés ou relations minimal. Des classes et propriétés plus raffinées et en plus grand nombre induiraient en effet une complexité et un coût de maintenance non nécessaires étant donné le type de contenus et de métadonnées envisagés.

- Une possibilité d'évolution : le modèle élaboré dans le cadre du présent travail répond à des spécifications initiales, destinées au déploiement expérimental d'un processus d'enrichissement des contenus limité aux entités. Si celles-ci constituent les éléments informatifs centraux de la production, d'autres présentent également un intérêt non négligeable dans des perspectives d'exploitation comparable et doivent pouvoir être intégrés au modèle d'ancrage des métadonnées lors de développements à venir. Il s'agit notamment des événements ainsi que des fonctions existant en tant que relations entre personnes et organisations, traités sur le même plan que les entités nommées dans les contenus textuels par de nombreux travaux en Extraction d'Information. L'intégration de ce type d'information au modèle défini ici doit ainsi être rendue possible par le formalisme adopté. Le langage OWL est considéré dans cette optique comme adéquat de par ses possibilités d'expression et de manipulation des axiomes ontologiques.

- Un périmètre de données adapté aux contenus et au métier de l'agence : la population de l'ontologie adoptée comme modèle doit correspondre à un ensemble d'entités dont la notoriété et l'importance au niveau de l'actualité et des domaines traités sont considérées comme pertinentes par les journalistes. Plutôt qu'un recensement visant à une exhaustivité dont les critères de satisfaction seraient difficilement formulables, cette population vise davantage l'adéquation à ces critères de pertinence, ramenant sa taille à des quantités de l'ordre de plusieurs milliers. À titre de comparaison, les ensembles de données généralistes mis à disposition par le réseau des Linked Data, tels que DBpedia et Wikipedia, peuvent compter plusieurs centaines de milliers d'entités ; les données publiées par le NYT sur ce réseau correspondent quant à elles à environ 10 000 entités, de types comparables à ceux envisagés pour les métadonnées destinées à enrichir les contenus AFP (cf. section 1.2).

Muni de ces spécifications, le modèle ainsi mis à disposition dans la chaîne de production de l'agence tient lieu de ressource référentielle spécifique à l'enrichissement des contenus en métadonnées et peut ainsi être désigné comme le *référentiel de métadonnées* de l'AFP, nommé AMO (*AFP Metadata Ontology*). Son caractère ontologique peut être plus ou moins mis en avant selon les traitements et la place qu'ils accordent à sa structure conceptuelle, par opposition à une valorisation plus directe de l'ensemble de données qui y sont rassemblées sous forme d'instances.

Population et maintenance du référentiel de métadonnées Afin de refléter les contenus produits par l'AFP, le référentiel de métadonnées fait l'objet d'une population, autrement dit d'une définition d'un ensemble d'instances permettant d'ancrer les métadonnées potentielles. Cette population peut s'envisager selon deux axes :

- Par lot, de façon statique et périodique : les entités identifiées par l'Annotation Sémantique de contenus archivés peuvent être régulièrement extraites et proposées comme nouvelles entrées du référentiel. Des groupes d'entités importées de ressources existantes, considérées comme pertinentes à l'égard des spécifications de domaine de l'AFP, peuvent également donner lieu à une population régulière.

- Au fil de la production : l'Annotation Sémantique menée sur les contenus pour leur enrichissement simultanément à leur production identifie de façon constante des entités qui peuvent être proposées comme nouvelles entrées du référentiel.

Dans les deux cas, l'ajout d'entrées au référentiel nécessite au préalable de déterminer si les entités identifiées dans les contenus font déjà l'objet d'une instance du référentiel ou si une nouvelle entrée doit être créée.

La population du référentiel de métadonnées, notamment au fil de la production, ainsi que les éventuelles extensions et modifications apportées au modèle donnent lieu à une nécessaire activité de maintenance pour laquelle un service d'administration doit être envisagé. Cette administration est spécifiquement concernée par les éléments remontés par l'enrichissement au fil de la production, pouvant donner lieu à de nouvelles instances du référentiel.

Ressources Le référentiel de métadonnées est destiné à la couverture des éléments considérés comme les plus pertinents pour la description des contenus de l'AFP. Il constitue ainsi une cible des traitements, à partir desquels les éléments informatifs devant donner lieu à des métadonnées lui sont associés. L'ensemble des entités mentionnées dans les contenus textuels n'est cependant pas nécessairement limité à la population du référentiel en tant qu'il peut dépasser son cadre de sélection fondé sur la pertinence. Afin d'identifier les entités mentionnées dans ces contenus, avant d'évaluer leur adéquation au statut de métadonnées, les outils d'identification doivent disposer de ressources plus exhaustives, proposant des instances en nombre et diversité à même de couvrir de façon maximale la production de l'AFP. Cette configuration correspond à la mise en œuvre de l'Annotation Sémantique et aux ressources correspondantes, présentées au chapitre 3. Il découle de cette coexistence entre ressources génériques employées par les traitements et référentiel cible la duplication dans ce dernier d'un sous-ensemble d'éléments des premières. On peut observer que la population du référentiel peut également jouer un rôle, parallèlement aux ressources génériques, dans le processus de sélection et d'identification : les informations et connaissances qu'il encode pour chaque instance correspondant à des métadonnées dans des contenus déjà traités peuvent en effet intervenir en tant qu'éléments de contextualisation supplémentaires pour la résolution des divers problèmes posés par la tâche d'identification.

Linked Data Les objectifs liés à l'enrichissement des contenus de l'AFP sont en partie définis par le cadre proposé par le Web Sémantique. Les métadonnées obtenues à partir de la production ainsi que le référentiel correspondant s'inscrivent également dans ce cadre en tant que tels : leur conformité avec les standards du Web Sémantique rend possible leur publication sous la forme d'ensemble de données sur le réseau des Linked Data, de façon comparable à d'autres acteurs notoires de l'information et du Web, tels que le NYT.

Les éléments de cas d'utilisation et de méthodologie présentés dans ce chapitre font l'objet de présentations détaillées dans la suite de ce mémoire. L'approche proposée pour l'identification des entités nécessaire à l'établissement de métadonnées est décrite dans le chapitre 5, aux côtés des

différents ressources employées dans sa mise en œuvre. Le chapitre 6 est consacré au système conçu selon cette approche et donnant lieu au module d'enrichissement répondant au besoin formulé par l'AFP. Le référentiel de métadonnées ainsi que les différents aspects de sa population sont abordés au chapitre 7, qui présente et évalue également un certain nombre d'objectifs applicatifs visés par l'enrichissement de contenus.

Chapitre 5

Approche de l'identification d'entités dans les contenus textuels de l'AFP

> Avoir un système borne son horizon ; n'en avoir pas est impossible. Le mieux est d'en posséder plusieurs.
>
> Raymond Queneau

En tant que généralisation méthodologique de l'Annotation Sémantique, le Liage d'Entités (chapitre 3, section 3) se présente comme le processus par lequel les entités destinées à jouer le rôle de métadonnées pour l'enrichissement sont identifiées en termes formels et relativement à des ressources établies. L'enrichissement de contenus textuels peut ainsi être vu comme l'application de l'Annotation Sémantique à un contexte d'utilisation particulier tel que celui de l'AFP, dans laquelle la méthode du Liage intervient comme composant spécifique à l'opération d'identification. Cette approche donne lieu au développement d'un système dont le composant central repose sur l'identification mais dont d'autres aspects, relatifs au présent contexte de travail, doivent également être pris en compte. Le traitement des données de l'AFP requiert en effet un certain nombre d'adaptations autour des points suivants :

- La production de l'AFP concernée par l'enrichissement consiste en contenus de genre journalistique sous la forme de données textuelles brutes, associées à un certain nombre d'éléments de description au niveau des documents, tels que présentés au chapitre précédent (chapitre 4, section 2). De telles données impliquent plusieurs niveaux de traitements dont l'interaction peut s'avérer problématique.

- L'apport du système développé réside notamment dans son aspect automatique, qui porte sur l'ensemble des sous-tâches essentielles liées à l'enrichissement : sélection des éléments destinés à constituer des métadonnées et identification des entités sous-jacentes.

- Les ressources associées au système développé doivent correspondre à une couverture thématique adéquate au vu des domaines traités par l'AFP, autrement dit l'actualité généraliste. Formellement, elles doivent se conformer aux standards du Web Sémantique afin de garantir l'adéquation des contenus enrichis au paradigme de publication associé.

Ce chapitre propose une description de l'approche adoptée pour l'élaboration d'un système d'identification appelé Nomos, dont le fonctionnement et le processus de développement seront abordées dans le chapitre suivant. Les caractéristiques de l'approche considérée ici tiennent d'une part aux emprunts techniques et méthodologiques qu'elle effectue, notamment au niveau de

l'Annotation Sémantique et du Liage d'Entités, et d'autre part à un déploiement concret, déterminé par le cas d'utilisation AFP. Les contributions dérivant de cette approche sont les suivantes :

- Intégration des propositions à l'état de l'art issues des travaux en Liage d'Entités à l'Annotation Sémantique.
- Traitement de données en français : les systèmes d'Annotation Sémantique et de Liage tels que ceux présentés au chapitre 3 portent sur l'anglais et dans une certaine mesure sur le chinois dans le cas de la dernière édition de TAC ; le français ne fait l'objet d'une évaluation en Annotation Sémantique que dans le cas du système Wikimeta. Il semble donc important de souligner que les travaux présentés ici correspondent à des développements spécifiques pour le français. Le chapitre consacré au développement de Nomos pourra déterminer dans quelle mesure la tâche d'identification et ses méthodes sont spécifiques à la langue traitée.
- Traitement de contenus textuels bruts : les contenus textuels bruts qu'il s'agit de traiter dans le cadre applicatif de l'AFP requièrent la mise en place d'une chaîne globale, comprenant l'ensemble des prétraitements nécessaires à l'Annotation Sémantique. La question du repérage des éléments destinés à constituer des métadonnées au sein de ces contenus constitue en effet un problème non négligeable abordé à divers degrés par l'Annotation Sémantique et non pris en compte dans le Liage de TAC/KBP.

La section 1 de ce chapitre présente les différents aspects de l'approche adoptée, qui donnent lieu à des développements spécifiques relativement à la mise en œuvre de l'Annotation Sémantique et l'intégration du Liage. Les deux sections suivantes (2 et 3) font état des ressources utilisées par le système élaboré : il s'agit d'une part des corpus de développement et des connaissances nécessaires au processus d'identification des entités, et d'autre part des outils existants faisant l'objet d'une intégration dans la présente approche. Un système initial d'identification, développé antérieurement dans le cadre du même travail, est également présenté, notamment dans le but de déterminer les points d'amélioration incombant au système Nomos.

1 Reconnaissance et identification d'entités : une approche jointe pour la production de métadonnées à partir de contenus textuels bruts

1.1 Reconnaissance de mentions d'entités

1.1.1 Nécessité d'une reconnaissance automatique

Les contenus concernés par la tâche d'enrichissement étant constitués de données textuelles brutes, l'acquisition de métadonnées procède d'une phase de sélection des éléments textuels pouvant donner lieu à des métadonnées, avant une seconde étape d'identification de ces éléments en termes d'entités. Autrement dit, l'Annotation Sémantique doit faire intervenir une phase de reconnaissance des mentions qu'il s'agit d'identifier. Comme cela a été évoqué dans le chapitre 3, l'Annotation Sémantique ne traite pas cette étape comme un problème particulier. Elle revient dans le système Spotlight à considérer toutes les variantes d'entités recensées dans la ressource associée ; le système Wikimeta intègre quant à lui de façon explicite un composant de Reconnaissance d'Entités Nommées (REN), dont les résultats donnent ensuite lieu à l'Annotation Sémantique, qui fait cependant seule l'objet d'une évaluation [CGO11]. Dans les deux cas, tout élément repéré est considéré comme devant recevoir une annotation, quelle que soit la méthode ayant conduit à sa sélection. Du côté du Liage d'Entités dans le cadre de la campagne TAC-KBP [MD09 ; Ji+10 ; JGD11], cette sélection est de fait ignorée, dans la mesure où les mentions à aligner (une seule

mention par document du corpus d'évaluation) sont présentées en tant que requêtes aux systèmes participants.

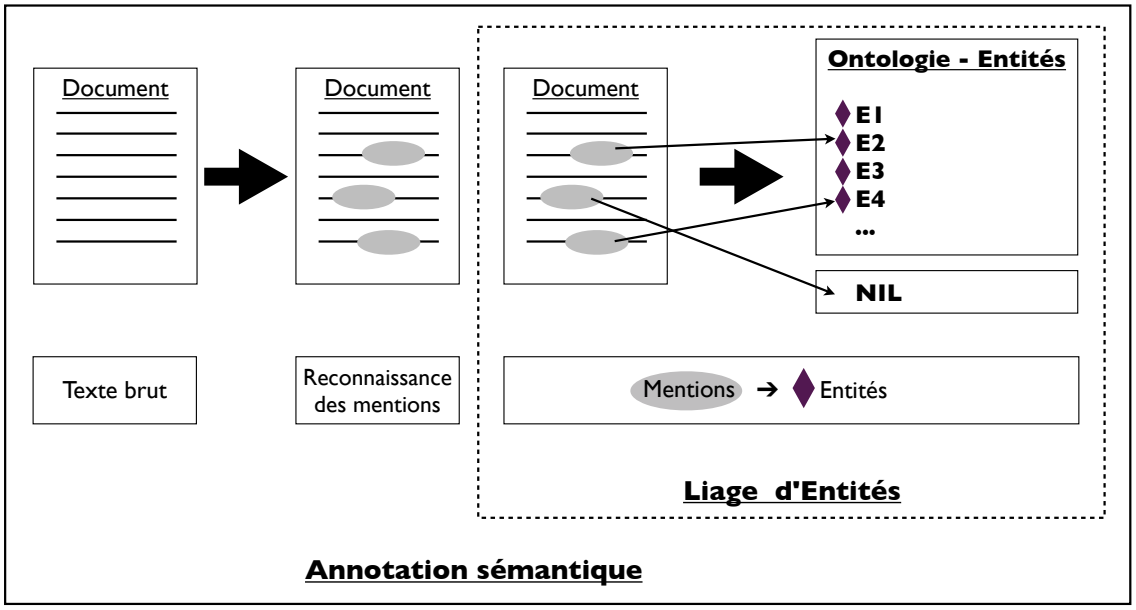


Figure 5.1 : Périmètre des tâches réalisées en Annotation Sémantique et en Liage.

Le traitement de contenus textuels dans le contexte de l'enrichissement en métadonnées trouve ainsi dans l'Annotation Sémantique une réponse méthodologique consistant en une prolongation directe de la REN, sous sa forme classique de module d'Extraction d'Information ou réalisée par d'autres moyens, tandis que ce traitement est absent en Liage (figure 5.1). Notre tâche relevant d'une configuration d'Annotation Sémantique sur des données textuelles brutes, le déploiement d'une étape de reconnaissance automatique de mentions d'entités est nécessaire afin de présenter à l'étape d'identification les éléments à même de constituer des métadonnées. L'intégration de la REN et du Liage dans un système d'Annotation Sémantique peut donner lieu à la distribution des tâches suivantes :

Reconnaissance d'Entités Nommées La REN résulte en une segmentation du texte donné en entrée et un marquage des segments correspondant à des mentions d'entités, au terme d'une analyse à deux niveaux :

- Reconnaissance des segments textuels constituant des dénotations d'entités, avec désambiguïsation totale en cas de découpages concurrents (chevauchements, imbrication, sous-analyse...)
- Typage des mentions selon le modèle adopté, avec désambiguïsation lorsque les ressources font état de variantes lexicales identiques pouvant correspondre à plusieurs types sémantiques, ou lorsque plusieurs règles de reconnaissance relatives à plusieurs types sont applicables ; les ambiguïtés de type correspondent aux cas de polysémie (homonymie ou métonymie), non visibles entre plusieurs entités de même type au niveau de la reconnaissance.

Liage Le liage considère chaque mention d'entité comme une requête munie d'un contexte — le document correspondant — à aligner vers l'une des entités recensées dans la BC à disposition, ou à reconnaître comme dénotation d'une entité ne figurant pas dans cette

base (cas NIL). Cet alignement fait intervenir un ensemble de caractéristiques sémantiques dérivées du contexte de la mention d'une part, et des connaissances disponibles pour les entités de la base d'autre part, leur similarité constituant le critère essentiel de calcul des probabilités d'alignement.

1.1.2 Propagation d'erreurs

En intégrant de façon explicite et nécessaire une étape de traitement automatique avant la tâche d'identification elle-même, la configuration de notre tâche pose la question de la relation entretenue entre ces deux niveaux. En effet, les résultats de la première étape doivent être envisagés en termes de taux de réussite et d'erreurs, propres à tout traitement et *a fortiori* automatisé. Il s'agit alors principalement de prendre en compte une possible propagation d'erreurs de la REN vers l'étape d'identification : les erreurs pouvant être retournées par un module de REN, relevant de la précision dans le cas de faux positifs et du rappel dans le cas de mentions non détectées, produisent en effet nécessairement des résultats incorrects au niveau de l'identification et de l'ajout de métadonnées.

Les sorties de REN peuvent constituer des faux positifs à deux égards : il s'agit d'une part de segments incorrectement étiquetés comme mentions d'entités, tels que *CV* (abréviation de *curriculum vitae*) identifié comme mention de type ORGANISATION dans

(23) L'entreprise avait employé 12 personnes parmi les 135 qui avaient laissé un CV.

par le système SxPipe/NP (cf. *infra*, section 3.1.1). D'autre part, une erreur de segmentation peut mener à la détection d'une mention incorrecte au niveau d'une autre mention à repérer, générant ainsi à la fois un faux positif et une mention non détectée ; c'est le cas avec le segment *Pasteur*, reconnu comme mention de type PERSONNE dans

(24) Sanofi Pasteur est ainsi venu recruter une cinquantaine d'ouvriers en intérim

par SxPipe/NP, masquant ainsi la mention de type ORGANISATION *Sanofi Pasteur*. Un cas de reconnaissance partielle comme

(25) Le président Barack Obama a approuvé un accord de coopération nucléaire civile

ou le segment *Obama* est reconnu comme mention de type PERSONNE, là où la mention à détecter correspond au segment *Barack Obama*, pose de façon plus problématique la question de la correction du résultat : il est ici à proprement parler partiel et plus difficilement qualifiable de faux que dans l'exemple précédent.

Qu'il s'agisse de faux positifs ou de correspondances partielles, de telles erreurs introduisent dans une configuration en cascade des données erronées affectant de façon particulièrement tangible la précision : un faux positif retourné par la REN donne lieu à un processus d'identification non pertinent, dont le résultat est nécessairement incorrect — la correction de l'identité assignée à une telle mention étant sans objet ; un faux positif est alors également introduit au niveau des métadonnées obtenues à l'issue de la tâche. Pour ce qui concerne le rappel, une mention non retournée n'est de fait pas traitée par le module chargé de l'identification — les résultats de la tâche à cet égard ne sont alors pas caractérisés par un bruit supplémentaire mais par un silence équivalent à celui du module de REN. Comme cela a été évoqué au chapitre 4, un cadre applicatif tel que celui de l'AFP induit une attention plus particulière portée aux performances en termes de précision, étant donnée des performances de rappel jugées satisfaisantes.

Le problème de la propagation d'erreurs au niveau d'une application séquentielle de modules est notamment formulé par Stoyanov et al. [Sto+09] dans le cadre de la résolution de coréférence, qui partage avec la tâche d'identification d'entités la nécessité de disposer préalablement des

éléments (groupes nominaux, noms propres, pronoms) constituant la cible de la résolution. Les auteurs de cette étude soulignent notamment la difficulté à évaluer de façon conclusive les systèmes reposant sur des données manuellement annotées, fournissant les éléments cibles ainsi que les données linguistiques utiles sans erreur. La comparabilité entre systèmes est également discutable, selon que l'un dispose de données de référence, par exemple celui de McCallum et Wellner [MW04], alors que l'autre repose partiellement ou totalement sur un repérage automatique des éléments cibles, par exemple celui de Yang et al. [Yan+03] ; on constate ainsi des écarts de performance de l'ordre de 20 points de F-mesure, avec un score de 91,5 pour le premier et de 71,3 pour le second, sans qu'il soit possible de conclure à la moindre qualité de ce dernier.

Cette situation renvoie au mode d'évaluation de la tâche de Liage dans TAC-KBP ainsi que dans plusieurs systèmes d'Annotation Sémantique présentés au chapitre 3 (section 2), où seule est mesurée la correction des liens établis au niveau des mentions dénotant effectivement des entités. Pour le Liage, il est alors malaisé de prédire leurs performances relativement à une configuration plus réaliste, où les mentions ne sont pas données préalablement à l'identification. Pour l'Annotation Sémantique, l'évaluation de correction des liens sur le seul ensemble des mentions correctes ne donne pas une vue générale des performances relativement à la tâche globale. Il est en revanche possible d'estimer ces performances en associant les résultats habituellement constatés pour la REN, jugés très satisfaisants avec des scores de l'ordre de 90% en F-mesure pour anglais, au taux de correction en Liage de l'ordre de 85% : une application intégrant ces deux sous-tâches aboutirait à une F-mesure de l'ordre de 76%, à même de relativiser les bonnes performances de chaque module examiné séparément et présentant une marge de progression non négligeable. La prise en charge de la relation existant entre REN et identification à proprement parler apparaît alors comme un sujet de réflexion et de propositions méthodologiques pertinent dans la perspective d'une minimisation de l'impact des erreurs d'un module à l'autre. Ce constat est d'autant plus valable pour une application de l'Annotation Sémantique au français, pour lequel la REN présente des résultats généralement inférieurs à ceux de l'anglais et peut donc nuire à la qualité de l'identification de façon plus prégnante.

1.2 Reconnaissance et identification jointes : modularité et niveaux d'analyse

Comme dans les configurations d'Annotation Sémantique de systèmes tels que Spotlight et Wikimeta, l'association de la REN et de l'identification peut prendre la forme d'une cascade de modules exécutés séquentiellement, comme l'illustre la figure 5.1. Afin de prendre en compte la possible propagation d'erreurs induites par le module de reconnaissance des mentions, des modifications de la configuration en cascade peuvent être envisagées.

Dans une première hypothèse, l'exécution séquentielle est maintenue comme telle, le second module prenant à sa charge une forme d'évaluation des résultats retournés par le premier. Un système doté d'une telle cascade peut être vu comme moins *naïf* que la configuration usuelle dans la mesure où il considère de possibles cas de faux positifs et peut être qualifié d'abstentionniste : une possibilité d'échec de l'identification peut y être définie et se traduire par l'élimination des mentions non liées, par exemple à l'aide d'un classifieur spécialisé ou d'un niveau de similarité minimal à atteindre pour toutes les hyppothèses de Liage. La question de la détection des cas de correspondances partielles se pose nénamoins : il peut sembler difficile de simplement rejeter un alignement de *Obama* (cf. *infra*, exemple 25) vers la représentation du président américain dans un tel contexte. Enfin, la distinction entre faux positifs et correspondances partielles serait par ailleurs gommée, le second cas pouvant donner lieu à une élimination mais difficilement à un rétablissement des frontières de mentions correctes.

Une seconde hypothèse, que nous proposons d'explorer plus spécifiquement dans le présent travail, consiste à envisager les processus de reconnaissance et d'identification de façon jointe. Dans une telle configuration, la reconnaissance des mentions est directement associée au proces-

sus d'alignement : celui-ci est exécuté de façon à déterminer la pertinence des éléments textuels au titre de mentions simultanément au calcul de l'alignement le plus probable en termes d'entités. Cette concomitance vise à optimiser les résultas de chaque processus en fonction de l'autre, et réciproquement : la sélection d'éléments textuels en tant que mentions est évaluée en termes de probabilité d'identification, dont la précision est ainsi améliorée. On retrouve l'idée d'une approche jointe par exemple chez Denis et Baldridge [DB09] dans le cadre d'une tâche de résolution de coréférence où l'attribution d'un type sémantique — personne, lieu ou organisation — aux mentions d'entités, qu'il s'agit de regrouper par unités référentielles, est évaluée simultanément au calcul de regroupement. Chaque sous-tâche — repérage des éléments anaphoriques, typage des mentions d'entités et regroupement par coréférence — donne lieu à un classifieur spécialisé ; les prédictions de chaque classifieur sont ensuite filtrées par l'application de contraintes transitives suivant un processus d'optimisation linéaire en nombres entiers. Cette méthode témoigne de son efficacité pour la tâche de résolution de coréférence mais présente une complexité computationnelle non négligeable.

Concrètement, une approche jointe peut être envisagée sous une forme contournant la complexité induite par un calcul simultané, c'est-à-dire joint à proprement parler, des deux niveaux d'analyse en jeu. Nous proposons en effet de conserver une configuration modulaire et séquentielle correspondant à ces deux niveaux, en établissant cependant une répartition des opérations relatives à chaque étape se distinguant de l'approche en cascade. Il s'agit de s'éloigner d'une situation où les résultats du module de reconnaissance de mentions sont déterminés et transmis comme tels au module d'identification, qui pourraient éventuellement les remettre en cause et s'abstenir de traiter un sous-ensemble jugé non pertinent (cf. première hypothèse). Au lieu de cela, une approche modulaire jointe convertit l'étape de reconnaissance de mentions en une sélection non définitive d'éléments candidats au statut de mentions d'entités, à la façon d'un filtrage préliminaire ; le processus d'identification évalue ensuite ce statut lors du calcul de probabilité d'alignement pour chaque candidat. Autrement dit, l'étape de reconnaissance s'applique en conservant une partie des ambiguïtés d'analyse. Habituellement levées pour la production du résultat final de REN, ces ambiguïtés le sont à l'issue de l'identification, qui détermine les entités les plus probablement dénotées et, par transitivité, la segmentation du texte au niveau de leurs mentions.

La répartition des décisions affectant d'une part la reconnaissance des mentions comme telles et d'autre part l'identification des entités dénotées implique alors de réviser la distribution entre REN et Liage proposée ci-dessus (cf. 1.1.1). Dans une configuration jointe et modulaire, où la reconnaissance déterminée des mentions est prise en charge au niveau de l'identification, cette répartition peut être envisagée ainsi :

Reconnaissance d'Entités Nommées L'étape de reconnaissance établit une sélection préliminaire de segments textuels pouvant constituer des mentions d'entités, limitée à des critères d'ordre surfacique. En cas d'ambiguïtés de découpage, les mentions alternatives sont conservées. Les ambiguïtés de type sont également maintenues. Une telle indétermination permet de reporter à l'étape d'identification la décision concernant l'analyse la plus probable.

Liage L'ensemble des segments retournés par l'étape de reconnaissance est considéré. En cas d'alternatives d'analyse au niveau de segments concurrents, la probabilité d'alignement entre mentions et entités, établie à l'aide critères sémantiques, permet de déterminer l'analyse la plus probable en termes de reconnaissance. Autrement dit, la désambiguïsation du niveau textuel est dérivée d'informations relevant des connaissances liées aux entités elles-mêmes et à leur probabilité de mention dans un contexte donné, par opposition à une désambiguïsation classique de REN s'appuyant essentiellement sur des critères d'ordre surfaciques.

Cette répartition repose sur l'idée que la levée de certaines ambiguïtés concernant le niveau

textuel des mentions est étroitement liée à leur interprétation dénotationnelle : l'entité à laquelle il est effectivement fait référence est associée à un contexte global, c'est-à-dire au niveau du document, à partir duquel il est possible de reconnaître de possibles mentions comme étrangères à ce contexte et donc non pertinentes au niveau de la zone d'ambiguïté. La REN permet donc de réduire dans un premier temps l'espace de recherche des éléments à identifier en tant qu'entités par application de règles de reconnaissance valables localement, tandis que leur interprétation en termes de sémantique référentielle contraint ensuite davantage cette validité selon des critères propres à l'identification et non aux modalités de reconnaissance surfacique. En ce qui concerne le typage des mentions par l'étape de reconnaissance, la décision est également reportée à l'étape d'identification : les ambiguïtés de type relevant essentiellement de l'homonymie, il s'agit bien là d'une ambiguïté d'ordre référentiel que l'identification permet de résoudre, un type étant associé aux entités vers lesquelles les mentions sont à aligner. Le type peut cependant être dérivé avec un degré élevé de certitude par les règles de reconnaissance, notamment par l'observation d'indications surfaciques ; le type d'une mention placée en concurrence d'analyse avec d'autres mentions de types différents peut donc constituer une contrainte d'interprétation utile lors de l'identification : la conformité du type entre mention et entités candidates peut ainsi être prise en compte dans le choix d'alignement et ainsi favoriser la reconnaissance de la mention adéquate étant donnée une zone d'ambiguïté.

L'approche proposée ici consiste donc à ramener la reconnaissance de mentions à une sélection partielle et essentiellement surfacique visant au repérage de l'ensemble des éléments potentiellement porteurs de dénotation ; les ambiguïtés à ce niveau sont conservées dès lors que leur résolution implique la prise en compte de critères d'ordre sémantique et contextuel. L'étape d'identification correspondant au Liage est quant à elle augmentée, dans la mesure où le processus d'alignement induit également une analyse au niveau des mentions. Le schéma d'Annotation Sémantique correspondant au cas général (figure 5.1) est alors modifié comme l'illustre la figure 5.2. La distinction entre faux positifs à proprement parler et correspondances partielles est maintenue dans cette configuration : l'étape d'identification peut aboutir à l'élimination d'une mention dont le caractère dénotationnel est jugé très faible ou nul, comme dans le cas de *CV* dans l'exemple 24 ; dans un cas comme celui de l'exemple 25, les deux mentions *Barack Obama* et *Obama* peuvent être prises en compte et éventuellement alignées sur la même entité ; la première peut ensuite être privilégiée sur la seconde en fonction de critères intégrés à la tâche d'identification.

1.2.1 Reconnaissance et identification : ambiguïté et lectures

Dans la configuration modulaire et jointe décrite précédemment, les modalités d'application de la REN jouent un rôle central : il est en effet nécessaire que le module de REN adopté procède à l'identification des segments pouvant constituer des mentions, sans que ce repérage n'entame la levée d'ambiguïtés relevant de critères hors de portée de cette étape d'analyse surfacique. La REN qu'il s'agit ici d'employer doit donc retourner une analyse de l'axe syntagmatique textuel *non-déterministe* ou *ambigue*, correspondant aux différentes *lectures* qu'il est possible de définir sur cet axe en termes de segmentation et de mentions d'entités. Ces lectures sont représentées par des *zones d'ambiguïté* au sein desquelles un même segment textuel peut faire intervenir différentes analyses relativement à ses découpages en mentions ainsi qu'à leur typage. Chacune de ces analyses correspond à un *chemin* défini par une séquence de mentions de longueur égale ou supérieure à 1 ; les chemins se distinguent les uns des autres suivant l'organisation de la séquence qu'ils présentent, en termes de découpage et de typage. La figure 5.3 illustre une zone d'ambiguïté (des états 1 à 3) ainsi constituée au sein de l'exemple 24.

Les données textuelles traitées par le système d'identification selon cette approche sont ainsi soumises à une première analyse portant sur les différentes lectures en jeu au niveau de leur segmentation en mentions. On peut observer, comme sur la figure 5.3, que l'ensemble de la chaîne

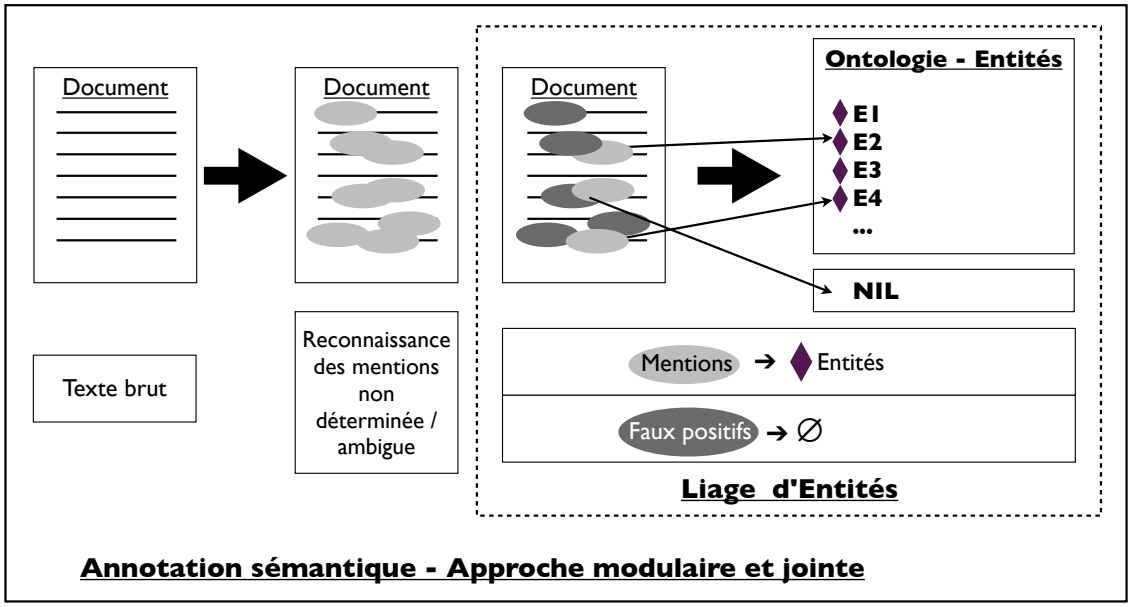


FIGURE 5.2 : Périmètre modifié des tâches réalisées en Annotation Sémantique et en Liage dans la configuration modulaire jointe.

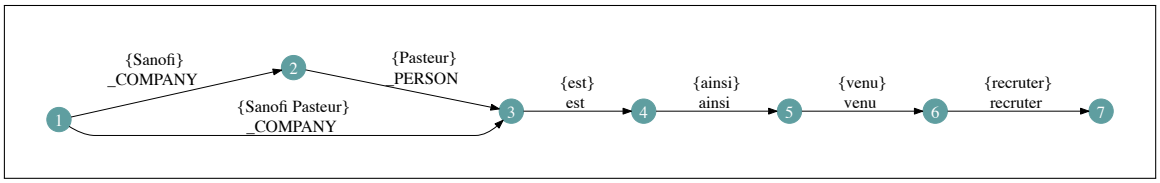


FIGURE 5.3 : Reconnaissance de mentions : exemple de zone d'ambiguïté (états 1 à 3).

considérée n'est pas concernée par cette analyse, autrement dit que le processus de reconnaissance ne donne lieu à des mentions d'entités qu'au niveau de certains segments, présentés sous forme de zones pertinentes pour l'étape suivante d'identification.

Cette représentation des données textuelles correspond à la structure des automates finis : pour une phrase donnée, l'analyse retournée par le module de reconnaissance correspond à un automate dont certaines zones, comprenant au moins une lecture interprétable comme entité, peuvent être ambigues. Hors de ces zones, les transitions de l'automate correspondent aux *mots* de la phrase, qu'il s'agisse de tokens, de formes ou de termes, selon le prétraitement effectué, et ne sont pas traitées par le module d'identification. Les zones concernées par des mentions sont normalisées de façon à constituer une disjonction de chemins de même longueur : elles présentent un seul état initial et un seul état final. Au sein de ces zones, au moins deux chemins comprennent donc une ou plusieurs transitions correspondant à des mentions d'entités, et peuvent également présenter des transitions non associées à des mentions. Cette normalisation permet de comparer aisément une à une les différentes lectures représentées ; elle est également utile dans les traitements impliquant des intersections d'automates, dont il sera question dans la suite de ce travail.[1] Différentes configurations de ces zones d'ambiguïtés sont présentées par la figure 5.4.

1. L'opération de normalisation accroit le nombre d'états des automates traités, mais cette augmentation reste limitée dans la présente configuration et ne donne pas lieu à des difficultés computationnelles notables.

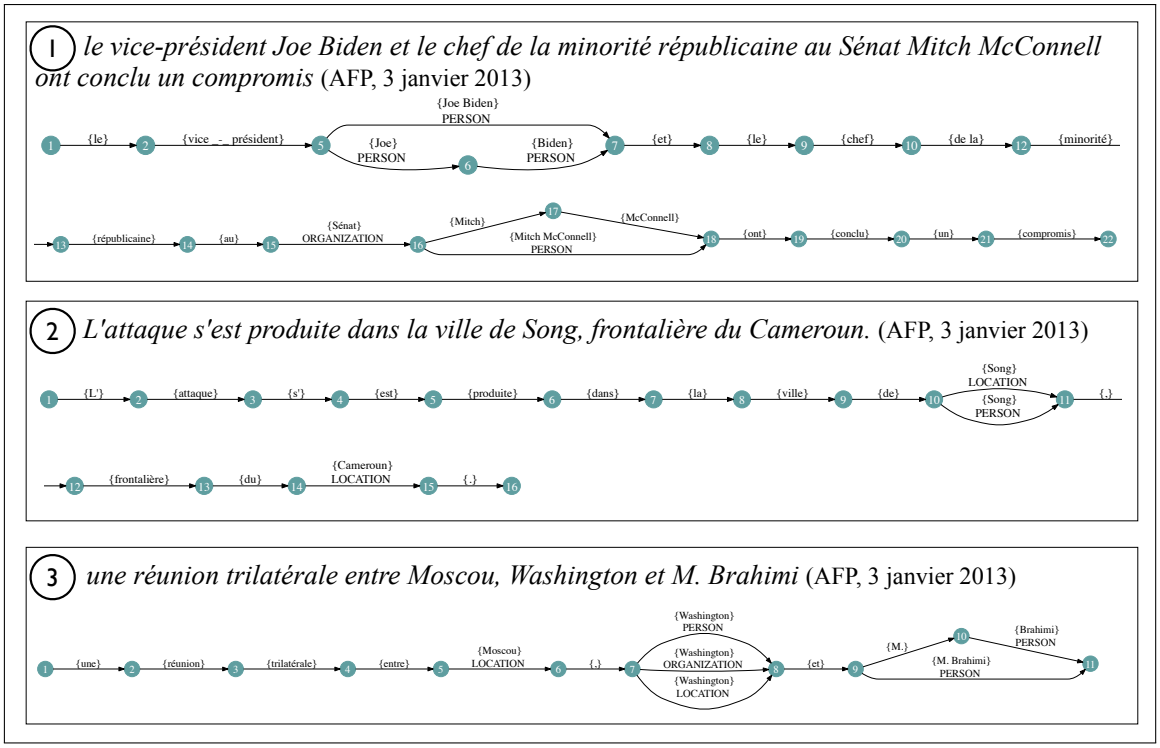


Figure 5.4 : Reconnaissance de mentions : exemples de configurations de zones d'ambiguïté (1 : états 5 à 7 et 16 à 18 ; 2 : états 10 à 11 ; 3 : états 7 à 8 et 9 à 11).

Lors de l'étape d'identification, les mentions de chaque zone constituée par la REN ambigue sont soumises à l'opération d'alignement usuelle en Liage d'Entités (cf. chapitre 3, section 3) : un ensemble d'entités candidates, disponibles dans la ressource associée au système, est défini relativement à la mention considérée ; les éléments contextuels d'occurrence de cette mention, dérivés du document courant, sont ensuite comparés aux connaissances rassemblées pour chacune des entités candidates afin d'établir l'entité dénotée par la mention avec le plus haut degré de probabilité. À l'issue de cette opération exécutée pour chaque mention pour tous les chemins de la zone, chacun d'eux est alors décoré par les informations d'identification associées à chacune de ses transitions ayant le statut de mention. La conservation d'ambiguïtés au niveau des lectures en termes de mentions doit permettre de déterminer la lecture la plus probable en fonction des entités les plus probablement dénotées : une désambiguïsation finale doit alors intervenir afin de sélectionner cette lecture. Cette sélection peut être effectuée à l'aide d'une mesure attribuée à chaque chemin, dérivée des informations relatives aux entités résultant de l'alignement des mentions de ce chemin ; les modalités de cette dérivation constitue un point méthodologique abordé dans le chapitre 6. Le résultat des opérations successives de reconnaissance et de sélection, au niveau des entités ainsi que des mentions, consiste ainsi en un automate entièrement désambiguïsé, dont les transitions présentant des mentions sont associées à des entités identifiées au sein de la ressource considérée, ou à l'entité spéciale NIL, comme l'illustre la figure 5.5.

L'approche décrite ici permet d'éliminer un certain nombre de lectures retournées par le module de reconnaissance ; les lectures erronées représentent ainsi des faux positifs qui, s'ils avaient été retournés de façon déterminée et non conjointement avec les autres analyses possibles, auraient donné lieu à une identification sans objet. Comme évoqué précédemment, le statut de faux positif peut être attribué à des mentions dont les frontières ou le type sont erronés, par exemple *Orange Alain Labé* en tant que nom de personne dans l'exemple 5.5, mais également

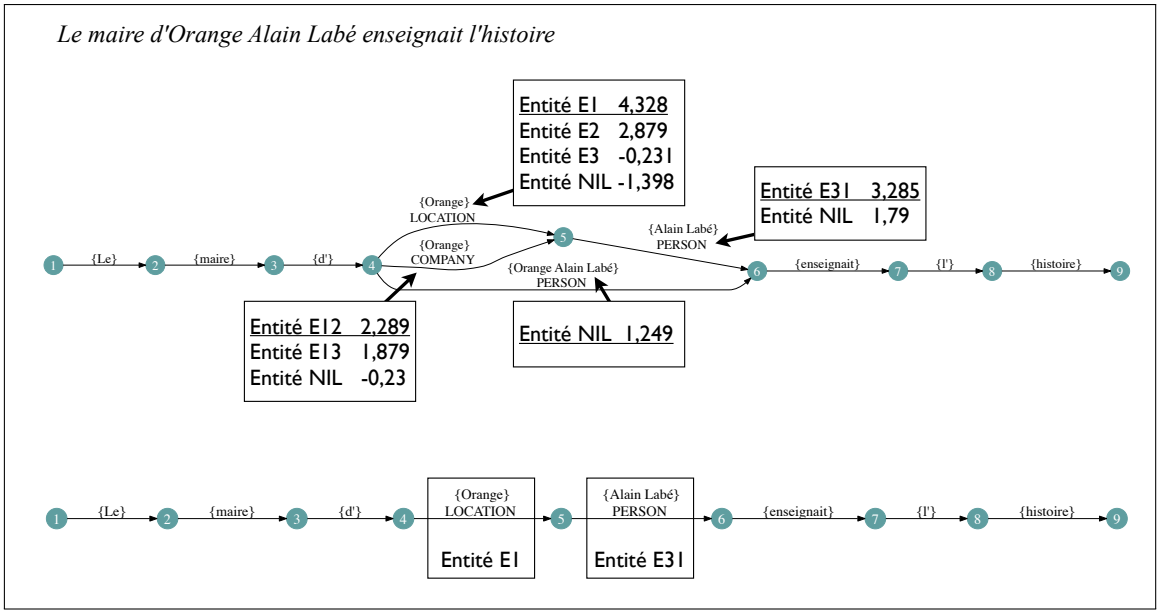


FIGURE 5.5 : Haut : zone d'ambiguïté (états 4 à 6) décorée des informations d'identification. Bas : Lecture désambiguïsée selon les informations d'identification.

et à plus proprement parler à celles pour lesquelles le segment textuel correspondant n'est pas concerné par une telle lecture. Dans un cas tel que celui illustré par la figure 5.6, le processus d'identification devrait éliminer l'analyse proposée : la mention *CDI* de type ORGANISATION peut par exemple être alignée avec la Commission du droit international (organe de l'ONU), puis être éliminée en raison d'une dénotation peu probable étant donné le contexte. Chaque mention est alors à aligner avec l'une des entités candidates correspondantes, avec l'entité spéciale NIL ou avec l'analyse négative. Il est utile d'observer ici que l'identification d'une entité comme NIL diffère d'une élimination de lecture en tant que mention d'entité, ce que le système élaboré à partir de l'approche considérée ici doit prendre en compte. Les modalités de sélection à ce niveau ainsi que la méthode de désambiguïsation des lectures en fin d'analyse font l'objet d'une étude dans le chapitre 6, consacré au système Nomos.

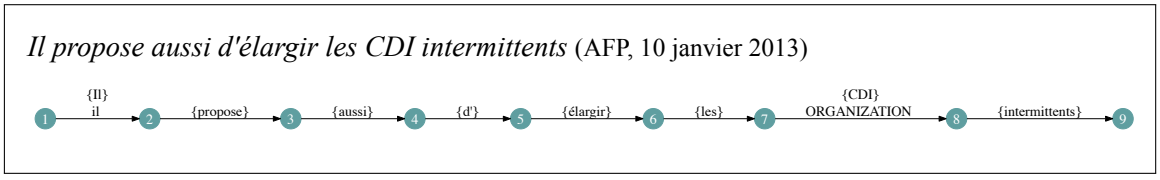


FIGURE 5.6 : Reconnaissance de mentions : exemple de faux positif.

L'idée générale consistant à envisager une tâche de façon jointe plutôt que sous forme séquentielle est présente dans les recherches et la littérature liées au TAL, dès lors que cette tâche repose sur des traitements préalables. Il s'agit notamment de l'étiquetage en parties du discours ou de l'analyse syntaxique, dans des configurations de traitement de l'oral ou de l'écrit. Dans ces domaines, la décomposition en sous-tâches (segmentation puis étiquetage, étiquetage puis analyse syntaxique, etc.) prive chaque niveau d'analyse locale de connaissances plus globales qui seraient nécessaires à l'obtention d'une solution optimale. Ces connaissances ne sont en ef-

fet accessibles qu'au niveau supérieur — étiquettes morpho-syntaxiques pour la segmentation, structures syntaxiques pour l'étiquetage, par exemple. La conservation d'ambiguïtés à un niveau d'analyse donné apparaît alors comme un moyen de reporter les décisions affectant ce niveau à un stade où les connaissances nécessaires sont accessibles. La tâche intègre alors de façon jointe les différentes analyses relevant du problème traité. La décomposition en sous-tâches demeure néanmoins pertinente : elle permet d'appliquer des méthodes efficaces pour les décisions ne relevant que du niveau local et ainsi de réduire l'espace de recherche du niveau supérieur aux décisions relevant de connaissances plus avancées. On retrouve ce paradigme chez Shafran et al. [Sha+11], qui l'appliquent à l'étiquetage morpho-syntaxique de sorties de reconnaissance de la parole sous forme de DAG (*directed acyclic graph*, graphe orienté acyclique, ou treillis), ou chez Chappelier et al. [Cha+99] qui effectuent une analyse syntaxique à l'aide de grammaires hors-contexte sur des données similaires. L'analyseur syntaxique FRMG (La Clergerie [LC05]) repose sur une segmentation du texte donné en entrée sous forme de DAG, réalisée par la chaîne de traitement surfacique SxPipe [SB08] qui sera évoquée dans la suite de ce travail (cf. *infra*, section 3.1). La chaîne Macaon de Nasr et al. [Nas+11] combine plusieurs modules de traitement, des sorties de reconnaissance vocale à l'analyse en constituants surfaciques (*chunking*), chacun d'eux prenant en charge les hypothèses multiples générées par le précédent.

1.3 Mise en œuvre de l'approche modulaire jointe

L'approche modulaire jointe proposée ici pour la réalisation de l'Annotation Sémantique destinée à l'acquisition de métadonnées peut être récapitulée par les points suivants :

Données textuelles et lectures Les données textuelles brutes à partir desquelles il s'agit d'obtenir des métadonnées sont analysées en fonction des différentes lectures possibles étant données les mentions d'entités repérables en surface. Cette analyse est permise par l'emploi d'un module de REN retournant une segmentation non déterminée et ambigue du texte d'entrée.

Identification informée L'étape d'identification est similaire dans son fonctionnement au Liage tel que défini par la tâche de Population de Bases de Connaissances (chapitre 3, section 3). Il met en jeu les composants principaux de génération de candidats puis de sélection du candidat le plus probable pour l'alignement. Sa portée est en revanche modifiée dans la présente approche : toutes les mentions retournées par la reconnaissance sont alignées, pour toutes les lectures constituées, dont au plus une seule est correcte. Les cas de faux positifs purs doivent par ailleurs donner lieu à un alignement négatif ou sans objet afin d'éliminer les lectures non dénotationnelles. La cible de l'identification consiste ainsi en un ensemble d'entités — définies ou non dans le cas NIL — augmenté du cas négatif où l'objet même de l'identification est absent.

Résolution jointe Organisée en deux modules séquentiels de reconnaissance et d'identification, l'approche considérée ici présente néanmoins un fonctionnement joint : la non-détermination de l'étape de reconnaissance permet une évaluation de ses résultats par l'étape d'identification selon des critères propres à ce niveau d'analyse. La résolution finale affecte à la fois la reconnaissance et l'identification, étant donnée une méthode d'ordonnancement des lectures enrichies d'informations d'identification.

Reconnaissance d'Entités Nommées ambigue Afin de constituer un ensemble de lectures possibles correspondant aux différents découpages et typages en termes de mentions pour un segment textuel donné, le module de REN peut

- se présenter sous la forme d'un système capable de retourner une analyse ambigue de façon inhérente,

- ou combiner plusieurs systèmes de REN dont l'union donne potentiellement lieu à des analyses différentes, à partir desquelles les zones d'ambiguïté peuvent être constituées. On peut observer à cet égard que la combinaison de plusieurs systèmes de REN constitue potentiellement un moyen de filtrage préalable dans l'étape de reconnaissance, notamment si seule l'intersection des résultats de ces systèmes est présentée à l'étape d'identification ; il est en effet possible de supposer que cette intersection constitue un ensemble d'analyses au degré de probabilité plus élevé que des résultats retournés de façon isolée par un système, le consensus en la matière pouvant être interprété comme un indice de confiance. Dans ce cas, l'élimination de faux positifs repose davantage sur cette sélection par intersection que sur la capacité du module d'identification à repérer les dénotations fausses.

La modularité de l'approche proposée permet, au niveau de la reconnaissance, d'utiliser n'importe quel système de REN existant, et par extension l'ensemble des combinaisons formées à partir des résultats de plusieurs systèmes. Dans le cadre applicatif de l'AFP, étudié dans le présent travail, cette configuration rend possible l'intégration de divers systèmes, qu'il s'agisse de travaux réalisés dans le cadre du logiciel libre, notamment par des équipes de recherche académique, ou issus de développements industriels commandités ou acquis par l'agence dans le cadre de partenariats commerciaux. La section 3 du présent chapitre présente deux systèmes de REN que nous proposons d'intégrer au système d'identification automatique d'entités Nomos faisant l'objet du chapitre 6.

Bien que les performances en termes de REN seule ne constituent pas la cible de notre travail, principalement concerné par la qualité de l'identification des entités mentionnées dans les contenus textuels de l'AFP, il est important d'observer que la présente approche lie étroitement cette qualité à cette de la reconnaissance des mentions. Les résultats du système développé selon cette approche pourront donc donner lieu à une évaluation de la REN induite.

Ordonnancement des lectures À l'issue des deux étapes de reconnaissance et d'identification, le texte donné en entrée se présente sous la forme d'un automate dont les zones ambigues sont décorées par des informations relatives aux entités sur lesquelles les différentes mentions sont alignées, associées à une mesure d'ordonnancement indiquant pour chaque mention l'entité la plus probablement dénotée. Il s'agit alors d'effectuer un second ordonnancement, au niveau de ces lectures, afin d'obtenir un automate dont les zones d'ambiguïté sont converties en une séquence de transitions unique par sélection de la lecture placée au premier rang de cet ordonnancement. Celui-ci peut être directement dérivé des informations associées aux lectures, chacune d'elles, dans une zone donnée, présentant une ou plusieurs entités munies d'un score, correspondant à sa probabilité d'alignement. Dans le cas où une transition de la lecture considérée correspond, au terme de l'identification, à un faux positif, cette analyse comporte elle aussi un score. Chaque chemin de zone d'ambiguïté peut alors recevoir un score calculé à partir des scores atomiques de chacune de ses transitions. L'ordonnancement de la zone est alors fonction des scores de chacun de ses chemins. Les modalités de calcul du score au niveau des chemins seront abordées lors de la description du système Nomos au chapitre suivant.

Afin de compléter la présentation de l'approche que nous proposons d'adopter pour la mise en œuvre de la tâche d'enrichissement de contenus textuels de l'AFP, les sections suivantes font état des ressources utilisées aux différents stades de développement et de déploiement du système d'identification correspondant. Il s'agit d'abord de corpus de développement, d'évaluation et de tests, similaires aux données qu'il s'agit de traiter dans le cas d'application, ainsi que des

connaissances relatives aux entités constituant les cibles de l'identification. En second lieu, le module de reconnaissance est constitué, dans ce système, d'outils existants choisis pour leur adéquation avec la configuration adoptée et leur disponibilité.

2 Ressources : Corpus et connaissances

2.1 Corpus d'apprentissage et d'évaluation

L'approche proposée pour la réalisation de l'enrichissement de contenus textuels à l'AFP comporte notamment un module d'identification reposant sur le Liage d'Entités, qui se présente comme une tâche de classification supervisée (cf. chapitre 3, section 3.2). Sa mise en œuvre requiert donc la disponibilité de données similaires à celles devant être traitées afin de nourrir l'apprentissage automatique du modèle d'alignement et d'évaluer ses performances. De telles données doivent constituer un ensemble de référence relativement à cette tâche, autrement dit se présenter sous la forme de corpus annotés ou *de référence* (*gold standard*). Les annotations requises visent les entités mentionnées ainsi que les informations d'alignement les concernant, en regard des ressources décrites ci-après — modèle et connaissances associées à la tâche.

2.1.1 Corpus de dépêches AFP et annotation

Un corpus de référence a été constitué à partir du fil de dépêches de l'AFP, généraliste et en français. Il couvre l'ensemble des catégories IPTC (cf. chapitre 4, section 2.2) à l'exception des dépêches marquées des sujets REL (religion et croyance) et WEA (météo).

Ce corpus de référence, nommé GAFP, comporte 96 dépêches datées des mois de mai et juin 2009. Chaque dépêche se présente selon le modèle introduit au chapitre 4 (section 2.1) et comporte donc

- des informations relatives à sa production : date, heure, lieu — le pays étant indiqué sous la forme du code ISO-3166-3 correspondant, la ville et éventuellement la zone géographique par un label,
- des informations de catégorisation, sous forme de *slugs* ou mots-clés et de codes de sujets IPTC,
- un titre
- un contenu sous forme de paragraphes.

Le corpus compte

- 881 paragraphes, soit en moyenne 9 par dépêche,
- 1 043 phrases, soit en moyenne 1,2 phrase par paragraphe,
- 37 134 tokens, soit en moyenne 35 tokens par phrase.

La distribution des catégories IPTC couvertes est donnée à la table 5.1. Les codes alphabétiques des catégories correspondent à la table reproduite à l'annexe A. Le corpus GAFP présente 153 associations entre dépêche et catégories, plusieurs codes pouvant être attribués à une dépêche.

POL	ECO	CLJ	CLT	UNR
39	23	20	15	14
HTH	**SOI**	**ENV**	**HUM**	**SOC**
12	6	6	5	3
SCI	**SPO**	**EDU**	**LIF**	**DIS**
3	3	2	1	1

TABLE 5.1 : Distribution des catégories IPTC sur les 96 dépêches du corpus de référence GAFP.

L'annotation de ce corpus est accomplie de façon à fournir les informations nécessaires à l'apprentissage de la fonction d'alignement étant donné un ensemble de mentions et un ensemble d'entités pouvant être dénotées, le cas des entités absentes des ressources considérées étant également pris en compte. Les éléments sujets à l'annotation sont les mentions d'entités de types PERSON (personnes), LOCATION (lieux) et ORGANIZATION (organisations), correspondant au modèle décrit dans la section suivante. Chaque élément annoté indique les positions de début et de fin de mention au sein du texte à l'aide d'une balise de format XML dont le nom « ENAMEX » est emprunté aux standards des campagnes MUC (cf. chapitre 2, section 3.1). Cette balise indique à l'aide d'un attribut l'identifiant de l'entité dénotée en correspondance avec la ressource d'entités mise à disposition, Aleda, décrite au chapitre 3 (section 1.3) et présentée pour notre tâche à la section suivante. En cas d'absence de l'entité dénotée dans cette ressource, l'identifiant *null* est utilisé afin d'identifier le cas spécial NIL. Le type et le nom normalisé de l'entité dénotée sont également indiqués par des attributs, correspondant aux informations équivalentes spécifiées pour l'entité considérée dans la ressource. Dans le cas NIL, un type parmi les trois possibles et manifestement adéquat ainsi qu'un nom normalisé sont ajoutés. La figure 5.7 présente ce format d'annotation pour quelques passages extrait du corpus GAFP.

```
"Beaucoup de ce que nous avons fait n'aurait pu être accompli par des robots", a souligné de
son côté <ENAMEX id="1000000001164112" type="Person" name="John Grunsfeld">John Grunsfeld</ENAMEX>,
qui a dirigé trois des sorties spatiales.

Le président <ENAMEX id="1000000000167398" type="Person" name="Barack Obama">Barack Obama</ENAMEX>
pourrait dévoiler un plan de réforme de l'immmigration détaillé à l'issue de cette réunion qu'il
présidera, selon cette source.

Une autre étude réalisée par des chercheurs néerlandais de la <ENAMEX id="2000000006951830"
type="Organization" name="Vrije Universiteit Amsterdam">Vrije Universiteit</ENAMEX> d'
<ENAMEX id="2000000002759794" type="Location" name="Amsterdam">Amsterdam</ENAMEX> en 2008
avait montré qu'une carence en vitamine D pourrait augmenter le risque de dépression ou
d'autres problèmes psychiatriques chez les personnes âgées.

<ENAMEX id="null" type="Person" gender="m" name="Mohammed Abdullah Warsame">Mohammed Abdullah
Warsame</ENAMEX>, 35 ans, habitant de <ENAMEX id="2000000005037649" type="Location"
name="Minneapolis">Minneapolis</ENAMEX> (<ENAMEX id="2000000005037779" type="Location"
name="Minnesota">Minnesota</ENAMEX>, nord), encourt 15 ans de prison et une amende de
250.000 dollars.
```

FIGURE 5.7 : Format d'annotation du corpus de référence GAFP.

Il convient d'observer que cette annotation a été réalisée de façon semi-automatique, par application d'un module de REN puis d'un système initial d'identification, qui sera présenté à la section 3.2, sur le contenu brut du corpus. Les résultats de cette première annotation automatique ont ensuite été corrigés et complétés manuellement par ajout des éléments non repérés automatiquement, suppression des faux positifs et correction des annotations erronées,

en termes de reconnaissance comme d'identification. Cette opération a été accomplie sans appel à des annotateurs et par une seule personne. Les mesures d'accord usuellement associées à la production de données de référence dans un cadre d'annotation par plusieurs annotateurs humaines sont donc non applicables ici.

L'annotation ainsi réalisé aboutit à un corpus de référence dont les caractéristiques sont rassemblées dans les tables 5.2 et 5.3 relativement à la distribution des mentions et entités par dépêche et par type. Les tables 5.4 et 5.5 indiquent le nombre d'entités candidates pouvant être associées à chaque mention distincte lors de l'alignement, les candidats correspondant aux entités pour lesquelles la mention est incluse dans l'ensemble des variantes connues pour ces entités. Ce taux d'ambiguïté est donné plus spécifiquement pour les mentions dénotant des entités correspondant au cas NIL. Dans les tables 5.2 à 5.5, les termes *mention*, et *mention distincte* réfèrent à une occurrence d'entité et une chaîne de caractère unique, respectivement ; le terme *entité* désigne une référence unique ; les qualificatifs *connu* et *inconnu* s'appliquent respectivement à des entités présentes dans la ressource associée à l'annotation et absentes de cette ressource ; associés à des mentions, ils s'appliquent aux entités dénotées par les mentions. La table 5.3 indique le taux de mentions d'entités non recensées par la BC (cas NIL ou inconnues). Ce taux de 18% est particulièrement faible en comparaison avec celui de l'ensemble de données de la tâche de Liage dans TAC-KBP (57% pour l'édition de 2009 [MD09]). On peut plus généralement observer que les dépêches de l'AFP tendent à mentionner en majorité des entités à forte notoriété et récurrentes [2], par opposition au corpus de TAC-KBP où les mentions à aligner sont explicitement sélectionnées sur des critères de difficulté, dont la *confusabilité* (ambiguïté, synonymie, référent absent de la base, etc., cf. chapitre3, section 3.2).

	moy.	**min.**	**max.**
Mentions	15	2	63
connues	12		
inconnues	3		
Entités	6	2	33
connues	4		
inconnues	1		

Table 5.2 : Distribution des mentions et entités par dépêche dans le corpus GAFP.

	Type	**Connues**	**Inconnues**	**Total**
Mentions	PERSON	281	151	432
	LOCATION	652	28	680
	ORGANIZATION	312	111	423
	Total	**1 245 (81%)**	**290 (19%)**	**1 535**
Entités	PERSON	126	99	225
	LOCATION	222	19	241
	ORGANIZATION	117	58	175
	Total	**465 (73%)**	**176 (27%)**	**641**

Table 5.3 : Distribution des mentions et entités par type dans le corpus GAFP.

2. La répartition des mentions d'entités dans les corpus de l'AFP semble en effet suivre une distribution de Pareto ou zipfienne, dans laquelle une minorité d'entités à forte notoriété représente la majorité des mentions, et inversement. Cette observation, notamment faite par les journalistes à l'origine des dépêches, nécessiterait cependant d'être étayée par des comptes systématiques d'entités sur un large corpus, ce qui n'est envisageable qu'à l'aide d'une méthode d'identification automatique particulièrement sure.

# Entités candidates	min	min \ 0	max	moy.	moy. \ 0
par mention distincte (toutes les entités)	0	1	56	2	3
par mention distincte (entités inconnues seulement)	0	1	47	6	0,98

TABLE 5.4 : Taux minimaux (à partir de 0 et à partir de 1), maximaux et moyens d'ambiguïté entre candidats (entités de la ressource) par mention distincte dans le corpus GAFP.

# Entités candidates	0	1	>1	≥ 1	Total
Mentions distinctes (toutes les entités)	244	341	164	505	749
Mentions distinctes (entités inconnues seulement)	178	10	16	26	204

TABLE 5.5 : Distribution des mentions distinctes selon le taux d'ambiguïté (nul, égal à 1, supérieur à 1) entre candidats (entités de la ressource) dans le corpus GAFP.

Représentation des dépêches pour le Liage L'alignement entre mentions et entités candidates relève principalement d'une comparaison des contextes d'occurrences des premières avec les connaissances mises à disposition sur les secondes dans une base de connaissances (BC). Le contexte d'une mention est ainsi dérivé du document dans lequel elle apparaît : la représentation qui en est faite détermine un ensemble d'attributs, dont la mention hérite. Cette dérivation, aboutissant aux attributs présentés à la table 5.6, concerne les points suivants :

Contenu textuel Chaque dépêche présente un vocabulaire représenté sous la forme d'un sac de mots, pondéré de deux façons. Le texte de la dépêche est soumis à un prétraitement usuel : segmentation en tokens, filtrage des tokens non pertinents (mots grammaticaux, mots les plus fréquents du français, tokens non lexicaux) et des hapax, stemming par troncation des terminaisons courantes du français [3]. Chaque token finalement obtenu est associé au nombre de ses occurrences au sein du document d'une part — ce qui correspond à la représentation de document habituellement effectuée en Recherche d'Information — et à un score établi à l'aide d'un test de Student ou test *t*. Ce dernier identifie les tokens présent dans le document considéré avec une fréquence statistiquement significativement plus élevée que dans un ensemble de documents de référence [4]. Le résultat du test *t* effectué sur chaque élément du sac de mot permet ainsi de lui attribuer un degré ou score de saillance dans le document courant, et donc une statut de descripteur dont la pertinence correspond à ce score. On note l'ensemble lexical pondéré par le nombre d'occurrences locales D_{bow}, et l'ensemble lexical pondéré par le score de saillance D_{sbow}.

Mentions L'ensemble des mentions d'entités apparaissant dans une dépêche, noté $D_{mentions}$, constitue un élément de représentation de ce document au même titre que le vocabulaire dérivé de son contenu textuel. Le rôle informatif de cet ensemble n'est cependant pas limité au niveau lexical caractérisant le vocabulaire d'une dépêche. Les mentions d'entités en tant que groupe de co-occurrences peuvent en effet être considérées les unes par rapport aux autres comme éléments de contexte mutuel ; autrement dit, les co-occurrences de mentions traduisent les rapports entretenus entre les entités ainsi dénotées, relativement au thème et à l'information traités par la dépêche.

3. Cette méthode est choisie pour sa simplicité, notamment comparée à la lemmatisation, et permet en outre de capturer une partie de la dérivation des formes traitées. Les terminaisons considérées pour la troncation sont *-e, -s, -t, -x, -es, -et, -nt, -ent.*

4. Le corpus de référence est l'ensemble des contenus textuels de l'encyclopédie Wikipedia.

Catégories thématiques La production de l'AFP étant catégorisée selon la taxonomie de l'IPTC (cf. chapitre 4, section 2.2), chaque dépêche est associée à une ou plusieurs catégories de cette taxonomie, formant l'ensemble D_{iptc}.

Slugs En relation avec la catégorisation IPTC, le rédacteur attribue à chaque dépêche un ensemble de slugs (cf. chapitre 4, section 2.2) noté D_{slugs}, jouant le rôle de mots-clés dans les processus d'indexation et de diffusion de la production. Les slugs peuvent être empruntés à une liste fermée, établie par correspondance avec la taxonomie IPTC, ou librement choisis par le rédacteur.

Date et lieu de rédaction Des métadonnées spéciales renseignent le lieu de rédaction de chaque dépêche, par indication d'un identifiant de pays selon la norme ISO-3166-3, ainsi que d'un nom de ville et éventuellement de zone géographique (région), administrative (département) ou non définie (cf. chapitre 4, section 2.3). La date de rédaction est également indiquée à l'aide d'une métadonnée. On dispose donc pour chaque dépêche des éléments $D_{country_code}$, D_{city}, D_{area} et D_{date}.

Données	Exemples
D_{bow}	au|1 troi|1 noyad|1 premier|1 veuv|1 entr|1 arafa|1 tourcoing|1 poi|2 ind|1 avenir|1 coloni|1 csm|1 afghanistan|1…
D_{sbow}	emira|1.74 obama|1.74 examen|1.41 exportatric|1.01 fissil|1.01 torturer|1.01 retraitem|1.01 energi|1.01 enrichissem|1.01 uranium|1.01…
$D_{mentions}$	{*Barack Obama, Emirats arabes unis, Maison Blanche, Obama, Emirats, Washington Post, Etats-Unis, Golfe, Iran* }
D_{iptc}	{POL, UNR}, {ECO, CLJ}, {REL, CLT, POL}
D_{slugs}	{USA, Emirats, nucléaire, énergie}, {USA, Birmanie, politique, SuuKyi}…
{$D_{country_code}$, D_{city}, D_{area} }	{FRA, Paris}, {USA, Chicago, Illinois}, {AFG, Base française de Nijrab}
D_{date}	21 janvier 2012, …

Table 5.6 : Attributs de représentation de dépêches AFP pour la tâche d'identification.

Le développement et l'évaluation du système Nomos, présenté au chapitre 6, s'appuie sur le corpus GAFP. Des expériences pourront également être menées sur un second corpus : le Corpus Arboré de Paris 7 (French Treebank, Abeillé et al. [ACT03]) a en effet été annoté manuellement selon la même procédure que celle décrite pour le corpus GAFP. Les caractéristiques du corpus obtenu, nommé ici GFTB, sont présentées à l'annexe C. Parmi elles, on note une proportion de mentions d'entités absentes de la base Aleda significativment supérieure à celle constatée pour le corpus GAFP (27% des mentions, et 44% des entités uniques, contre 27% dans le corpus GAFP). Il s'agit en effet d'un corpus dont les documents datent des années 1990 : la couverture de ce corpus par Aleda, dérivée de ressources créées ultérieurement, présente un certain nombre de lacunes. L'impact de ce nombre accru de cas NIL à identifier, en comparaison avec le corpus GFTB, constitue l'un des points à examiner dans le cadre d'expériences menées à partir du corpus GFTB pour le développement du système Nomos.

2.2 Entités et connaissances pour l'identification

L'identification automatique d'entités repose notamment dans l'approche proposée ici sur l'application de la méthode générale du Liage présentée au chapitre 3 (section 3). Il s'agit principalement de pondérer les hypothèses de dénotation d'une entité, répertoriée dans un ensemble préalablement constitué, par une mention donnée en fonction de similarités observables entre le contexte de cette mention et les informations associées à l'entité considérée au sein de la base de connaissances (BC) adoptée pour la tâche.

La constitution de cette BC revêt un aspect important dans la conduite de cette tâche, dans la mesure où les informations mises à disposition doivent permettre une comparaison pertinente avec le contexte d'occurrence des mentions à aligner. Il s'agit ainsi de se munir de connaissances de type similaire ou comparable à celui des données traitées. La BC considérée doit également, et ce avec le même degré d'importance, présenter une couverture en termes d'entités permettant de maximiser les alignements informatifs avec les mentions présentes dans la production textuelle de l'AFP, autrement dit de retourner dans le plus grand nombre de cas possibles une identification d'entité plutôt qu'une référence de type NIL. Comme cela a été évoqué, le domaine à couvrir à ce niveau est d'ordre général, et les entités mentionnées dans les corpus de l'AFP pour lesquelles une identification constitue une information pertinente présentent une notoriété établie dans l'espace public.

Dans le cadre de la tâche de Liage définie par la campagne d'évaluation TAC-KBP, la BC mise à disposition des participants consiste en une dérivation de l'encyclopédie en ligne Wikipedia, présentée au chapitre 3 (section 1.3). Celle-ci présente en effet un nombre d'entités et une distribution en termes de domaines adaptés aux corpus à traiter, en grande partie formés à partir de documents journalistiques et non restreints à un domaine particulier. Chaque entité répertoriée dans cette base bénéficie de plus de l'ensemble des informations rassemblées à son sujet au sein de l'article encyclopédique dont elle fait l'objet ; ces informations relèvent de caractéristiques associées aux entités, telles que les attributs biographiques ou typologiques, mais également d'une forme de contextualisation des entités, grâce au contenu textuel de l'article ainsi que des mentions faites de l'entité considérée dans les autres articles de l'encyclopédie. On peut observer que, dans le cadre de TAC, l'inventaire d'entités et la BC constituent un seul et même ensemble, exclusivement dérivé de Wikipedia.

La présente approche propose de s'appuyer, comme dans le cadre de TAC-KBP, sur Wikipedia afin de constituer la BC nécessaire au fonctionnement du système à élaborer, mais également sur d'autres ressources. Plus précisément, il s'agit de définir d'une part un ensemble d'entités considéré comme adéquat de par son périmètre quantitatif et thématique, Aleda, et d'autre part une BC associée, Nomos-KB, établissant pour chaque entité inventoriée les informations nécessaires à leur alignement lors de l'identification.

2.2.1 Aleda

La version française d'Aleda utilisée ici est à ce jour formée de deux ensembles d'entités, établis à partir de Wikipedia pour les entités de type PERSON et ORGANIZATION [5] et de GeoNames, également présenté au chapitre 3 (section 1.3), pour les entités de type LOCATION. Le périmètre des entités disponibles est ainsi en adéquation avec les besoins de la tâche puisqu'il hérite de la couverture de

5. Aleda définit un type COMPANY pour les entreprises, que nous intégrons en pratique au type ORGANIZATION. Cette base comprend également des entités de type PRODUCT (produits et marques), WORK (œuvres, films, romans...) et FICTIONCHAR (personnages de fictions), obtenues à partir de Wikipedia, que nous ne considérons pas dans la présente tâche d'identification. Les informations relatives à la population d'Aleda dans ce chapitre sont donc entendues à l'exclusion de ces entités.

GeoNames, particulièrement exhaustive et riche en termes de caractéristiques des lieux recensés, ainsi que de Wikipedia. Comptant environ 920 000 entités, Aleda est comparable par sa taille à la BC fournie aux participants de TAC-KBP.

La base Aleda associe à chaque entité un ensemble d'attributs de nature statique dérivés des ressources correspondantes, listés à la table 3.7 et reproduite ici (table 5.7). Aleda présente par ailleurs un ensemble de variantes lexicales pour chaque entité, également dérivées de Wikipedia et GeoNames et, pour les noms de personnes, augmentées de variantes calculées à partir de la structure en prénom, nom de famille et autres noms tels que les *middle names* américains. La définition de ces variantes lexicales en relation avec les entités, intervenant lors de la construction de la base Aleda et donc indépendamment du système pour lequel elle est utilisée, permet de disposer immédiatement et sans temps de calcul supplémentaire de l'index inversé utile pour la constitution de l'ensemble des entités candidates lors de l'alignement des mentions (cf. chapitre 3, section 3.3).

Les tables 5.8 et 5.9 présentent quelques exemples d'entités et de variantes recensées par Aleda. La table 5.10 rappelle la distribution des entités par type ; les tables 5.11 et 5.12 rendent compte, pour les 801 003 entités et 972 646 variantes d'Aleda, des taux de synonymie et de polysémie respectifs.

Attribut Aleda	**Attribut Wikipedia**	**Attribut GeoNames**
Identifiant	[nombre entier unique]	[nombre entier unique]
Nom canonique	titre d'article	nom canonique (souvent nom en anglais)
Type	typage Wikipedia	GeoNames → type LOCATION
Poids	taille de l'article	nombre d'habitants
Description	premiers mots du résumé	-
Lien	URI Wikipedia préfixe d'URI : `http://wikipedia/wiki/` préfixe d'URI abrégé : `wp/`	URI GeoNames préfixe d'URI : `http://geonames.org/` préfixe d'URI abrégé : `geon/`
Sous-type	-	correspondance (cf. table 3.6 p. 95)
Code pays	-	code pays
Longitude	-	longitude
Latitude	-	latitude

Table 5.7 : Correspondance entre attributs Aleda et Wikipedia ou GeoNames.

2.2.2 Nomos-kb

Parallèlement à l'inventaire d'entités fourni par Aleda, une BC à proprement parler est donc constituée, principalement à partir de Wikipedia et de façon comparable aux travaux réalisés par les différents participants à la tâche de TAC-KBP, référencés au chapitre 3 (section 3). La présente tâche d'identification étant à réaliser sur des données en français, l'édition linguistique française de Wikipedia est adoptée. Dans cette base, appelée Nomos-kb, les entités de type PERSON et ORGANIZATION répertoriées dans Aleda sont associées à des connaissances dérivées de Wikipedia, c'est-à-dire des articles dont elles sont le sujet ainsi que des contextes de chacune de leurs occurrences au sein d'autres articles. Les entités de type LOCATION, non obtenues à partir de Wikipedia mais de GeoNames, donnent lieu à des informations de type différent, décrites ci-après.

Si Wikipedia constitue une source d'information riche au sujet des entités, sa qualité de corpus encyclopédique la distingue nettement des corpus à traiter dans la tâche d'identification,

ID et nom	Attributs	Variantes
1000000000001054 Émile Benveniste	**type** : PERSON **poids** : 15 **lien** : `wp/Émile_Benveniste` **descr.** : [...] *linguiste français* [...]	Benveniste E. Benveniste Emile Benveniste É. Benveniste Émile Benveniste
2000000000745044 Istanbul	**type** : LOCATION **sous-type** : CITY **poids** : 11 174 257 **code pays** : TR **long./lat.** : 28,949 66, 41,013 84 **lien** : `geon/745044`	Istanbul Byzance
2000000003017382 Republic of France	**type** : LOCATION **sous-type** : COUNTRY **poids** : 64 768 389 **code pays** : FR **long./lat.** : 46, 20 **lien** : `geon/3017382`	France Republique Française
1000000003065020 Parti radical de gauche	**type** : ORGANIZATION **poids** : 37 **descr.** : [...] *parti politique français* [...] **lien** : `wp/Parti_radical_de_gauche`	Parti radical de gauche Parti Radical de Gauche Parti Radical de gauche Parti radical de Gauche PRG Mouvement des Radicaux de Gauche

TABLE 5.8 : Exemples d'entrées de la base Aleda.

Entités				
ID	**Type**	**Nom canonique**		
2000000002510769	LOCATION	Kingdom of Spain		
1000000000050915	PERSON	Michael Jordan		
1000000000680078	PERSON	George W. Bush		
2000000005379513	LOCATION	Orange (California)		
1000000000059373	ORGANIZATION	Orange		
Variantes				
ID	**Variante**	**FirstName**	**MidName**	**LastName**
2000000002510769	Espagne	–	–	–
1000000000050915	M. Jordan	M.	–	Jordan
1000000000050915	Michael Jordan	Michael	–	Jordan
1000000000050915	Jordan	–	–	Jordan
1000000000680078	George Walker Bush	George	Walker	Bush
1000000000680078	George Bush	George	-	Bush
2000000005379513	Orange	–	–	–
1000000000059373	Orange	–	–	–

TABLE 5.9 : Exemples d'entrées de la base Aleda : structure des variantes lexicales.

PERSON	ORGANIZATION	LOCATION	Total
304 158	59 652	465 926	**801 003**

Table 5.10 : Distribution des entités d'Aleda par type.

# Variantes par entité	# Entités
1	620 565
2	109 291
> 2	71 147
# max. = **102**	1

Table 5.11 : Nombre d'entités d'Aleda associées à 1, 2, plus de 2 et 102 variantes (nombre maximal d'associations).

# Entités par variante	# Variantes
1	903 753
2	47 119
> 2	21 774
# max. = **246**	1

Table 5.12 : Nombre de variantes d'Aleda associées à 1, 2, plus de 2 et 246 entités (nombre maximal d'associations).

en termes d'organisation et de structuration des documents mais également de distribution du lexique et des entités mentionnées. Afin que les connaissances obtenues soient comparables aux contextes de mentions lors de l'alignement, leur dérivation à partir des articles de Wikipedia doit correspondre à la représentation des dépêches de l'AFP adoptée dans cette tâche, décrite précédemment (table 5.6).

La collecte d'informations relatives aux entités est effectuée selon ce schéma à partir des éléments structurants des articles (cf. aussi la figure 3.4). Les connaissances ainsi rassemblées correspondent alors aux contextes c d'occurrences de mentions m, permettant une comparaison point à point entre des entités candidates e et m lors du processus d'identification. L'ensemble des articles de Wikipedia est considéré pour la construction de Nomos-KB ; cet ensemble distingue les articles concernant les entités recensée par Aleda, nommés ici e-articles, des articles dits généraux ou g-articles. Les éléments structurants d'articles pertinents pour cette collecte sont les suivants :

Titre Chaque article de Wikipedia est identifié par un titre, correspondant à un nom normalisé ou canonique pour les e-articles. En cas d'ambiguïté entre plusieurs sujets d'articles de même nom et donc entre articles de même titre, une propriété discriminante est indiquée entre parenthèses à la suite de ce nom. On trouve par exemple les titres *François Morel (acteur)* ou *Les Verts (France)*. Pour les entités concernées par les e-articles, cet élément parenthésé, noté e-`titlepar`, est souvent informatif dans la mesure où il indique par exemple la profession ou la qualité de personnalités (*chanteur*, *acteur*, *homme politique*), le secteur d'activité ou le type d'entreprises et d'organisations (*maison d'édition*, *informatique*). Le e-`titlepar` constitue alors un élément de contexte lexical revêtant un caractère descriptif particulièrement saillant pour l'entité concernée e ; on peut en effet supposer que la présence du e-`titlepar` d'une entité candidate e dans c constitue un trait positivement discriminant quant à la probabilité de dénotation de e par m.

Contenu textuel Le contenu de l'article lui-même constitue un contexte lexical ou vocabulaire associé à l'entité, de façon identique au contenu des dépêches pour les mentions d'entités.

Un traitement similaire à celui des dépêches est donc appliqué aux articles de Wikipedia (e-articles et g-articles) afin d'obtenir un sac de mots, pondéré en termes de nombre d'occurrences (ensemble E_{bow1}) et de saillance (ensemble E_{sbow1}). Chaque entité e d'e-articles est ainsi associée à un E_{bow1} et un E_{sbow1}, dont les éléments munis des scores les plus élevés en termes de test t peuvent être considérés comme des descripteurs pertinents de e. De façon similaire avec ce qui a été décrit pour le e-`titlepar`, les descripteurs les plus saillants peuvent ainsi contribuer à évaluer la probabilité d'alignement de m avec e selon que ces descripteurs sont ou non dans c.

Catégories Les articles de Wikipedia sont en grande majorité associés à des catégories d'ordre thématique ou descriptif et de granularité souvent très fine mais ne correspondant pas à un modèle sémantique défini. On trouve par exemple les catégories *Événement récent, Mathématicien du XX^{e} siècle, Vicomte (Belgique)* ou *Lauréat de la médaille Fields* associées à l'article concernant le mathématicien Pierre Deligne. Ces catégories ne constituent dont pas en tant que telles une information comparable à celles que présentent les dépêches de l'AFP avec les slugs. Elles font en revanche l'objet, pour leur utilisation dans Nomos-KB, d'une normalisation permettant d'obtenir des formes lexicales proches du e-`titlepar` ; les catégories suivantes :

> *Linguiste français, Sénateur du Nebraska, Militaire né à Metz, Architecture gothique aux Pays-Bas*

sont normalisées en :

> *linguiste, sénateur, militaire, architecture*

par une normalisation consistant à conserver uniquement le premier mot du terme désignant la catégorie en lettres minuscules. Les formes obtenues constituent alors, de façon similaire au statut du e-`titlepar`, des descripteurs relatifs aux entités e (ensemble E_{cats1}), pouvant s'avérer pertinents dans le processus d'identification par comparaison avec le contexte c.

Afin de permettre un rapprochement structurel des connaissances concernant les entités avec la forme des dépêches de l'AFP, les catégories d'articles ainsi normalisées peuvent être mises en correspondance avec les slugs décrits précédemment. Les catégories et slugs représentent en effet un type d'information de même nature, proche du principe des mots-clés et ne présentant pas de sémantique définie dans un modèle formel particulier. La correspondance entre catégories et slugs [6] a été concrètement obtenue par la mise en relation des termes identiques dans chacune des deux listes (*bourse* et *bourse*) ou par association manuelle des termes jugés synonymes dans ce cadre particulier (*sport* et *sport, sportif*). Certaines entités de Nomos-KB présentent ainsi, en plus d'un ensemble de catégories associées, un ensemble de slugs lorsque des correspondances sont possibles (ensemble E_{slugs1}). Lors du processus d'identification, ces catégories et slugs d'entité peuvent ainsi faire l'objet de comparaisons avec les slugs de c, afin d'établir le degré de similarité entre e et m à ce niveau [7]. La liste des catégories de Wikipedia pour lesquelles une relation de correspondance a été établie avec des slugs de l'AFP, au nombre de 378, est reproduite à l'annexe A, table A.1.

6. La liste des slugs AFP est définie en relation avec la taxonomie thématique de l'IPTC. Elle est reproduite à l'annexe A (tables A.1 à A.5).

7. Les slugs attribués aux dépêches par les journalistes ne sont pas limités à la liste fermée définie relativement à la taxonomie IPTC, tout terme jugé pertinent pouvant être utilisé. La comparaison entre slugs hors liste et catégories peut néanmoins être réalisée par association des termes identiques.

Wikilinks (1) Comme expliqué au chapitre 3, les mentions de sujets recensés dans Wikipedia au travers des articles de l'encyclopédie se présentent, au sein des contenus textuels, sous la forme de balises spéciales ou *wikilinks*, indiquant pour une mention donnée le lien interne de l'article dédié au sujet correspondant. L'article concernant le logicien Gottlob Frege mentionne ainsi Betrand Russel, auquel correspond également un article dans l'encyclopédie :

> [...] où il tente de dériver l'arithmétique de la logique, que
> `<a href="/wiki/Bertrand_Russell" title="Bertrand Russell">`**Russell**`</a>`
> lui fait parvenir [...]

Dans les cas où ces liens renvoient à des e-articles, les wikilinks en question constituent des mentions d'entités identifiées et sont ici notés e-wikilinks.

Dans chaque e-article, l'ensemble des e-wikilinks de l'article constitue, comme les mentions présentes dans les dépêches de l'AFP, une forme de représentation du document au même titre que le vocabulaire dérivé de son contenu textuel. Chaque entité e référencée par ces e-wikilinks, notée ewl, se présente comme un descripteur pertinent pour l'entité e faisant l'objet de l'article en question, à un niveau non seulement lexical mais également référentiel. On peut considérer, lors du processus d'identification, que la présence de ces ewl dans c, sous la forme des mentions correspondantes, augmente la probabilité d'alignement de m avec e. Dans les autres articles de l'encyclopédie Wikipedia (e-articles ou g-articles), un contexte du même type peut être constitué pour chaque entité faisant l'objet d'un e-wikilink. Nomos-KB recense ainsi toutes les co-occurrences d'e-wikilinks renvoyant à des entités d'Aleda. Chaque e est alors associée à ensemble d'entités qualifiées de *parentes*, au premier degré pour celles dont mention est faite au sein de l'article concernant l'entité considéré (ensemble E_{ewl1}), et au second degré pour les parentes observées en co-occurrence avec cette entité au travers de l'ensemble des articles de Wikipedia (ensemble E_{ewl2}).

Articles connexes Un certain nombre d'articles de Wikipedia référencent de façon explicite, dans une section spéciale, les articles dits *connexes*, dont le sujet est en relation particulièrement proche avec le sujet courant. On trouve par exemple des liens vers les articles concernant Bertrand Russel ou la philosophie du langage sous la rubrique « Articles connexe » de l'article au sujet de Gottlob Frege. 14 924 e-articles présentent entre 1 et 20 liens vers des e-articles connectés, qui sont recensés dans Nomos-KB en tant qu'entités parentes au premier degré pour chacune des entités concernées (ensemble E_{rel}), s'ajoutant ainsi aux co-occurrences dérivées des e-wikilinks (E_{ewl1} et E_{ewl2}).

Catégorisation thématique Les catégories d'articles utilisées dans Wikipedia ne sont pas aisément assimilables à des classes thématiques telles que celles proposées pour la publication journalistique par l'IPTC et se rapprochent davantage de mots-clés. Dans la même perspective d'un rapprochement structurel des connaissances concernant les entités avec la forme des dépêches de l'AFP, il est en revanche possible de munir les articles de Wikipedia, et par conséquent les entités concernées par les e-articles, d'informations relatives à la catégorisation thématique de l'IPTC : un modèle de classification obtenu par apprentissage supervisé à partir de corpus de l'AFP, pour lesquels les catégories IPTC sont indiquées, permet d'assigner à chaque article de Wikipedia (e-articles et g-articles) une ou plusieurs de ces catégories. Les modalités de l'acquisition de ce modèle par apprentissage automatique sont exposées à l'annexe A. Les entités d'e-articles disposent ainsi dans Nomos-KB d'informations thématiques (ensemble E_{iptc1}) comparables aux contextes d'occurrences des mentions à aligner lors du processus d'identification. La distribution de ces entités en termes de catégories IPTC est également donnée à l'annexe A.

Wikilinks (2) Les e-wikilinks apparaissant au travers des articles de Wikipedia (e-articles et g-articles) permettent également d'associer aux entités ainsi mentionnées certaines des informations propre à chacun de ces articles. Ceux-ci présentent en effet en tant que tels un vocabulaire pondéré notamment en termes de saillance, comme évoqué précédemment, ainsi que des informations relatives à leurs catégories, slugs et sujets IPTC, après application des traitements évoqués précédemment. Parallèlement aux entités parentes de E_{wl2}, les indications de saillance lexicale, catégories, slugs et sujets IPTC de cet article sont également associées à toute entité e faisant l'objet d'un e-wikilink. Ces indications viennent enrichir les connaissances de même type déjà collectées pour e dans le contexte de l'article qui lui est dédié — ces dernières étant considérées comme primaires (ensembles E_{bow1}, E_{sbow1}, E_{cats1}, E_{slugs1}, E_{iptc1}, E_{ewl1} et E_{ewl2}) — avec les ensembles E_{bow2}, E_{sbow2}, E_{cats2}, E_{slugs2}, E_{iptc2}, de statut secondaire.

Mentions Les occurrences d'une même entité e par le bias d'e-wikilinks au travers de l'ensemble des articles de Wikipedia (e-articles et g-articles) font intervenir différentes variantes lexicales dénotant e. Le nombre d'associations entre l'une de ces variantes et une entité e dans les e-wikilinks est reporté dans Nomos-KB. On dispose ainsi pour chaque entité e d'un ensemble de variantes (ensemble E_{vars}) pondéré en fonction du nombre d'emplois de chaque variante dans Wikipedia. On obtient également indirectement un nombre total d'occurrences de chaque entité, noté E_{freq}, pouvant être considéré comme le reflet d'une certaine popularité, à l'image de l'attribut *poids* intégré à Aleda — calculé quant à lui relativement à la taille de l'article concernant une entité. Les mentions sont également recensées en tant que telles, indépendamment des entités qu'elles dénotent ; le nombre d'occurrences d'une chaîne de caractères en tant que mention, noté M_{freq}, dans un corpus de référence tel que Wikipedia, peut en effet constituer un indicateur utile dans le processus de reconnaissance de mentions et de repérage des faux positifs.

La table 5.13 récapitule la nature des connaissances ainsi rassemblées dans Nomos-KB pour chaque entité de type PERSON et ORGANIZATION dans Aleda, à partir de l'article la concernant spécifiquement d'une part, et des autres articles de l'encyclopédie d'autre part (concernant ou non une autre entité d'Aleda). Les modalités d'utilisation de ces connaissances en relation avec les contextes d'occurrence des mentions à aligner seront précisées au chapitre 6, dans le cadre de la description fonctionnelle du système d'identification d'entités proposé.

Lieux Les entités de type LOCATION, importées dans Aleda à partir de GeoNames et non de Wikipedia, présentent quant à elles des connaissances relatives à leur emploi dans la production de l'AFP. Il est en effet possible d'établir une correspondance entre certaines métadonnées des dépêches de l'AFP et ces entités par le biais des attributs qui les caractérisent dans Aleda.

Comme cela a été évoqué précédemment, chaque dépêche est associée à un lieu de production, sous la forme d'un code ISO-3166-3 de pays noté $D_{country_code}$, ainsi que d'un nom de ville noté D_{city} et éventuellement de zone noté D_{area}. L'information $D_{country_code}$ peut être directement mise en relation avec l'entrée d'Aleda correspondante par le biais de son attribut ISO-3166-2, et donc avec les variantes lexicales recensées pour cette entité. On dispose donc pour une dépêche donnée non seulement d'un $D_{country_code}$, mais également d'une identification du pays concerné relativement à Aleda, l'entité correspondante étant notée $D_{country}$; l'ensemble des variantes lexicales correspondant à ce pays peut alors être associé à la dépêche et est noté $D_{country_vars}$.

Les métadonnées concernant les villes et zones (D_{city} et D_{area}) se présentent sous la forme de chaînes de caractères, interprétées comme des mentions d'entités. Ces entités peuvent figurer dans Aleda avec un sous-type correspondant (CITY pour D_{city} et REGION, COUNTRYDIVISION pour

ArticlesWikipedia	Connaissances	Description
e-**articles** (entités Aleda)	E_{bow1}	Vocabulaire (sac-de-mots) pondéré par nombre d'occurrences
	E_{sbow1}	Vocabulaire (sac-de-mots) pondéré par saillance (test t
	E_{cats1}	Catégories normalisées
	E_{slugs1}	Correspondances de E_{cats1} avec les slugs de l'AFP
	E_{iptc1}	Catégorisation thématique selon la taxonomie IPTC
	E_{rel}	Entités parentes au premier degré
	E_{ewl1}	Entités parentes directes (articles connexes)
e-**articles** **&** g-**articles**	E_{bow2}	Vocabulaire (sac-de-mots) secondaire pondéré par nombre d'occurrences
	E_{sbow2}	Vocabulaire (sac-de-mots) secondaire pondéré par saillance (test t)
	E_{cats2}	Catégories secondaires normalisées
	E_{slugs2}	Correspondances secondaires de E_{cats2} avec les slugs de l'AFP
	E_{iptc2}	Catégorisation thématique secondaire selon la taxonomie IPTC
	E_{ewl2}	Entités parentes au second degré
	E_{vars}	Variantes associées à l'entité, pondération par le nombre d'association dans Wikipedia
	E_{freq}	Nombre d'occurrences total de l'entité dans Wikipedia
	M_{freq}	Nombre d'occurrences total des mentions d'entités dans Wikipedia

Table 5.13 : Nomos-kb : connaissances rassemblées à partir de Wikipedia pour les entités de type person et organization.

D_{area}, cf. table 3.5 p. 94) mais ne sont cependant pas identifiables de façon directe dans la mesure où une même variante lexicale peut renvoyer à plusieurs entités, y compris de même type. Les mentions renseignées constituent donc un ensemble lexical, et non référentiel, venant s'ajouter à l'ensemble des variantes lexicales correspondant au pays renseigné ($D_{country_vars}$) pour former l'ensemble noté $D_{locvars}$.

Chaque dépêche est ainsi associée à un ensemble lexical spécifique aux informations de localisation, $D_{locvars}$, et à un pays, identifié relativement à Aleda ($D_{country}$). La BC Nomos-kb fait état des correspondances entre codes ISO-3166-3 et pays recensés dans Aleda. Lors du processus d'identification, ces informations peuvent être intégrées à la mesure de la similarité existant entre contextes c d'occurrences de mentions m et une entité candidate e, en supposant que la probabilité de dénotation de e par m est accrue si :

- e étant une entité de type location, le code ISO-3166-2 renseigné pour e dans Aleda correspond au pays renseigné par $D_{country}$,
- e étant une entité de type person ou organization, l'ensemble $D_{locvars}$ de c présente une intersection significative avec les ensembles E_{bow1}, E_{bow2} (ou E_{sbow1} et E_{sbow1}) de e,
- e étant une entité de type person, organization, l'ensemble $D_{locvars}$ de c présente une intersection significative avec les variantes lexicales recensées pour les entités des ensembles

E_{rel}, E_{ewl1} et E_{ewl2}.

Il est utile d'observer que les informations de localisation données pour une dépêche sont propres au lieu de rédaction, qui n'est pas systématiquement identique au lieu de déroulement de l'événement médiatique reporté (cf. chapitre 4, section 2.3). La similarité pouvant être observée à ce niveau relève doit donc donner lieu à une interprétation prudente, ce dont il sera discuté au chapitre 6.

3 Ressources : Outils

La mise en œuvre de l'approche modulaire jointe proposée pour l'identification automatique d'entités dans les données de l'AFP se traduit, au niveau du module de reconnaissance des mentions décrit précédemment (cf. *infra*, section 1), par l'utilisation d'outils existants développés pour la Reconnaissance d'Entités Nommées (REN). La tâche d'identification en elle-même, pour laquelle le système spécialisé Nomos est développé (chapitre 6), a par ailleurs fait l'objet de travaux antérieurs dans le même cadre collaboratif — AFP et équipe-projet Alpage —, à partir desquels un système initial d'identification, principalement destiné à jouer le rôle de *baseline* a été mis au point.

3.1 Reconnaissance d'Entités Nommées

La configuration découlant de l'approche proposée ici pour l'identification automatique d'entités implique qu'un certain nombre de décisions concernant le niveau de segmentation du texte d'entrée en mentions d'entités soient reportées à l'étape de Liage. L'analyse en termes de REN est alors envisagée sous une forme non déterministe, dans laquelle certaines ambiguïtés usuellement levées dans le résultat final des systèmes de REN sont maintenues.

Des résultats de cette forme peuvent être obtenus par l'utilisation d'un système formalisant de façon inhérente ou indirecte l'ambiguïté, ou par l'emploi combiné de plusieurs systèmes. Leurs sorties sont dans ce cas regroupées par les opérations d'union — tous les résultats étant alors considérés — ou d'intersection — seuls les résultats commun aux différents systèmes sont pris en compte dans la suite des traitements. L'intersection peut ici être envisagé comme un premier filtrage de faux positifs, le consensus de plusieurs procédures distinctes étant vu en relation avec la probabilité de correction du résultat en question.

Le choix des systèmes adoptés repose également sur leur caractère libre et disponible, en association avec les possibilités d'adaptation qu'ils présentent en termes de paramétrage et de configuration relativement à un cadre applicatif déterminé.

Les deux systèmes faisant l'objet d'une intégration au module de reconnaissance pour la tâche d'identification répondent à ces critères, l'un faisant usage de techniques symboliques, l'autre reposant sur des techniques numériques et d'apprentissage automatique. Ces deux orientations sont considérées comme performantes dans le cadre de la REN, les approches symboliques faisant montre d'une plus grande efficacité en termes de précision, à l'inverse des approches numériques qui s'avèrent plus robustes relativement à des données bruitées et meilleures en termes de rappel, comme cela nous l'avons souligné dans [BSS11].

3.1.1 SxPipe/NP

Le premier système intégré au module de reconnaissance, SxPipe/NP, repose sur une approche symbolique de la REN et fait partie de la chaîne d'analyse linguistique de surface SxPipe [8]. L'ana-

8. SxPipe, comme l'ensemble des outils développés dans le cadre de la chaîne de traitement linguistique de l'équipe-projet Alpage, est disponible sous licence Cecill-C, compatible avec LGPL et accessible *via* le projet `lingwb`

lyse de données par SxPipe vise les corpus textuels bruts, dans l'objectif d'un prétraitement utile à des modules d'analyse plus profonde tels que les analyseurs syntaxiques. Les principes, l'architecture ainsi que le fonctionnement de SxPipe sont décrits par Sagot et Boullier [SB08]. La chaîne SxPipe est de type modulaire et présente trois aspects principaux de traitement des données textuelles : segmentation en phrases et mots, correction orthographique et reconnaissance d'entités nommées, celles-ci étant entendues en un sens très large (dates, heures, quantités, adresses email et URL, etc.). La modularité de SxPipe correspond à l'application séquentielle de traitements à plusieurs niveaux, les décisions prises par chaque module pouvant affecter les suivants. Afin d'éviter des analyses erronées pouvant se propager à l'ensemble de la chaîne, plusieurs modules, dont le module de REN, sont capables de produire des sorties ambigues, ainsi que de lire des entrées du même type.

Les trois aspects de traitement pris en charge par SxPipe se traduisent par les groupes de modules correspondant :

- Un premier ensemble de modules transforme l'entrée sous forme textuelle en flux de tokens, ceux-ci étant définis comme des séquences de caractères séparés des autres par l'espace ou un signe de ponctuation ; cette segmentation, qui concerne également les frontières de phrases, attribue à chaque token un identifiant de position unique et est déterministe. Cette première phase de traitement est également chargée de la reconnaissance des entités nommées entendues au sens large et dont les caractéristiques peuvent être définies au niveau des caractères ; il s'agit notamment des URL, adresses postales et électroniques, dates, horaires, nombres, etc.

- Le flux de tokens issu de la segmentation est transformé en graphe de formes, celles-ci étant définies comme des unités linguistiques syntaxiquement atomiques ou élémentaires, c'est-à-dire non analysables syntaxiquement ; la définition des formes, complexe et discutable selon les contextes, peut être ramenée par convention à un critère de présence ou d'absence de l'unité considérée dans une ressource lexicale de référence pour la question. Les formes sont simples ou composées selon qu'elles correspondent à un token unique ou à plusieurs tokens, respectivement. Des formes spéciales peuvent être définies dès lors qu'elles ne doivent pas faire l'objet d'une analyse syntaxique ou qu'elles constituent un élément informatif en tant que telles pour l'analyse syntaxique. Le graphe produit par SxPipe, qui modélise les ambiguïtés de reconnaissance des formes composées ainsi que des formes issues de la correction orthographique, se présente sous la forme d'un DAG (*directed acyclic graph*, graphe orienté acyclique) de formes, qui peut être écrit en tant qu'expression régulière :

  ```
  ({pomme} pomme {de} de {terre cuite} terre_cuite | ({pomme de terre}
  pomme_de_terre | {pomme} pomme {de} de {terre} terre) {cuite} cuite)
  ```

  ou sous une forme dépliée avec identification des transitions entre formes [9] :

  ```
  ##DAG BEGIN
   1 {pomme} pomme  2
   1 {pomme de terre} pomme_de_terre  4
   2 {de} de  3
   3 {terre} terre  4
   3 {terre cuite} terre_cuite  5
   4 {cuite} cuite  5
   ##DAG END
  ```

sur INRIAGforge. Sources du projet `lingwb` : `https://gforge.inria.fr/projects/lingwb/`. Site internet : `http://lingwb.gforge.inria.fr/`.

9. Des identifiants uniques sont associés à chaque token distinct, permettant de distinguer les tokens de contenu identique. Ils ne figurent pas dans ces exemples afin d'alléger la reproduction du format.

- À partir des DAG de formes, SxPipe permet la reconnaissance de motifs définis par l'utilisateur sous la forme de grammaires locales non contextuelles, produisant elles-mêmes des DAG de formes, possiblement ambigus. Chaque type de motifs à reconnaître fait l'objet d'un module du groupe nommé `dag2dag` et définit une grammaire ainsi qu'un analyseur lexical ou un lexique destinés à une analyse syntaxique réalisée par l'analyseur SYNTAX [BD88].

SxPipe peut être configuré, par activation et paramétrage sélectifs des modules des différents niveaux de traitement, selon les besoins particuliers au traitement d'un corpus donné, la langue concernée par les données et l'ensemble des modules de reconnaissance de motifs non contextuels définis. L'application de SxPipe pour une langue donnée requiert cependant l'installation du lexique Alexina correspondant [10], tel que le Le*fff* [Sag10] pour le français.

La REN intégrée à SxPipe se présente sous la forme d'un module, nommé NP, au sein de `dag2dag`, autrement dit d'une grammaire locale associée à un lexique spécialisé. Celui-ci consiste en une dérivation de la base d'entités Aleda, présentée précédemment (chapitre 3, section 1.3 et *infra*, section 2.2), plus précisément de l'ensemble des variantes lexicales définies dans Aleda pour les entités recensées. Il s'agit donc d'un lexique typé, chaque forme étant associée à une ou plusieurs étiquettes parmi PERSON, ORGANIZATION et LOCATION. Une centaine de règles, pour certaines munies d'indications de priorité, sont définies dans la grammaire et se répartissent en trois catégories :

- correspondance exacte avec un terminal typé du lexique,
- indices contextuels autour d'un terminal typé du lexique,
- indices contextuels autour d'une forme non recensée dans le lexique et présentant des indices internes pertinents.

La REN fournie par NP est ainsi centrée sur les entités nommées sous la forme de noms propres en adéquation avec les types sémantiques visés dans la tâche d'enrichissement des contenus textuels de l'AFP. Le troisième type de règles permet en outre d'obtenir des analyses en termes de mentions d'entités correspondant à des variantes non recensées par Aleda. Cette fonctionnalité joue un rôle important dans la prise en charge du cas spécial de dénotation NIL et illustre l'importance, dans une tâche telle que l'Annotation Sémantique, d'un outil relevant de l'Extraction d'Information et non limité à l'application directe de lexiques sur les données à traiter.

Les résultats de l'analyse réalisée par SYNTAX consistent en une forêt partagée, dans lesquelles plusieurs analyses concurrentes, notamment en termes d'imbrication, de croisement et d'ambiguïté sont représentées indépendamment les unes des autres. La forêt peut être filtrée :

- par élimination des analyses ne comportant aucun motif reconnu, si d'autres analyses concurrentes comportent un motif reconnu,
- par retenue des analyses correspondant à l'application de règles prioritaires en cas de concurrence entre plusieurs empans de texte de même longueur en termes de formes,
- par retenue des analyses de longueur maximale entre plusieurs motifs reconnus entre un état e_n et différents états $e_{m1}, \ldots, e_{mN}$,
- par élimination des analyses faisant intervenir des sous-motifs relativement aux analyses concurrentes.

10. Alexina est une architecture linguistique et informatique pour le développement de lexiques morphologiques et syntaxiques. Site internet `http://alexina.gforge.inria.fr/`

Un DAG de formes est alors construit par extraction des séquences de feuilles de la forêt filtrée, puis est minimisé. Les informations relatives aux noms propres reconnus et typés sont indiquées au niveau des formes du DAG, comme l'illustre la figure 5.8.

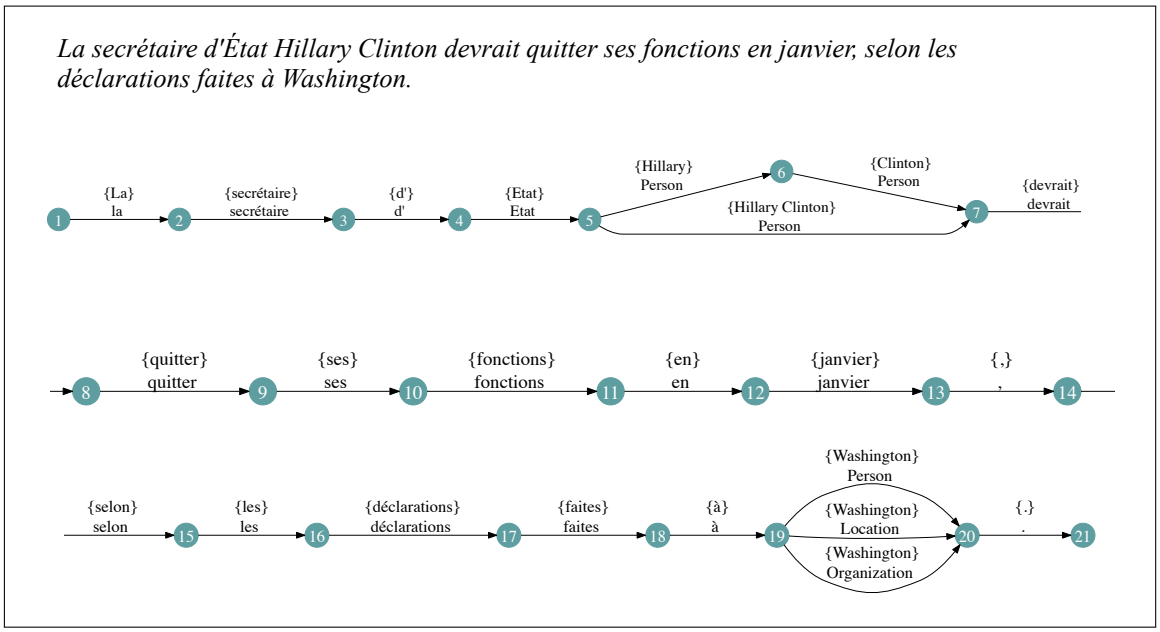


Figure 5.8 : DAG de formes produit par SxPipe/NP.

On peut noter que l'on obtient, en sortie de SxPipe, un format de représentation des données équivalent au format décrit précédemment pour l'approche modulaire jointe (cf. *supra*, section 1.2.1) : seules les ambiguïtés d'analyses issues du module NP sont considérées, à l'exclusion de celles concernant les autres formes des DAG[11], et les zones d'ambiguïté des DAG sont normalisées en disjonction de chemins de même longueur, avec un seul état initial et un seul état final. La configuration de SxPipe élaborée pour la présente tâche fait donc appel aux composants suivants :

- sélection du français comme langue d'analyse, en association avec le lexique Le*fff*,
- segmentation en phrases et en tokens,
- formation des DAG de formes, éventuellement ambigus,
- reconnaissance des noms propres de type person, organization et location par le module de `dag2dag` NP

puis élimine les analyses relatives aux formes en dehors des analyses retournées par NP. Le format final consiste ainsi en un DAG de tokens et de formes, celles-ci ne concernant que les noms propres reconnus par NP et conservant les informations relatives aux tokens qui les constituent.

En tant que module intégré à SxPipe, NP ne présente pas de fonctionnalité de désambiguïsation. Dans le cadre de travaux initiaux réalisés à partir de ce module de REN pour le traitement de corpus de l'AFP, un module spécial, npNormalizer, a été développé afin de retourner une analyse univoque en termes de REN à la suite de NP ; il est décrit ci-après, dans la section 3.2.

11. Les autres ambiguïtés à ce niveau concernent notamment la segmentation en unités lexicales, où la séquence de tokens `de la` peut être segmentée de façon ambigüe, la lecture *de_la* (déterminant partitif) étant en concurrence avec la lecture *de la* (préposition suivie d'un déterminant défini).

3.1.2 LIANE

Le système de REN LIANE [BC10], développé dans le cadre de la campagne d'évaluation ESTER (cf. chapitre 2, section 3.1), repose sur une approche hybride avec l'emploi de deux modèles, génératif et discriminant, pour l'apprentissage automatique à partir du corpus ESTER. LIANE est construit autour des deux sous-tâches essentielles de la REN, consistant d'une part en une segmentation du texte d'entrée relativement aux frontières des entités nommées et d'autre part en une classification sémantique de ces entités nommées, selon une typologie ramenée ici aux types PERSON, ORGANIZATION et LOCATION. La réalisation de ces deux sous-tâches peut être vue de façon jointe si l'on considère la classification comme un processus appliqué au mots ou tokens. Chacun d'eux reçoit alors une étiquette de type et une étiquette du modèle BIO, indiquant sa position au sein de la séquence formant une entité nommée — étiquette *B* (*begin*) pour la position initiale et I (*inside*) pour les autres —, les mots ou tokens n'appartenant pas à une telle séquence recevant l'étiquette vide et la position notée *O* (*outside*).

La REN ainsi considérée comme une tâche d'étiquetage de mots peut faire usage de toutes les méthodes développées pour l'étiquetage en parties du discours, notamment numériques. Les approches principales en la matière sont de type génératif, avec notamment les Modèles de Markov Cachés (*Hidden Markov Models*, HMM) [BSW99] ou discriminant avec notamment les modèles a maximum d'entropie [Bor+98] ou les Champs de Markov Aléatoires (*Conditional Random Field*, CRF) [ML03].

L'approche présentée par LIANE est hybride en tant qu'elle met en œuvre, successivement :

- un processus génératif à l'aide d'un modèle HMM pour la prédiction d'étiquettes syntaxiques (parties du discours) et sémantiques (type parmi PERSON, ORGANIZATION et LOCATION dans la configuration présente) au niveau de chaque mot ou token,
- puis un processus discriminant reposant sur les CRF, chargé de déterminer les frontières des expressions formant des entités nommées sur le modèle BIO, et d'assigner un type global à la séquence ainsi identifiée.

Le format des données d'apprentissage se présente donc selon le modèle BIO, comme l'illustre la figure 5.9.

```
investiture     NFS       O
aujourd'hui     ADV       B-TIME
à               PREPADE   O
Bamako          LOCATION  B-LOCATION
Mali            LOCATION  B-LOCATION
du              PREPDU    O
président       NMS       O
Amadou          PERSON    B-PERSON
Toumani         PERSON    I-PERSON
Touré           PERSON    I-PERSON
```

FIGURE 5.9 : Format BIO de représentation des données d'apprentissage de LIANE

L'étiquetage des mots ou tokens à l'aide d'un modèle HMM préalablement à l'étiquetage final à l'aide de CRF est motivé par

- la facilité d'intégration de multiples sources d'information dans les estimations de probabilité des mots sachant les classes,
- la tolérance des HMM au bruit dans les données d'apprentissage, dès lors que la fréquence relative des événements modélisés est respectée,

- la désambiguïsation partielle de la chaîne par l'étiquetage en parties du discours et en types sémantiques, qui permet de simplifier la tâche l'étiquetage du modèle CRF, concentré sur les problèmes de frontières et de classification globales.

L'emploi d'un modèle numérique tel que les CRF permet d'obtenir un ensemble d'analyses, chacune étant munie d'une probabilité quant aux frontières et au typage des entités nommées reconnues et un même segment pouvant être concerné par plusieurs analyses en concurrence. Les sorties du système LIAne correspondent à des DAG qu'il est ensuite possible de normaliser comme cela est fait pour les DAG de SxPipe/NP. Chaque DAG ainsi obtenu peut par ailleurs être considéré comme associé à une pondération, chaque chemin d'analyse en concurrence avec d'autres recevant un score correspondant à la probabilité émise par le modèle CRF de reconnaissance. La désambiguïsation des résultats de LIAne peut enfin être réalisée par sélection de l'analyse ayant reçu le score maximal.

3.2 Identification initiale et baseline

Dans le cadre de travaux initiaux menés en collaboration entre l'équipe-projet Alpage et l'AFP, la chaîne de traitement SxPipe a été utilisée afin de prendre en charge les besoins préliminaires d'analyse et d'exploration de corpus, conduisant notamment au développement effectif du module SxPipe/NP. Celui-ci ne présentant cependant pas de fonctionnalité de désambiguïsation et se limitant à une analyse surfacique propre à la REN, un module *ad hoc* a été mis au point afin de rendre NP utilisable à court terme à des fins d'Extraction d'Information. Le nom de ce module, npNormalizer, reflète son objectif de départ : d'abord conçu dans le but de désambiguïser les sorties de NP et d'effectuer une normalisation des entités nommées — principalement des noms de personnes — au niveau des documents, npNormalizer a ensuite intégré une fonctionnalité équivalente au Liage tel que défini au chapitre 3 (section 3) afin de répondre à la nécessité prégnante d'associer la REN de NP à une Annotation Sémantique reposant sur des ressources d'entités identifiées [SS10]. On dispose donc, avec SxPipe et npNormalizer, d'une chaîne initiale accomplissant l'identification des entités dans les corpus de l'AFP, que l'approche et le système étudiés dans le présent travail proposent d'améliorer.

Cette amélioration est souhaitée pour plusieurs raisons :

- Le module npNormalizer est, au niveau formel, étroitement associé à SxPipe et à NP, et ne peut prendre en charge l'analyse de sorties d'autres systèmes de REN. L'approche proposée ici est quant à elle modulaire et s'intéresse à la possibilité d'utiliser tout système de REN considéré comme performant et adapté à la tâche.

- Le fonctionnement de npNormalizer repose sur un ensemble d'heuristiques, simulant partiellement les critères d'analyse généralement envisagés par le Liage. Ces heuristiques traduisent l'intuition générale selon laquelle les entités les plus notables ou populaires ont une probabilité plus grande d'être effectivement dénotées, étant donnés une mention et un ensemble d'entités candidates. Ce critère est cependant exploité de façon isolée et favorise ainsi systématiquement les entités les plus populaires *a priori*. Une telle stratégie permet, dans le cadre de travaux initiaux et exploratoires, de prendre en compte la distribution particulière des entités dans les contenus journalistiques généralistes de l'AFP : la majorité des mentions d'entités correspondent en effet usuellement, dans de tels corpus, à un ensemble réduit d'entités à forte notoriété. Une désambiguïsation presque uniquement fondée sur la notoriété fournit ainsi une couverture non négligeable et une bonne précision des liens vers les entités dénotées. On peut cependant observer qu'une telle procédure répond de façon limitée à la tâche à réaliser, qui appelle la mise en œuvre d'une approche plus sophistiquée

et à même de prendre en charge le phénomène dénotationnel dans son ensemble. Le fonctionnement général ainsi que l'évaluation qui a pu être faite du module NPNORMALIZER sont décrits ci-après.

- Le développement de NPNORMALIZER est très fortement guidé par une adéquation aux données de l'AFP. Plus précisément, les retours réguliers des utilisateurs en termes d'erreurs donnent lieu à des corrections immédiates et locales, souvent apparentés à des suppressions et ajouts manuels de règles spécifiques aux cas particuliers indiqués par les rapports d'erreurs, sans prise en compte au niveau plus général du mode d'analyse lui-même. D'autre part, NPNORMALIZER reflète de façon systématique le corpus GAFP (cf. *supra*, section 2.1) qui est utilisé pour son développement et son évaluation.

- On peut considérer qu'un système reposant sur une approche numérique du phénomène à traiter devrait être plus robuste, généralisable et adaptable qu'un système reposant sur des heuristiques tel que NPNORMALIZER.

Fonctionnement Comme dans l'approche proposée ici, NPNORMALIZER est associé à la base d'entités Aleda, ainsi qu'aux variantes qu'elle définit pour chaque entité recensée. De façon comparable au Liage tel qu'envisagé pour cette approche (cf. *infra*, section 1), un DAG correspondant à un automate modélisant chaque phrase à traiter est présenté à NPNORMALIZER, qui considère dans un premier temps toutes les mentions d'entités y figurant, telles que retournées par SxPipe/NP, et procède à leur Liage. Cette étape résulte en une entité choisie pour chaque mention, le cas NIL étant également considéré. Chaque zone d'ambiguïté concernant des mentions est ensuite désambiguïsée par sélection du chemin contenant les entités les plus probablement dénotées par les mentions correspondantes, relativement aux entités choisies dans les autres chemins.

La première étape de traitement de NPNORMALIZER, conformément au cadre général du Liage, établit en premier lieu un ensemble d'entités candidates pour chaque mention par la mise en correspondance du contenu de ses tokens avec les variantes lexicales recensées dans Aleda. Leur ordonnancement repose ensuite sur un ensemble d'heuristiques traduisant le rôle primordial accordé au critère de popularité dans la résolution du phénomène dénotationnel. Ces heuristiques consistent principalement à assigner un poids à chaque candidat sous la forme d'un score numérique. Ce score est calculé à partir du poids indiqué par Aleda pour une entité donnée, soumis à une normalisation permettant de rendre comparables la popularité des entités issues de Wikipedia, correspondant à la taille de l'article correspondant, et le nombre d'habitants renseignés pour les lieux issus de GeoNames. Lorsque l'ensemble des candidats retourné par la requête sur Aleda est vide, l'entité spéciale NIL est assigné à la mention avec un score fixe. Le cas NIL n'est donc pas envisagé pour toutes les mentions à lier.

À partir du DAG décoré des informations de Liage pour les mentions, chacune d'elles étant associée à une entité la plus probablement dénotée, les différents chemins en situation d'ambiguïté dans chaque zone concernée par des mentions font l'objet d'une pondération permettant de désambiguïser la zone par sélection du chemin au score le plus élevé. Ce score est calculé à partir de celui des entités choisies pour chaque mention, modifié en tenant compte du statut des mentions. Ainsi, les entités choisies pour des mentions :

- apparaissant en tête de phrase,

- figurant sur une liste préétablie de façon manuelle et selon des jugements liés à la probabilité forte d'une chaîne de caractère de ne pas constituer une mention d'entité,

- figurant en tant qu'entrée d'un lexique général — ici le Le*fff*,

- dont le type est ORGANIZATION

voient leur score diminuer. À l'inverse, les entités choisies pour des mentions de type LOCATION sont favorisées, ainsi que les entités dont le sous-type indiqué par Aleda est CITY, CAPITAL ou ADMINISTRATIVEDIVISION.

NPNORMALIZER est par ailleurs directement appliqué en sortie de SxPipe/NP et dispose d'informations spécifiques à certaines des règles ayant donné lieu aux analyses retournées. Les analyses résultant de règles considérées comme plus sures que les autres, pour une même zone d'ambiguïté, sont favorisées. Il s'agit notamment de règles faisant intervenir des indices multiples : indices contextuels, indices internes et intervention d'une variante lexicale recensée dans Aleda. De telles informations ne sont pas transmises en sortie de SxPipe/NP dans le cadre du système développé ici : elles induiraient en effet des annotations différentes selon les règles appliquées, en contradiction avec la généralité attendue de tout module de REN employé à ce niveau. Il importe en effet que les analyses retournées par les différents modules de REN éventuellement combinés soient comparables.

Il est important de noter que toutes les valeurs, assignations et variations de scores d'entités dans NPNORMALIZER sont établies manuellement.

Évaluation Le module NPNORMALIZER est évalué à l'aide du corpus de référence GAFP, présenté précédemment. Comme cela devra être le cas pour le système d'identification automatique proposé dans la suite de ce travail, l'évaluation porte à la fois sur la qualité du Liage et sur la reconnaissance des mentions d'entités en tant que telles, celle-ci étant en partie dépendante des choix effectués lors de la seconde étape. Les métriques utilisées correspondent donc

- d'une part à la précision, au rappel et à la F-mesure usuellement appliquées à la REN et notées P_{REN}, R_{REN} et F_{REN}, avec

$$F_{\text{REN}} = \frac{2 \times P_{\text{REN}} \times R_{\text{REN}}}{(P_{\text{REN}} + R_{\text{REN}})}$$

- d'autre part au taux de correction de l'étape de Liage étant donné l'ensemble des mentions correctement reconnues, noté Acc. Acc est égal au nombre de mentions correctement reconnues et identifiées (cas NIL inclus) divisé par le nombre de mentions correctement reconnues.

- De ces deux types de mesures sont dérivés une précision, un rappel et une F-mesure de la tâche globale, notées P_{ALL}, R_{ALL} et F_{ALL}, comme cela sera discuté plus en détail lors de l'évaluation du système présenté au chapitre 6, avec

$$P_{\text{ALL}} = P_{\text{REN}} \times Acc$$
$$R_{\text{ALL}} = R_{\text{REN}} \times Acc$$

  et

$$F_{\text{ALL}} = \frac{2 \times P_{\text{ALL}} \times F_{\text{ALL}}}{(P_{\text{ALL}} + R_{\text{ALL}})}$$

Les résultats de l'évaluation, reportés à la table 5.14 [12], montrent que la base heuristique de NPNORMALIZER est efficace et constitue une baseline déjà élevée. Ces bonnes performances ne doivent cependant pas occulter les insuffisances de NPNORMALIZER évoquées ci-dessus, notamment en

12. L'évaluation de la REN tient compte de la correction des frontières de mentions d'entités, les correspondances partielles étant considérées comme des faux positifs. La correction du type est en revanche ignorée au niveau de la REN, pour être intégrée à la performance de Liage : une mention correctement liée reçoit en effet le type de l'entité choisie, qui se substitue au type assigné par le module de REN.

termes de possibilités de généralisation : le développement de ce module est en effet étroitement lié à la connaissance du corpus de référence GAFP, ainsi qu'aux réponses immédiates à apporter en termes de correction d'erreurs à court terme pour les usages exploratoires des contenus de l'AFP [13]. On peut également constater, comme cela a été évoqué à la section 1.1, que de bons résultats au niveau du Liage et de la REN aboutissent à un score global moins satisfaisant (en-deçà de 75%).

P_{REN}	R_{REN}	F_{REN}	Acc	P_{ALL}	R_{ALL}	F_{ALL}
87,75%	78,22%	82,71%	88,78%	77,90%	69,44%	73,43%

TABLE 5.14 : Résultats d'évaluation de NPNORMALIZER sur l'ensemble du corpus de référence GAFP.

Au-delà des résultats d'évaluation reposant sur des mesures numériques usuelles en la matière, il convient d'explorer la typologie des erreurs générées par NPNORMALIZER en termes de reconnaissance de mentions et de Liage afin de déterminer. Ces erreurs sont en effet régulières, du fait de la nature systématique des manipulations effectuées aux deux niveaux, et peuvent, même en nombre absolu réduit, provoquer un effet de bruit faisant de NPNORMALIZER une solution non satisfaisante au problème de l'identification d'entités. Le système présenté dans le chapitre suivant, Nomos, pourra être comparé à NPNORMALIZER à la fois au niveau des scores obtenus quant aux métriques d'évaluation de la tâche d'identification (performances de la REN, taux de correction du Liage et performances au niveau global), mais également relativement à des tâches applicatives, abordées au chapitre 7,

13. Il convient également d'observer que ces résultats sont difficilement comparables à l'état de l'art en REN pour le français, tel que la campagne ESTER 2 a notamment permis de l'établir ; les données d'évaluation de NPNORMALIZER sont en effet limitées à des dépêches d'agence, tandis que les corpus d'ESTER 2 sont constitués de transcriptions de l'oral, également du domaine journalistique mais *a priori* plus difficiles à traiter.

Chapitre 6

Un système d'identification d'entités : Nomos

Le système Nomos procède à l'identification automatique d'entités selon l'approche proposée au chapitre précédent et dans la perspective d'une utilisation pour l'enrichissement des contenus de l'AFP. Afin d'expliciter les hypothèses à partir desquelles le développement de Nomos est envisagé, la section 1 propose une analyse des différentes configurations de l'identification, de la situation décrite pour la tâche de Liage dans le cadre de TAC-KBP (chapitre 3, section 3) à celle du traitement de contenus textuels bruts dont il s'agit dans le présent travail. La section 2 présente les composants de Nomos, conçus et agencés de façon à prendre en charge les aspects de l'identification spécifiques à un tel traitement. La section 3 fait état des expériences menées afin de tester la validité et l'efficacité de l'approche choisie relativement à cette configuration.

1 Configurations de l'identification

Nous proposons de qualifier de *naïve* la configuration de l'identification dans laquelle toute requête présentée au système est considérée comme une mention d'entité, sans que ce statut ne soit remis en cause, par opposition à une configuration *informée*, dans laquelle le caractère dénotationnel des requêtes peut être infirmé.

1.1 Configuration naïve

La configuration naïve est inhérente à la tâche de Liage telle que définie dans le cadre de TAC-KBP, mais peut également concerner le cas de traitement de contenus textuels bruts où une étape de Reconnaissance d'Entités Nommées est envisagée.

1.1.1 Cas TAC-KBP étendu

L'identification d'entités dans le cadre de TAC-KBP se ramène à la tâche de Liage, dans laquelle une mention par document du corpus traité est soumise en tant que requête, qu'il s'agit d'aligner avec l'une des entités recensées dans la base de connaissances (BC) associée à la tâche ou avec le cas spécial NIL, lorsque que l'entité dénotée est absente de la BC. Nous étendons ce cas d'identification en considérant comme requête toutes les mentions d'entités présentes dans un document à traiter et non plus une seule.

Dans ce cas d'identification, les mentions à aligner sont *données* : aucune procédure particulière n'est requise pour leur obtention en vue de la réalisation de la tâche et toutes sont

effectivement dénotationnelles. Elles sont en pratique issues des données de référence constituées pour la tâche. Le Liage est alors mis en œuvre selon la méthode générale présentée au chapitre 3 (section 3.2) :

Sélection des candidats Un sous-ensemble d'entités de la BC est constitué afin de limiter l'espace de recherche pour l'alignement de chaque mention. Le cas NIL peut y être directement intégré ou considéré dans un second temps.

Ordonnancement des candidats Les entités candidates à l'alignement d'une mention sont ordonnées de façon à retourner l'entité dénotée par la mention au premier rang. Le rang d'un candidat est déterminé par un ensemble de facteurs, principalement sa similarité avec la mention et son contexte. Le cas NIL peut être retourné au rang 1, de façon directe par l'ordonnancement ou indirecte par invalidation de l'entité de la BC placée au rang 1.

On suppose dans ce cas que la dénotation peut être modélisée par des éléments de similarité contextuelle entre entités et mentions. Les différents facteurs usuellement pris en compte dans les méthodes présentées au chapitre 3 (section 3.3) peuvent à ce titre être intégrés à l'apprentissage d'un modèle qui devrait donner lieu à des résultats d'identification comparables à l'état de l'art rapporté dans le cadre de TAC-KBP. L'évaluation correspondante est de même type que celle en vigueur pour TAC-KBP : il s'agit d'une mesure d'exactitude (Acc), établie à partir du rapport entre les alignements corrects et l'ensemble des alignements à réaliser, telle que :

$$Acc = \frac{\text{nombre de mentions (requêtes) correctement alignées}}{\text{nombre total de mentions (requêtes)}}$$

dans laquelle on peut distinguer

$$Acc_{BC} = \frac{\text{nombre de mentions (requêtes) correctement alignées}}{\text{nombre total de mentions (requêtes) dénotant des entités de la BC}}$$

et

$$Acc_{NIL} = \frac{\text{nombre de mentions (requêtes) correctement alignées}}{\text{nombre total de mentions (requêtes) dénotant des entités hors BC (NIL)}}$$

1.1.2 Cas naïf sur corpus bruts

La configuration disposant des mentions à aligner n'est pas transposable à une conduite de la tâche d'identification dans des conditions plus réalistes, où l'entrée du système consiste en du texte brut et où une étape de REN est donc nécessaire à l'obtention des mentions. Cette REN retourne usuellement une analyse déterministe du texte donné en entrée, indiquant les frontières et le type des segments correspondant à des mentions d'entités. Les types considérés ici sont PERSON, ORGANIZATION et LOCATION. Ces résultats de REN sont potentiellement imparfaits et lacunaires.

On peut envisager l'application d'un modèle équivalent à celui du cas précédent pour l'identification de ces mentions obtenues par voie automatique. On suppose alors dans ce cas *naïf* qu'un tel modèle ne remet pas en cause le statut dénotationnel de ces requêtes et ignore ainsi les possibles erreurs d'analyse du module de REN ; la tâche d'identification peut alors présenter des erreurs relevant de la précision (liens établis sur des segments non dénotationnels)[1], ce qui n'est pas le cas dans la configuration de TAC-KBP. L'évaluation correspondante comporte plusieurs volets :

1. Des erreurs de rappel de la REN peuvent également entraîner des erreurs du même ordre (silence) au niveau de l'identification. Elles ne sont pas prises en compte ici mais seront discutées dans la suite de ce travail.

Reconnaissance d'Entités Nommées Cette étape peut être évaluée selon les métriques usuelles dans cette tâche : précision (P_{REN}), rappel (R_{REN}), F-mesure (F1, F_{REN}), tels que :

$$F_{REN} = \frac{2 \times P_{REN} \times R_{REN}}{(P_{REN} + R_{REN})}$$

Le calcul de la précision considère comme correctement reconnues les mentions dont les frontières sont exactement identiques à celles des données de référence : les correspondances partielles sont donc traitées comme des faux positifs. Symétriquement, les mentions des données de référence se trouvant alors sans correspondance dans les résultats du système diminuent d'autant le taux de rappel. Le type associé aux mentions par les systèmes de REN ne sont en revanche pas intégrés au calcul de la précision : la REN étant ici intégrée à une tâche d'identification, c'est au niveau des entités que la correction du type peut être évaluée ; elle l'est en fait de façon implicite et concomitante à l'évaluation des entités elles-mêmes, qui sont associées à un type dans les ressources utilisées. La correction du type est ici inhérente à la correction de l'alignement.

Liage L'exactitude des alignements peut être établie à partir d'un rapport du même type que dans le cas TAC-KBP (mesure Acc), mais les ensembles de calcul de ce rapport sont différents : cette exactitude ne peut en effet être définie que sur l'ensemble des mentions retournées par le module de REN et effectivement dénotationnelles, et non sur les faux positifs. L'évaluation d'un même modèle sur un corpus disposant des mentions et sur un corpus où les mentions sont obtenues automatiquement, si ces dernières sont différentes des mentions de référence, donne donc deux mesures d'exactitude qui ne sont pas strictement comparables.

Tâche globale d'identification Les mesures d'évaluation de la REN et du Liage ne permettent pas à elles seules de juger la performance de la tâche globale d'identification. En effet, une très bonne exactitude de Liage peut être atteinte sur un ensemble réduit de mentions correctement détectées ; à l'inverse, la REN peut donner de bons résultats sans que le Liage n'atteigne des scores satisfaisants. Nous introduisons donc une mesure supplémentaire afin d'évaluer l'identification à son niveau global. Cette mesure, décomposée en précision (P_{ALL}), rappel (R_{ALL}) et F-mesure (F1, F_{ALL}), combine l'exactitude du Liage à la précision et au rappel de la REN, telle que :

$$P_{all} = P_{REN} \times Acc$$
$$R_{all} = R_{REN} \times Acc$$

et

$$F_{all} = \frac{2 \times P_{all} \times R_{all}}{(P_{all} + R_{all})}$$

Dans le cas d'une REN parfaite, la mesure globale se ramène à l'exactitude du Liage avec

$$P_{all} = R_{all} = F_{all} = Acc$$

et le cas est alors équivalent au cas TAC-KBP étendu.

1.2 Configuration informée

Partant du cas naïf sur corpus bruts, une configuration informée est définie par la possibilité d'infirmer le statut dénotationnel des segments retournés comme mentions d'entités par l'étape

de REN. La caractère informé de cette approche réside dans la prise en compte des possibles erreurs de REN, se traduisant par une évaluation du statut dénotationnel par le composant de Liage lui-même.

1.2.1 Approche jointe en cascade

La configuration naïve sur corpus brut est ici modifiée pour tenir compte des éventuelles erreurs retournées par la REN. Une configuration *informée en cascade* est proposée : elle demeure déterministe, mais le composant de Liage peut infirmer le statut dénotationnel de chaque mention qui lui est présentée : le processus d'alignement est alors augmenté d'une valeur de retour possible, applicable dans les cas où le segment évalué est jugé non dénotationnel ; le statut de mention est alors éliminé. Cette valeur est notée NAE pour *Not An Entity* et peut être retournée au même titre que l'identifiant d'une entrée de la BC ou NIL.

En pratique, le cas NAE peut être traité de la même façon que NIL : directement intégré à l'ensemble des candidats à l'alignement et sujet au même ordonnancement, ou considéré dans un second temps en concurrence avec une solution de la BC ou NIL. L'introduction du cas NAE implique que les facteurs permettant de déterminer l'absence de dénotation soient intégrés à l'apprentissage du modèle d'alignement. Ces facteurs peuvent relever d'indices généralement pris en compte dans la REN — ambiguïté avec le lexique général, capitalisation de début de phrase plutôt que de nom propre... — ou d'éléments davantage liés au problème de la dénotation et de l'alignement — hypothèses d'alignement sur la BC faiblement supportées, notamment.

On suppose dans l'approche en cascade informée que la précision de la tâche globale (P_{ALL}) peut être améliorée relativement au cas naïf grâce à l'élimination de lectures non dénotationnelles avec NAE. Afin d'être efficace, cette remise en cause des résultats de la REN doit correctement identifier les cas de faux positifs sans éliminer les mentions effectivement dénotationnelles. Autrement dit, il est attendu que la précision (P_{ALL}) augmente sans que le rappel (R_{ALL}) ne se dégrade. De plus, par transitivité, la précision de la REN (P_{REN}) peut elle aussi être améliorée par l'identification des faux positifs ; cet effet de retour permet de caractériser cette approche comme jointe, dans la mesure où un composant du système (Liage) retourne des analyses concernant un composant antérieur (REN). L'efficacité de l'introduction de la solution NAE est évaluée à l'aide de mesures de précision (rapport entre faux positifs réels et requêtes considérées comme non dénotationnelles par le système) et de rappel (taux des faux positifs réels retournés par le système), ainsi que du F-score qui en dérive ; on note ces mesures P_{NAE}, R_{NAE} et F_{NAE}.

1.2.2 Approche modulaire jointe

L'élimination de mentions non dénotationnelles dans la configuration informée en cascade ne peut porter que sur les analyses retournées par la REN, qui est dans ce cas déterministe. Les décisions de la REN sont alors prises en fonction de critères et de connaissances limitées à son niveau d'analyse, essentiellement surfacique, alors que la désambiguïsation des frontières et types d'entités nommées peut dépendre de leur statut dénotationnel. Autrement dit, la lecture la plus probable en termes d'entités est en partie déterminée par des connaissances inaccessibles à un système de REN, mais disponibles lors du processus d'alignement. Celui-ci ne peut cependant pas s'appliquer de façon pertinente en l'absence des analyses correctes au niveau des mentions d'entités.

L'approche *modulaire jointe* considère donc la possibilité d'une REN non-déterministe, conservant un certain nombre d'ambiguïtés quant aux frontières et aux types des mentions repérées. Les analyses retournées par la REN forment alors des *lectures* au sein desquelles des zones d'ambiguïté présentent plusieurs mentions en concurrence (cf. *supra*, section 1). On suppose alors que

certaines décisions de désambiguïsation au niveau de ces zones peuvent être reportées au composant de Liage. Le choix de la lecture en termes de mentions est dépendant de l'alignement de chacune des mentions d'une zone donnée, l'alignement le plus probable dans une zone déterminant la lecture correcte. Comme dans la configuration informée en cascade, les résultats de REN sont affectés par les décisions prises au niveau du Liage, en ce qui concerne les éliminations de segments non dénotationnels mais également les choix de frontières et de types de mentions. L'efficacité de la solution NAE est également évaluée à l'aide des mesures P_{NAE}, R_{NAE} et F_{NAE}.

Les différentes configurations de la tâche d'identification ainsi que les caractéristiques des composants qu'elles impliquent sont récapitulées à la table 6.1. Les types de candidats possibles sont indiqués pour le Liage.

	Configuration naïve		**Configuration informée**	
	Cas TAC-KBP (1)	**Cas corpus brut (2)**	**Cascade (3)**	**Modulaire-jointe (4)**
REN	Mentions du corpus de référence	Automatique, déterministe	Automatique, déterministe	Automatique, non-déterministe
Liage	BC, NIL	BC, NIL	BC, NIL, NAE	BC, NIL, NAE
Éval. Liage	Acc	Acc	Acc	Acc
Éval. REN	$\perp$	P_{REN}, R_{REN}, F_{REN}	P_{REN}, R_{REN}, F_{REN}	P_{REN}, R_{REN}, F_{REN}
Éval. Globale	$\perp$	P_{ALL}, R_{ALL}, F_{ALL}	P_{ALL}, R_{ALL}, F_{ALL}	P_{ALL}, R_{ALL}, F_{ALL}

Table 6.1 : Configurations de la tâche d'identification.

2 Composants de Nomos

2.1 Module de Reconnaissance d'Entités Nommées et construction de lectures

Comme cela a été exposé au chapitre 5 (section 3), tout système de REN peut être intégré à Nomos pour l'obtention des mentions qu'il s'agit de lier à des entités. Les expériences présentées ici sont menées avec deux systèmes de REN, SxPipe/NP et LIAne, qui présentent deux approches et deux fonctionnements différents de la tâche de REN. Leur intégration concerne les configurations présentées ci-dessus, à l'exception du cas TAC-KBP où les mentions d'entités ne sont pas obtenues par voie automatique. Dans cette configuration, la REN employée correspond à l'annotation en entités nommées des données de référence disponibles pour la tâche : il s'agit dans notre contexte du corpus GAFP, dont les spécifications ont été détaillées au chapitre précédent (section 2.1). Dans les autres cas, la REN peut être déterministe (cas 2 et 3) ou non-déterministe (cas 4). La REN déterministe est réalisée à l'aide du système LIAne dans sa version non-ambigüe, où les meilleures analyses sont retournées (cf. chapitre 5, section 3.1.2)[2]. LIAne présente également une version non-déterministe, tout comme SxPipe/NP ; chacun de ces systèmes peut donc être employé dans la configuration 4.

Une REN non-déterministe peut également être obtenue par combinaison de ces deux systèmes : cette combinaison peut prendre la forme d'une union (tous les résultats de LIAne ajoutés à tous les résultats de SxPipe/NP) ou d'une intersection (seulement les résultats communs à LIAne et SxPipe/NP). Ces opérations sont réalisées sur les DAG normalisés de chaque système (cf. chapitre 5, sections 1.2.1 et 3.1), où les zones d'ambiguïté présentent des chemins de même longueur avec un seul état initial et un seul état final. L'intersection se ramène à un filtrage par exclusion de toutes les analyses en termes de mentions, dans ou hors des zones d'ambiguïté, ne figurant

2. Le système SxPipe/NP peut également fournir des analyses désambiguïsées après application du module npNormalizer présenté au chapitre précédent (section 3.2). Il n'est cependant pas considéré ici pour la REN déterministe dans la mesure où la désambiguïsation réalisée par npNormalizer procède déjà d'une approche jointe où la reconnaissance finale des mentions dépend en partie de leur alignement.

que dans l'un des systèmes. L'union des résultats des deux systèmes inclut les analyses retournées par l'un ou l'autre, les analyses retournées par les deux avec suppression des doublons, et de nouvelles zones d'ambiguïté dans les cas d'analyses distinctes recouvrant partiellement seulement les mêmes zones normalisées de part et d'autre.

À partir des données de référence disponibles pour notre tâche, c'est-à-dire le corpus GAFP, différentes analyses en termes de REN sont dérivées. La table 6.2 rapporte le nombre de mentions (notées ENAMEX) ainsi que de zones d'ambiguïtés (ou alternatives, notées ALT), dans le cas d'analyses non-déterministes, pour chacune des configurations de REN envisagées sur ces données. Dans le cas TAC-KBP, dont la REN est absente, ces quantités correspondent à l'annotation de référence du corpus GAFP. Les différentes configurations de REN sont notées :

Gold mentions de l'annotation de référence,

L1 résultats de LIANE dans sa version déterministe ou *1-best*,

LN résultats de LIANE dans sa version non-déterministe ou *n-best*,

SA résultats de SxPipe/NP dont les analyses sont ambigües,

LNSAU union des résultats de LN et SA,

LNSAI intersection des résultats de LN et SA.

	Gold	**L1**	**LN**	**SA**	**LNSAU**	**LNSAI**
# ENAMEX	1 535	1 669	1 784	1 803	2 364	1 210
# ALT	0	0	88	163	346	3

TABLE 6.2 : Nombre de mentions (ENAMEX) et de zones d'ambiguïté (ALT) dans les différentes configurations de REN sur le corpus GAFP.

Chaque configuration de REN donne lieu à différentes lectures du texte d'entrée. Les figures 6.1 à 6.4 présentent les lectures obtenues à l'issue de la REN à travers quelques exemples du corpus GAFP. On observe, d'après la table 6.2, que la configuration LNSAI (intersection des résultats non-déterministes de LIANE et SxPipe/NP) présente très peu de zones d'ambiguïté (3 sur l'ensemble du corpus GAFP) et moins de mentions que le total indiqué par les données de références. L'intersection tend manifestement à éliminer de possibles ambiguïtés de lecture, et dans certains cas les lectures non dénotationnelles ou incorrectes, mais également des lectures correctes. Cette configuration de REN peut être vue, dans le meilleur cas, comme un procédé d'élimination de mentions dès avant l'étape de Liage, qui prend en charge la détection de faux positifs dans les cas 3 et 4. L'efficacité de cette amélioration de la précision doit néanmoins être évaluée au regard du taux de lectures abandonnées à tort, autrement dit du taux de rappel finalement obtenu à l'aide de cette REN.

2.2 Liage d'Entités et apprentissage supervisé

2.2.1 Liage

Chaque configuration de REN présentée ci-dessus donne lieu à un ensemble de mentions, qui sont toutes concernées par le Liage. Celui-ci se présente cependant différemment selon les cas : dans le cas 1, toutes les mentions sont dénotationnelles et l'alignement de chacune avec une entité, issue de la BC ou NIL, est dans tous les cas un problème pertinent. Dans les autres cas (2 à 4), certaines mentions retournées ne sont en fait pas dénotationnelles et la question de leur

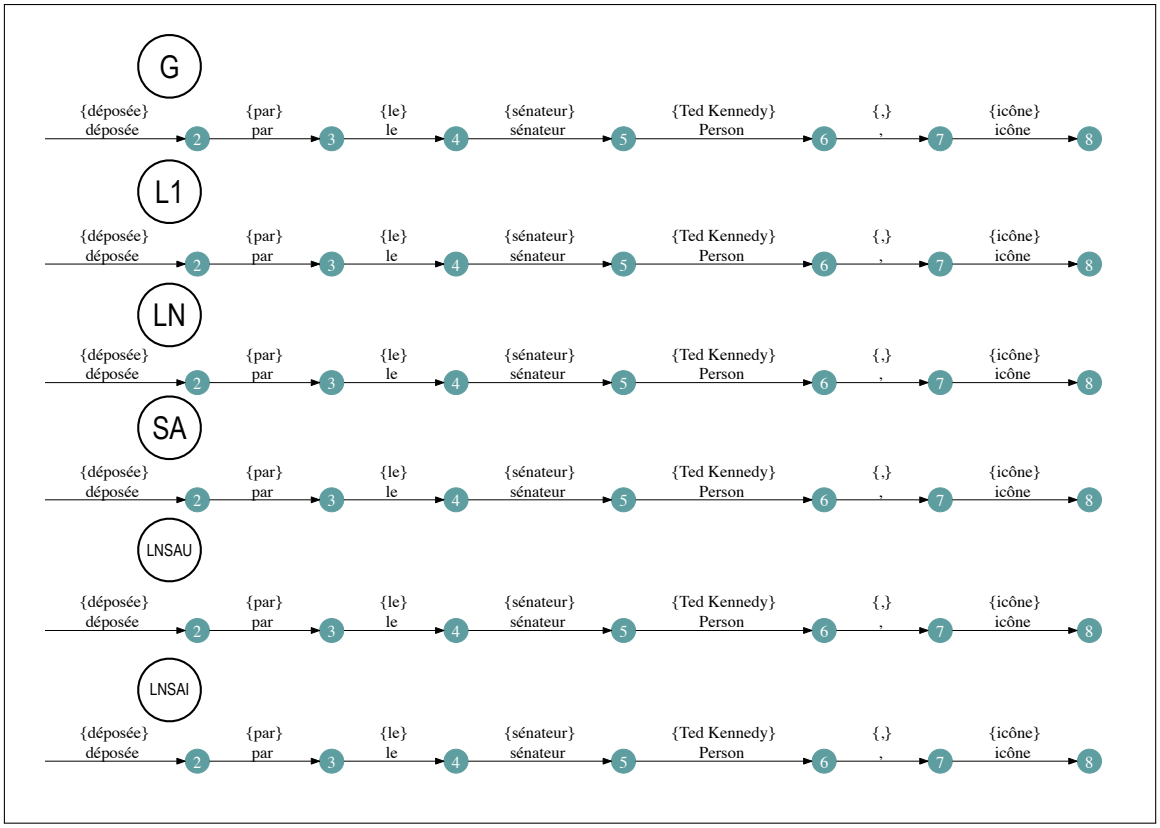


Figure 6.1 : Exemple de lectures dérivées des différentes configurations de REN.

alignement est alors sans objet. La méthode de Liage proposée ici est donc modifié relativement à la configuration de TAC-KBP. Il s'agit principalement de permettre au composant de Liage, dans les cas 3 et 4, de retourner une solution représentant l'absence de dénotation, c'est-à-dire NAE. Le cas 2 est pris en compte ici afin de vérifier la pertinence de cette solution.

La formalisation du Liage proposée dans la description de la tâche dans le cadre de TAC-KBP peut être modifiée ici de la façon suivante :
Soient

- M l'ensemble des mentions dénotationnelles, $M_{ref} \subset M$ l'ensemble des mentions, toutes dénotationnelles, du corpus de référence, et Z l'ensemble des segments non dénotationnels tel que $Z \wedge M = \emptyset$;
- une requête r pour l'alignement, qui peut être :
  - une mention $r \in M_{ref}$, c'est-à-dire une mention à lier dans un document d issu du corpus D de référence ; r est nécessairement dénotationnelle ($r \in M_{ref}$) puisqu'issue des données de référence ;
  - un segment non dénotationnel $r \in Z$ retourné par le module de REN (cas 2 à 4) ;
  - un segment r pouvant être interprété comme dénotationnel, retourné par le module de REN (cas 2 à 4), mais constituant une correspondance partielle avec une mention de M_{ref}, c'est-à-dire $r \in M$ mais $r \notin M_{ref}$; on note r_{ref} la mention de M_{ref} dont r est une correspondance partielle (sur la figure 6.2 par exemple, on a r le segment *Washington* dans l'analyse SA/LNSAU, $r \in M$ et $r \notin M_{ref}$, et r_{ref} est dans ce cas le segment *Washington Post*) ;

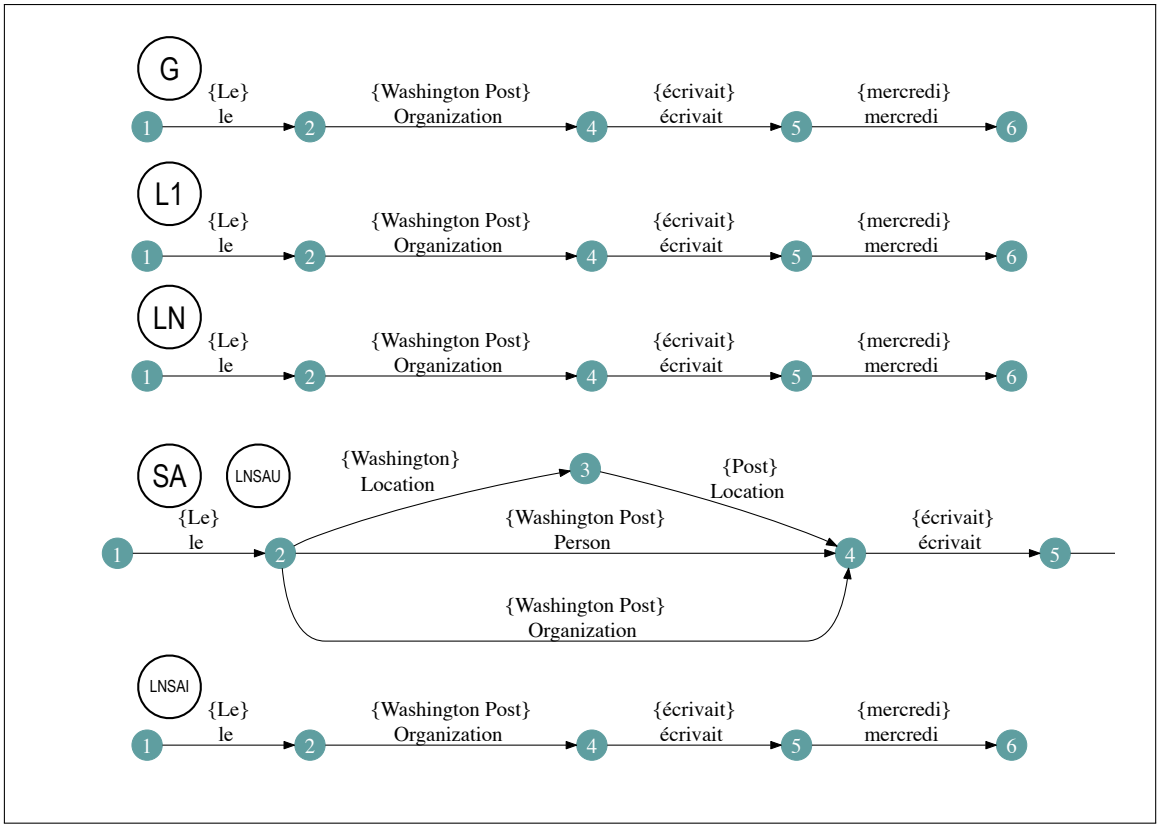


FIGURE 6.2 : Exemple de lectures dérivées des différentes configurations de REN.

- e_r l'entité associée à r dans les données de référence si $r \in M_{ref}$; $e_r = \perp$ si $r \in Z$; $e_r = \perp$ également si $r \notin Z$ et $r \notin M_{ref}$, bien que dans certains cas on puisse interpréter r avec le même alignement que r_{ref} ;
- e'_r la solution d'alignement retournée par le système ;
- $E = \{e_1, ..., e_n\}$ l'ensemble des entités constituant des entrées de la BC ; on a donc $n = |E|$;
- e_{out} l'entité spéciale NIL ;
- e_{nae} le cas spécial NAE introduit dans les cas 3 et 4 ;

on a, dans les cas 1 et 2 :

- $E_{ext} = E \cup \{e_{out}\}$
- $e_r \in E_{ext}$ si $r \in M_{ref}$; $e_r = \perp$ si $r \in Z$ dans le cas 2 ; $e_r = \perp$ si $r \in M$ et $r \notin M_{ref}$ dans le cas 2 ;
- $e'_r \in E_{ext}$
- dans le cas 2, si $r \notin M_{ref}$, e'_r est nécessairement différent de e_r qui est égal à $\perp$

et, dans les cas 3 et 4 :

- $E_{ext} = E \cup \{e_{out}, e_{nae}\}$

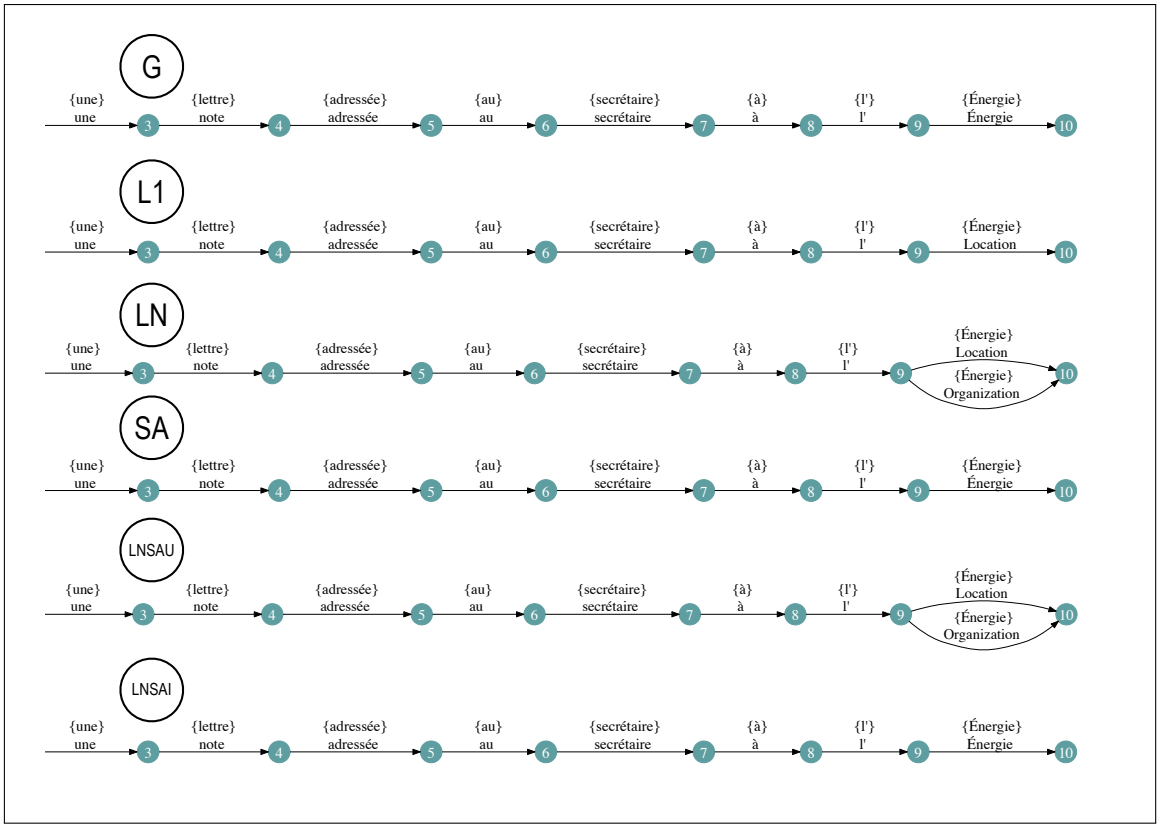


Figure 6.3 : Exemple de lectures dérivées des différentes configurations de REN.

- $e_r \in E_{ext}$; $e_r = e_{nae}$ si $r \notin M_{ref}$, c'est-à-dire si $r \in Z$ ou si $r \in M$ et $r \notin M_{ref}$
- $e'_r \in E_{ext}$

On définit la fonction d'alignement f :

$$f : M \cup Z \mapsto E_{ext}$$

comme

$$f(r) = \underset{e \in E_{ext}}{\operatorname{argmax}}\ g(r, e) = e'_r$$

où g, pour les e_i de E, est une fonction de quantification de la proximité entre une requête r et e_i, représentées par leurs sens respectifs, eux-mêmes dérivés d'une représentation des contextes correspondants. Pour e_{out} et e_{nae}, g tient compte d'autres critères définis pour ces cas. La fonction de quantification g peut par ailleurs être vue comme une fonction de score dont la fonction f utilise le résultat afin de déterminer e'_r.

On a ainsi, pour une requête r :

- pour tout $e \in E_{ext}$, un score s_r de e tel que $s_r(e) = g(r, e)$
- d'où $e'_r = f(r) = \operatorname{argmax}_{e \in E_{ext}}\ s_r(e)$

Le système Nomos intègre la méthode de Liage avec les spécifications suivantes :

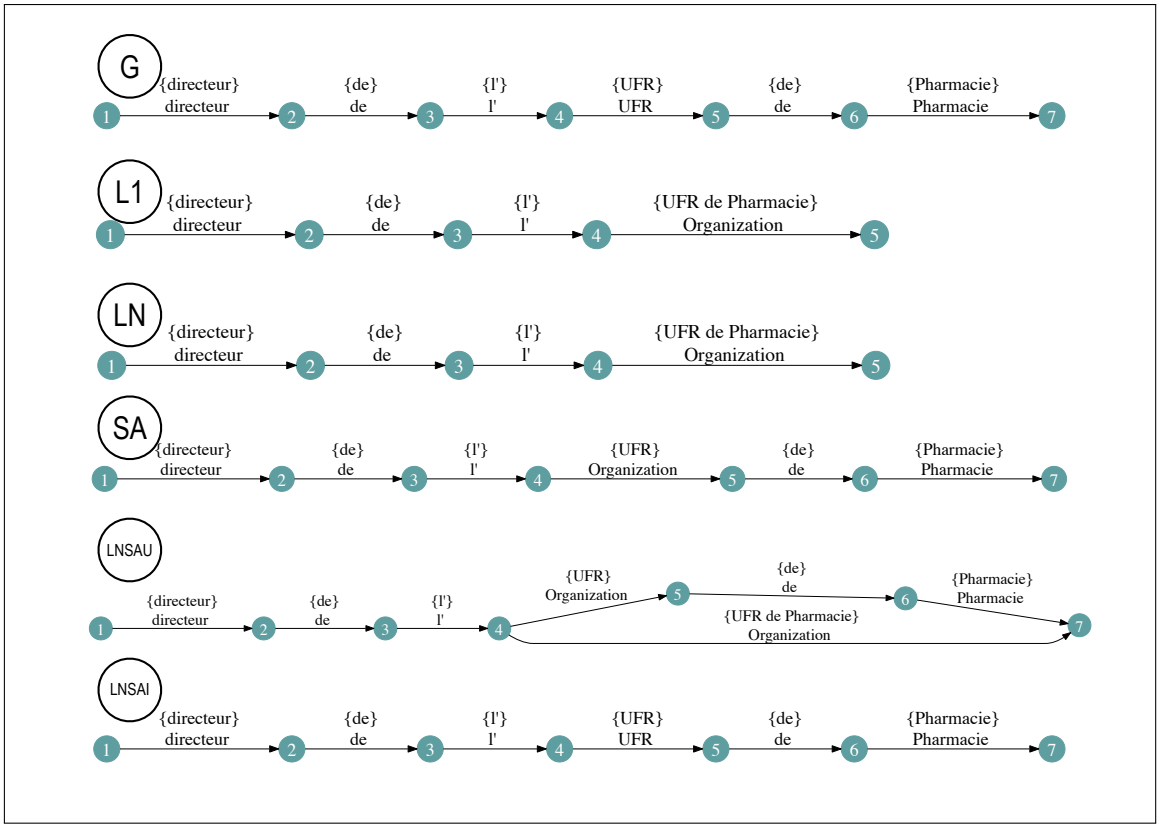


FIGURE 6.4 : Exemple de lectures dérivées des différentes configurations de REN.

Base de connaissances Les entités cibles du Liage (ensemble E) sont disponibles dans la base Aleda et les connaissances rassemblées sur ces entités sont mises à disposition dans la base Nomos-KB (chapitre 5, section 2.2).

Sélection des candidats Pour chaque mention à aligner, l'ensemble des candidats C est constitué à partir d'une requête sur la base Aleda : toutes les entités d'Aleda pour lesquelles la mention figure parmi les variantes lexicales sont sélectionnées. Cette étape repose sur la qualité de l'index des variantes d'entités dans Aleda et détermine un taux de rappel d'entités crucial pour la suite de la tâche. Si l'entité effectivement dénotée est présente dans Aleda mais non atteinte par ce procédé, l'alignement correct ne peut avoir lieu. La constitution de l'index des variantes est similaire à la plupart des travaux réalisés en Liage (cf. chapitre 3, section 3.3) et repose principalement sur les titres d'articles, redirections et désambiguïsation de Wikipedia, ainsi que sur les labels de lieux dans GeoNames. Le taux de rappel de candidats ainsi obtenu est de 88,01% sur le corpus GAFP. On peut d'ores et déjà observer que, pour les mentions dans M_{ref} à aligner sur une entité dans E, la correction obtenue à l'issue de l'analyse (Acc_{BC}) sur les mentions des données de référence ne peut pas dépasser ce taux : aucun alignement n'est en effet possible en dehors de l'ensemble de candidats sélectionnés pour une mention donnée. La mesure Acc_{BC} peut en revanche dépasser ce taux avec des configurations de REN automatique, où l'ensemble des mentions sur lesquelles elle s'applique, avec un rappel inférieur à 100%, n'est pas le même que les mentions de référence.

Le candidat NIL est ajouté à l'ensemble des candidats et ainsi considéré de façon concomitante dans la procédure d'ordonnancement. Dans les configurations informées (cas 3 et 4),

le candidat NAE est également ajouté.

Ordonnancement des candidats Les candidats de chaque mention sont ordonnés en fonction de critères touchant leur relation à la mention, notamment similarité pour les candidats issus de la BC. Ces critères forment pour chaque paire (m, c) un ensemble de traits intégrés à l'apprentissage du modèle de Liage et permettant de calculer son score.

2.2.2 Apprentissage supervisé

Seule l'étape d'ordonnancement des candidats est concernée par l'apprentissage supervisé d'un modèle. Dans les cas 1 et 2, ce modèle est similaire à ceux présentés dans le cadre de TAC-KBP et permet de choisir un candidat pour l'alignement d'une mention donnée, toutes les requêtes étant considérées comme mentions dénotationnelles (à juste titre dans le cas 1, mais parfois erronée dans le cas 2). Dans les cas 3 et 4, l'apprentissage du modèle intègre la notion de faux positif ; le Liage joue alors à la fois un rôle d'identification par alignement pour les requêtes dénotationnelles et de repérage des faux positifs pour les autres. Les cas de correspondance partielle (cf. exemples 24 et 25, page 156) donnent lieu à une représentation particulière dans l'apprentissage du modèle, comme cela sera exposé ci-après. On verra que dans les cas 2 à 4, le Liage est ainsi concerné à la fois par la validité de l'alignement des mentions et la correction du statut dénotationnel des mentions retournées par le module de REN.

Cas 1 Dans ce cas, toute requête r correspond nécessairement à une dénotation d'entité : $\forall r, r \in M$ et $\forall r, r \in M_{ref}$. Le modèle intègre les différents critères présentés ci-dessus pour la candidats BC et NIL sous forme de traits associés à chaque paire réunissant une mention m et un candidat c. L'apprentissage est envisagé selon l'approche introduite lors de l'étude du Liage dans le cadre de TAC-KBP (cf. chapitre 3, section 3.3.2), comme suit :

- Soient une requête r et un ensemble de candidats $C = C_a \cup C_b$ avec $C_a = \{c_1, ..., c_n\}$, $C_a \subset E$, où n est le nombre de candidats générés par la requête parmi les entrées de la BC et $C_b = \{e_{out}\}$

- Pour une paire (r, c_i), on définit un vecteur

$$v_{ri} = \phi(r, c_i) \in \mathbb{R}^d$$

où d est le nombre de traits considérés, avec

$$\phi(r, c_i) = \Big(\phi_1(r, c_i), \phi_2(r, c_i), ..., \phi_d(r, c_i)\Big).$$

- Pour une mention m, on calcule ainsi la séquence $V = [v_1, ..., v_n]$ de vecteurs de traits.

- On cherche à apprendre une fonction h telle que

$$h : M_{ref} \times E_{ext} \mapsto \{0, 1\} \quad \text{et} \quad h(r, c) = \begin{cases} 1 & \text{si } c = e_r \\ 0 & \text{sinon} \end{cases}$$

à l'aide de ϕ telle que

$$h(r, c) = h_\phi\Big(\phi(r, c)\Big)$$

- On génère les exemples d'entraînement pour h_ϕ à partir des éléments de V associés à une classe dans $\{0, 1\}$, indiquée par les données de référence de la tâche :

$$h_\phi(v_{ri}) = \begin{cases} 1 & \text{si } c = e_r \\ 0 & \text{sinon} \end{cases}$$

- On a donc pour chaque mention un nombre d'exemples égal à $n + 1$ (candidats issus de la BC auxquels s'ajoute e_{out}). Parmi ces exemples, seule la paire (r, c) où $c = e_r$ est étiquetée avec la classe positive, les autres recevant la classe négative.

Les exemples d'apprentissage (ensemble $Train$) et les instances soumises à la prédiction (ensemble $Pred$) se présentent de la façon suivante :
Pour une requête r_1 dans $Train$ dont l'alignement dans les données de référence est l'entité de la BC e_{34}, et dont l'ensemble de candidats est $C = \{e_{23}, e_{34}, e_{45}, e_{out}\}$, chaque paire (r_1, c_i) avec $c_i \in C$ est représentée par un vecteur $\phi(r_1, c)$ et associée à une classe :

$$\begin{array}{ll} \phi(r_1, e_{23}) & 0 \\ \phi(r_1, e_{34}) & 1 \\ \phi(r_1, e_{45}) & 0 \\ \phi(r_1, e_{out}) & 0 \end{array}$$

On adopte un mode d'apprentissage de l'ordonnancement point-à-point (*pointwise*), à l'aide d'un classifieur à maximum d'entropie [3]. Chaque exemple d'entraînement correspond à une paire (r, c) et est considéré indépendamment des autres par le classifieur. Lors de la prédiction, l'ordonnancement point-à-point tient compte du score assigné par le modèle à la paire (r, c), et non de sa classe. Pour une même mention r, les candidats c_i sont regroupés et ordonnés en fonction du score de la paire où chacun d'eux figure. Pour une requête r dans $Pred$ à aligner sur e_{34}, les paires (r, c_i) correspondantes formant les instances soumises à la prédiction reçoivent donc un score $s_r(\phi(r, c_i))$ et on cherche à obtenir

$$\begin{array}{rlll} & s_{r1}(\phi(r_1, e_{34})) & > & s_{r1}(\phi(r_1, e_{23})) \\ \text{et} & s_{r1}(\phi(r_1, e_{34})) & > & s_{r1}(\phi(r_1, e_{45})) \\ \text{et} & s_{r1}(\phi(r_1, e_{34})) & > & s_{r1}(\phi(r_1, e_{out})) \end{array}$$

Cas 2 Le cas 2 reprend la formalisation du Liage du cas 1. Les requêtes r ne sont en revanche pas nécessairement toutes dénotationnelles ni présentes dans les données de référence ($r \in M$ et $r \in M_{ref}$, ou $r \in M$ et $r \notin M_{ref}$ ou $r \in Z$). Trois approches sont alors possibles :

- Un modèle appris sur les données de référence, c'est-à-dire résultant du cas 1, peut être appliqué aux résultats de REN déterministe de L1.

- Un modèle distinct peut également être appris sur les sorties de L1, combinées aux données de référence. Dans ce second cas, pour les exemples où $r \notin M_{ref}$, aucune des paires (r, c) ne reçoit la classe positive. À la prédiction, la méthode point-à-point assigne cependant un score à chaque candidat $c \in E_{ext} = E \cup \{e_{out}\}$, dont le premier en vertu de ce score est retourné en tant que e'_r, comme avec le modèle issu du cas 1.

- Dans ces deux approches, les cas où $r \in Z$ ou $r \in M$ mais $r \notin M_{ref}$ donnent nécessairement lieu à un alignement incorrect puisque sans objet : les réponses possibles du système

3. Nous utilisons le logiciel *MEGA Model Optimization Package* de Daumé III [DI04], disponible à l'adresse `http://www.umiacs.umd.edu/~hal/megam`

(e'_r) sont en effet limitées à l'ensemble $E_{ext} = E \cup \{e_{out}\}$, tandis que e_r est égal à $\perp$. On peut envisager la prise en charge de $\perp$ par l'introduction d'un seuil appliqué aux scores $s_r(\phi(r, c_i))$, en-deçà duquel le système, alors qualifiable d'abstentionniste, ne retourne aucun candidat. Ce seuil peut reposer sur la limite entre classe positive et classe négative dans le classifieur, autrement dit 0,5 : on considère alors que toute requête pour laquelle aucun candidat n'obtient un score permettant de lui assigner la classe positive ne reçoit aucun alignement. Cette approche réintègre l'utilisation du classifieur en tant que moyen de discrimination binaire, en plus de son usage détourné pour la méthode d'ordonnancement point à point.

Les correspondances partielles sont ici traitées comme des faux positifs : une mention retournée par le module de REN, inexacte du point de vue des frontières relativement à la mention indiquée par les données de référence, reçoit lors de l'apprentissage du modèle la classe négative quel que soit le candidat avec lequel elle est associée. Cette procédure prend en compte le caractère erroné de la lecture obtenue avec une mention partielle, même si $r \in M$ et si son alignement correspond à l'entité indiquée dans les données de référence pour la mention complète (r_{ref}). Le label de classe, dont on peut rappeler qu'il ne constitue pas l'objectif final dans la méthode d'ordonnancement, exprime en effet dans Nomos la correction à la fois au niveau de l'alignement et de la REN. Ainsi, dans tous les cas où les mentions sont obtenues par un système de REN (cas 2 à 4), seuls les exemples d'entraînement où $r \in M_{ref}$ et où $e'_r = e_r$ reçoivent un label positif. On cherche dans le cas 2 à obtenir un score inférieur à 0,5 pour tout candidat d'une requête $rinM$ mais $r \notin M_{ref}$ ou $r \in Z$ (correspondance partielle ou faux positif). Avec l'introduction d'un seuil minimal, de tels cas ne devraient donc donner lieu à aucun alignement.

Cas 3 La prise en compte explicite des faux positifs donne lieu dans le cas 3 à un apprentissage du modèle de Liage sur les sorties du module de REN (ici L1, déterministe) combinées aux données de référence. On obtient ainsi des exemples similaires au cas 1 pour les mentions d'entités, autrement dit lorsque $r \in M_{ref}$, ainsi que des exemples de faux positifs et de correspondances partielles. L'introduction du candidat NAE permet d'assigner cette solution aux cas de faux positifs, autrement dit lorsque $r \in Z$. Les correspondances partielles sont traitées de façon similaire au cas 2 : aucune paire (r, c) où $r \in M$ mais $r \notin M_{ref}$ ne reçoit de classe positive, quel que soit le statut de c (BC, NIL ou NAE).

On a donc dans le cas 2 des requêtes $r \in M$ ou $r \in Z$ et pour chaque r un ensemble de candidats C avec $C_a = \{c_1, ..., c_n\}$, $C_a \subseteq E$, où n est le nombre de candidats générés par la requête parmi les entrées de la BC et $C_b = \{e_{out}, e_{nae}\}$.
Pour une requête $r \in Train$ et $r \in M_{ref}$ avec l'alignement e_{91}, et dont l'ensemble de candidats est $C = \{e_{23}, e_{38}, e_{91}, e_{out}, e_{nae}\}$, les exemples d'entraînement sont les suivants :

$$\begin{array}{ll} \phi(r_1, e_{23}) & 0 \\ \phi(r_1, e_{38}) & 0 \\ \phi(r_1, e_{91}) & 1 \\ \phi(r_1, e_{out}) & 0 \\ \phi(r_1, e_{nae}) & 0 \end{array}$$

Si $r \in Pred$ et r est à aligner sur e_{91}, on cherche à obtenir par l'ordonnancement point-à-point :

$$\begin{array}{lrcl} & s_{r1}(\phi(r_1, e_{91})) & > & s_{r1}(\phi(r_1, e_{23})) \\ \text{et} & s_{r1}(\phi(r_1, e_{91})) & > & s_{r1}(\phi(r_1, e_{38})) \\ \text{et} & s_{r1}(\phi(r_1, e_{91})) & > & s_{r1}(\phi(r_1, e_{out})) \\ \text{et} & s_{r1}(\phi(r_1, e_{91})) & > & s_{r1}(\phi(r_1, e_{nae})) \end{array}$$

Pour une requête $r \in Train$ et $r \in Z$, dont l'ensemble de candidats est $C = \{e_{12}, e_{out}, e_{nae}\}$, les exemples d'entraînement sont les suivants :

$$\begin{array}{ll} \phi(r_1, e_{12}) & 0 \\ \phi(r_1, e_{out}) & 0 \\ \phi(r_1, e_{nae}) & 1 \end{array}$$

Si $r \in Pred$ et r est à aligner sur e_{nae}, on cherche à obtenir par l'ordonnancement point-à-point :

$$\begin{array}{rlll} & s_{r1}(\phi(r_1, e_{nae})) & > & s_{r1}(\phi(r_1, e_{12})) \\ \text{et} & s_{r1}(\phi(r_1, e_{nae})) & > & s_{r1}(\phi(r_1, e_{out})) \end{array}$$

Une requête $r \in Train$ et $r \in M$ mais $r \notin M_{ref}$, dont l'ensemble de candidats est $C = \{e_{52}, e_{48}, e_{out}, e_{nae}\}$, pose le problème suivant : l'alignement de r n'est pas à proprement parler la cas NAE. Il ne s'agit en revanche d'aucun autre des cas possibles. On n'attribue donc la classe positive à aucun exemple et les exemples d'entraînement sont les suivants :

$$\begin{array}{ll} \phi(r_1, e_{52}) & 0 \\ \phi(r_1, e_{48}) & 0 \\ \phi(r_1, e_{out}) & 0 \\ \phi(r_1, e_{nae}) & 0 \end{array}$$

Si $r \in Pred$, peut alors envisager d'appliquer un critère d'élimination tel qu'un seuil de score en-deçà duquel la mention ne donne pas d'alignement. Alternativement, on cherche à obtenir un score supérieur à tous les autres avec NAE, qui ne traduit pas fidèlement la notion de correspondance partielle mais peut être considérée comme la plus satisfaisante. On cherche alors à obtenir :

$$\begin{array}{rlll} & s_{r1}(\phi(r_1, e_{nae})) & > & s_{r1}(\phi(r_1, e_{52})) \\ \text{et} & s_{r1}(\phi(r_1, e_{nae})) & > & s_{r1}(\phi(r_1, e_{48})) \\ \text{et} & s_{r1}(\phi(r_1, e_{nae})) & > & s_{r1}(\phi(r_1, e_{out})) \end{array}$$

Il s'agira pour ce cas de vérifier si la configuration du cas 4, avec une REN non-déterministe, permet l'elimination de lectures partielles par la sélection d'une lecture alternative et correcte.

Cas 4 La configuration du Liage dans le cas 4 est identique à celle du cas 3, et les exemples d'entraînement sont produits de la même façon. Au sein des zones d'ambiguïté, chaque mention donne lieu à un ensemble d'exemples correspondant à son ensemble de candidats, et seule la mention correspondant à la lecture correcte donne lieu à un exemple étiqueté avec la classe positive. Les lectures incorrectes sont traitées comme dans les configurations déterministes. Ainsi,

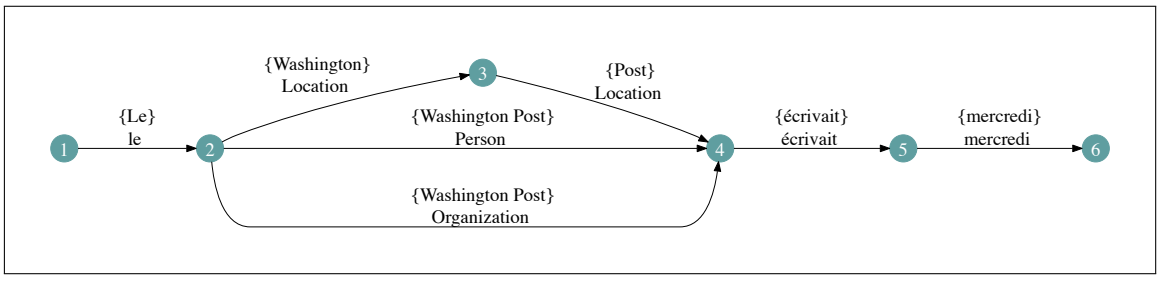


FIGURE 6.5 : Lecture dérivée de la configuration de REN SA.

l'analyse retournée par le module de REN SA reproduite à la figure 6.5 présente l'ensemble de

mentions {r_1 = *Washington* (LOCATION), r_2 = *Post* (LOCATION), r_3 = *Washington Post* (PERSON) et r_4 = *Washington Post* (ORGANIZATION)}. On a pour cette zone la mention *Washington Post* $\in M_{ref}$, avec un alignement sur l'entité e_{70} de type ORGANIZATION. Deux mentions sont retournées par SA avec des frontières correctes, dont une avec le même type que e_{70} [4]. Cette dernière (r_4) correspond donc à la mention de référence et est concernée par un Liage autre que NAE. Chaque mention r_i présente un ensemble de candidats tels que :

$$\begin{aligned} C_1 &= \{e_{93}, e_{36}, e_{out}, e_{nae}\} \\ C_2 &= \{e_{26}, e_{83}, e_{out}, e_{nae}\} \\ C_3 &= \{e_{43}, e_{out}, e_{nae}\} \\ C_4 &= \{e_{70}, e_{out}, e_{nae}\} \end{aligned}$$

On a ainsi pour cette analyse ambigüe, si $r_i \in Train$, les exemples d'entraînement suivants :

$$\begin{array}{llll} \phi(r_1, e_{93}) & 0 & \phi(r_2, e_{26}) & 0 \\ \phi(r_1, e_{36}) & 0 & \phi(r_2, e_{83}) & 0 \\ \phi(r_1, e_{out}) & 0 & \phi(r_2, e_{out}) & 0 \\ \phi(r_1, e_{nae}) & 0 & \phi(r_2, e_{nae}) & 0 \\ \phi(r_3, e_{43}) & 0 & \phi(r_4, e_{70}) & 1 \\ \phi(r_3, e_{out}) & 0 & \phi(r_4, e_{out}) & 0 \\ \phi(r_3, e_{nae}) & 0 & \phi(r_4, e_{nae}) & 0 \end{array}$$

Si $r_i \in Pred$, on cherche à obtenir, par l'ordonnancement point-à-point :

$$\begin{aligned} & s_{r4}(\phi(r_4, e_{70})) > s_{r4}(\phi(r_4, e_{out})) \\ \text{et } & s_{r4}(\phi(r_4, e_{70})) > s_{r4}(\phi(r_4, e_{nae})) \\ \text{et } & s_{r4}(\phi(r_4, e_{70})) > s_{r1}(\phi(r_1, e_{93})) \end{aligned}$$

et de même pour tous les autres $s_{ri}(\phi(r_i, c))$: $s_{r4}(\phi(r_4, e_{70})) > s_{ri}(\phi(r_i, c))$ avec $i \neq 4$ et $c \neq e_{70}$.

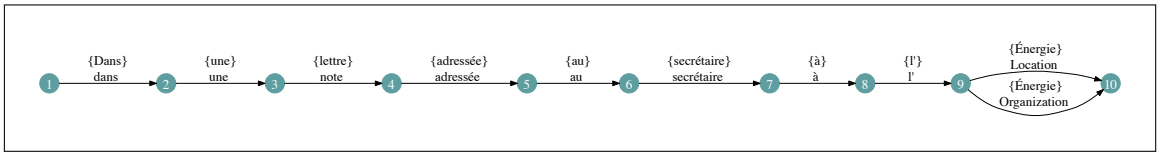


FIGURE 6.6 : Lecture dérivée de la configuration de REN LNSAU.

L'analyse retournée par le module de REN LNSAU reproduite à la figure 6.6 présente l'ensemble de mentions {r_1 = *Énergie* (LOCATION), r_2 = *Énergie* (ORGANIZATION)}. La zone correspondante ne présente aucune mention dans les données de référence. L'ensemble des candidats C_1 de r_1 est $\{e_{50}, e_{out}, e_{nae}\}$ et l'ensemble C_2 de r_2 est $\{e_{out}, e_{nae}\}$. Si $r_i \in Train$, les exemples

4. En pratique, les mentions de frontières identiques dont les types diffèrent sont traitées comme une mention unique, munies de deux types possibles. Lors de la sélection des candidats et de l'alignement, on ne contraint pas l'identité de type entre mention et candidat, lorsque le candidat est issu de la BC (les autres candidats, NIL et NAE, n'ayant pas de type). Le type finalement assigné à une mention est donc celui du candidat choisi lors de l'alignement, si celui-ci est issu de la BC. Il demeure identique aux types retournés par la REN si le candidat NIL est choisi.

d'entraînement sont les suivants :

$$\begin{array}{ll} \phi(r_1, e_{50}) & 0 \\ \phi(r_1, e_{out}) & 0 \\ \phi(r_1, e_{nae}) & 1 \\ \phi(r_2, e_{out}) & 0 \\ \phi(r_2, e_{nae}) & 1 \end{array}$$

Si $r_i \in Pred$, on cherche à obtenir, par l'ordonnancement point-à-point :

$$\begin{array}{rlcl} & s_{r1}(\phi(r_1, e_{nae})) & > & s_{r1}(\phi(r_1, e_{50})) \\ \text{et} & s_{r1}(\phi(r_1, e_{nae})) & > & s_{r1}(\phi(r_1, e_{out})) \\ \text{et} & s_{r4}(\phi(r_2, e_{nae})) & > & s_{r2}(\phi(r_2, e_{out})) \end{array}$$

avec

$$\begin{array}{rlcl} & s_{r1}(\phi(r_1, e_{nae})) & > & s_{r2}(\phi(r_2, e_{nae})) \\ \text{ou} & s_{r2}(\phi(r_2, e_{nae})) & > & s_{r1}(\phi(r_1, e_{nae})) \end{array}$$

2.2.3 Traits pour l'apprentissage

Les critères adoptés pour l'apprentissage du Liage s'appuient en grande partie sur la modélisation des informations collectées au sujet des entités dans Nomos-KB ainsi que sur celle des documents traités ; la notation de ces différentes informations suit l'introduction qui a été faite de cette modélisation au chapitre précédent (section 2). Ces critères forment des ensembles de traits différents pour les candidats issus de la BC, NIL et NAE :

Candidat BC Les critères considérés s'apparentent à la typologie présentée par Ji et al. [JGD11] à partir des principaux systèmes de Liage participant à TAC-KBP. Ils relèvent principalement de

- la notoriété de l'entité, modélisée notamment sous la forme d'un poids dérivé du nombre d'habitants pour les lieux et de la taille de l'article Wikipedia pour les personnes et organisations ; cette notoriété est l'approximation d'un sens *par défaut* ou d'une *probabilité a priori* de la mention ;
- la similarité surfacique entre le nom canonique de l'entité et la mention ;
- la similarité contextuelle entre l'entité, dont les éléments contextuels sont modélisés dans Nomos-KB, et la mention, dont le contexte est le document courant. Plusieurs éléments sont considérés à ce titre :
  - le contexte lexical, pondéré par une saillance dérivée d'un test t ;
  - le contexte thématique, indiqué par les slugs et les catégories IPTC ;
  - les entités co-occurrentes : plusieurs approches du Liage ([HSZ11 ; Rat+11 ; Hof+11]) effectuent de façon globale l'alignement de toutes les mentions d'un même document en intégrant une contrainte de cohérence au niveau des entités obtenues ; il est en effet vraisemblable que des mentions co-occurrentes dénotent des entités partageant de façon générale les mêmes contextes. Nomos accomplit l'alignement des mentions localement et n'intègre donc pas ce type de contraintes. Nomos-KB fournit en revanche des indications sur les co-occurrences d'entités dans le corpus Wikipedia (ensembles E_{ewl1}, E_{ewl2}, E_{rel}, cf. chapitre 5, section 2.2.2). À partir de ces ensembles, on peut dériver pour chaque entité un ensemble de *variantes* co-occurrentes, qui peuvent être recherchées dans le contexte de la mention ; ces variantes étant elles-mêmes des mentions à aligner, on ne connaît pas au préalable les entités qu'elles dénotent mais leur présence dans le contexte usuel des entités candidates peut être vue comme un facteur discriminant.

Candidat NIL Si des candidats de la BC sont définis pour la mention, leur degré de similarité générale constitue une information relative au cas NIL : cette similarité peut être élevée, faible ou nulle, ce qui pourra influer sur l'hypothèse d'un Liage hors BC.

Candidat NAE Une lecture non dénotationnelle de la mention peut être renforcée par l'absence de candidats issus de la BC, une similarité faible des candidats, l'ambiguïté de la mention avec le lexique général, sa position en tête de phrase, ou un score de confiance faible associé à la mention par LIAne, lorsque ce système participe à l'étape de REN.

On dérive de ces différents critères les traits d'apprentissage énumérés à la table 6.3. La spécification complète de chacun de ces traits est donnée à l'annexe B.

2.3 Lectures et ordonnancement

Dans les configurations déterministes (cas 1 à 3), le Liage procède à l'alignement de chacune des mentions (sélection de l'entité placée au rang 1 pour chaque mention) et aboutit directement à la lecture finale du texte donné en entrée. Au niveau des mentions alignées sur des entités issues de la BC ou NIL, le texte est augmenté des informations relatives à ces entités insérées sous forme d'annotations, comme l'illustre la figure 6.7 (cadre 1). Dans le cas 2, les mentions correspondant à de faux positifs peuvent donner lieu à l'introduction d'informations erronées (figure 6.7, cadre 2).

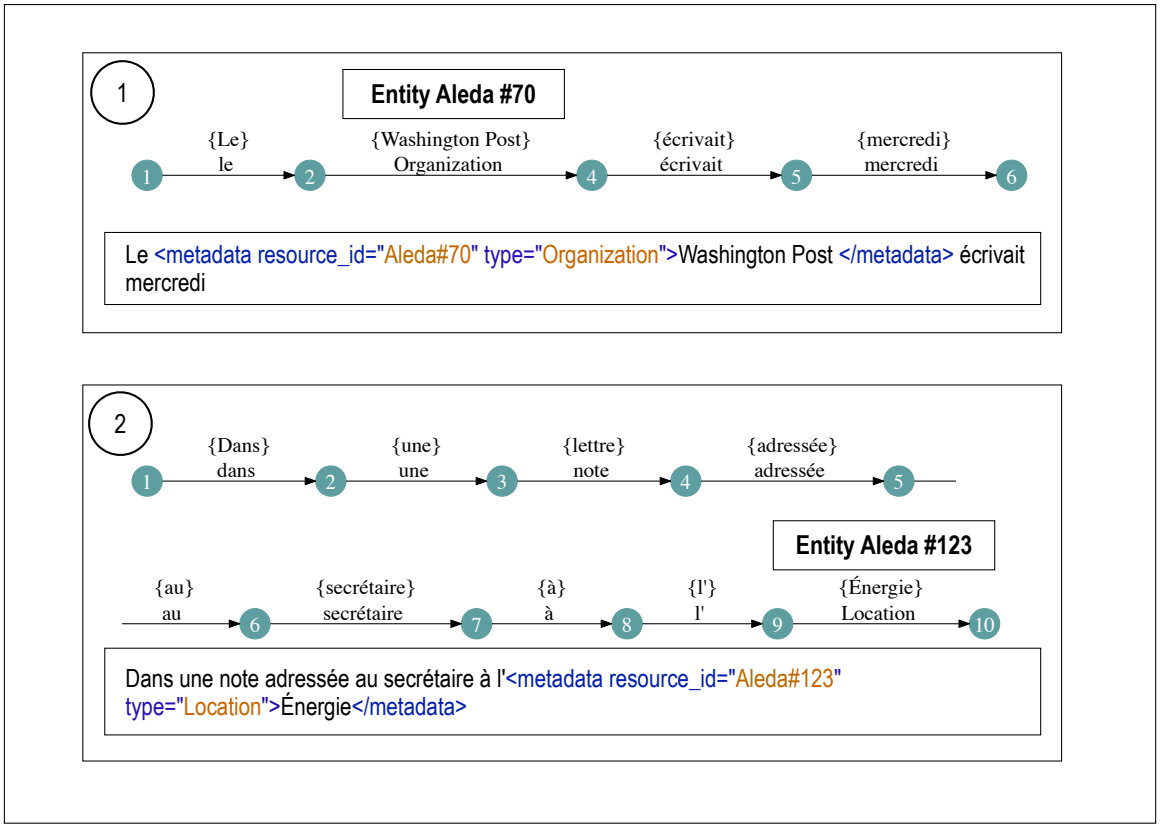


Figure 6.7 : Exemple de lectures obtenues dans les cas 1 et 2.

Dans le cas 3, les mentions pour lesquelles le candidat NAE est placé au rang 1 de l'alignement voient leur statut dénotationnel éliminé. La lecture obtenue indique ainsi qu'aucune entité n'est dénotée à son endroit, comme l'illustre la figure 6.8.

	Trait	Candidats	Variables[1]	Valeurs
[1]	c_wpw	BC [PER, ORG]	c	$\mathbb{R}_{\geq 0}$
[2]	c_geonw	BC [LOC]	c	$\mathbb{R}_{\geq 0}$
[3]	c_weightnull	BC [PER, ORG, LOC]	c	$\{0,1\}$
[4]	c_wp_occs	BC [PER, ORG]	c	$\mathbb{R}_{\geq 0}$
[5]	c_wp_occs_rel	BC [PER, ORG, LOC]	$c, c\star$	$\mathbb{R}_{\geq 0}$
[6]	c_wp_occs_m	BC [PER, ORG, LOC]	m, c	$\mathbb{R}_{\geq 0}$
[7]	c_titlepar_in_m_suw	BC [PER, ORG]	m, d, c	$\{0,1\}$
[8]	c_name_in_doc	BC [PER, ORG, LOC]	d, c	$\{0,1\}$
[9]	c_noname	NIL	$d, c\star$	$\{0,1\}$
[10]	c_name_m_dist	BC [PER, ORG, LOC]	m, c	$\mathbb{R}$ [0..1]
[11]	c_aleda_type_in_m_types	BC [PER, ORG, LOC]	m, c	$\{0,1\}$
[12]	c_gender_equals_m_gender	BC [PER]	m, c	$\{0,1\}$
[13]	c_name_m_longer_dist	BC [PER, ORG, LOC]	m, d, c	$\mathbb{R}$ [0..1]
[14]	c_multi_m	BC [PER, ORG, LOC]	$m\star, c$	$\{0,1\}$
[15]	c_iptc_d_iptc_sim	BC [PER, ORG]	d, c	$\mathbb{R}$ [0..1]
[16]	c_slugs_d_slugs_sim	BC [PER, ORG]	d, c	$\mathbb{R}$ [0..1]
[17]	c_sw_d_sw_sim	BC [PER, ORG]	d, c	$\mathbb{R}$ [0..1]
[18]	c_sw_d_locvars_sim	BC [PER, ORG]	d, c	$\mathbb{R}$ [0..1]
[19]	c_sw_m_suw_sim	BC [PER, ORG]	m, d, c	$\mathbb{R}$ [0..1]
[20]	c_peers_vars_d_mentions_sim	BC [PER, ORG]	d, c	$\mathbb{R}$ [0..1]
[21]	c_peers_vars_d_sw_sim	BC [PER, ORG]	d, c	$\mathbb{R}$ [0..1]
[22]	c_d_sim	BC [PER, ORG]	d, c	$\mathbb{R}$ [0..1]
[23]	c_sim_level	BC [PER, ORG], NIL, NAE	$d, c\star$	$\mathbb{R}$ [0..1]
[24]	c_countrydiv	BC [LOC]	c	$\{0,1\}$
[25]	c_countrycode_equals_d_countrycode	BC [LOC]	d, c	$\{0,1\}$
[26]	c_country_equals_d_country	BC [LOC]	d, c	$\{0,1\}$
[27]	c_size	NAE	$m, c\star$	$\mathbb{N}_{\geq 0}$
[28]	c_nil_sole	NAE	$m, c\star$	$\{0,1\}$
[29]	m_conf	NAE	m, d	$\mathbb{R}$ [0..1]
[30]	m_spos1	NAE	m, d	$\{0,1\}$
[31]	m_wpoccs	NAE	m	$\mathbb{R}_{\geq 0}$
[32]	m_doccs	NAE	m, d	$\mathbb{N}_{\geq 0}$
[33]	m_inlefff	NAE	m	$\{0,1\}$
[34]	m_inlefff_nocap	NAE	m	$\{0,1\}$
[35]	m_dagu	NAE	m, d	$\{0,1\}$
[36]	m_short	NAE	m, d	$\{0,1\}$
[37]	m_long	NAE	m, d	$\{0,1\}$
[38]	m_nolongerexists	NAE	m, d	$\{0,1\}$
[39]	m_multiner	NAE	m, d	$\{0,1\}$

TABLE 6.3 : Traits utilisés pour l'apprentissage supervisé du Liage.
1 : Le calcul des valeurs de traits dépend selon les cas et pour chaque paire de la mention (m), du candidat (c), de l'ensemble des candidats de la mention ($c\star$), du document (d), de l'ensemble des mentions du document ($m\star$).

Dans la configuration non-déterministe (cas 4), le Liage est appliqué à l'ensemble des mentions, et les lectures finales sont obtenues après un ordonnancement au sein des zones d'ambiguïté, en fonction du score des candidats (BC, NIL ou NAE) placés au rang 1 dans chaque chemin de ces zones. Hors des zones d'ambiguïté, une lecture est obtenue comme dans les cas 1 à 3.

Plus précisément, l'ordonnancement des chemins au sein de zones d'ambiguïté est obtenu à partir des scores assignés à chaque chemin d'une zone. Pour un chemin présentant une seule transition correspondant à une mention, le score est celui de l'entité (BC ou NIL) placée au rang

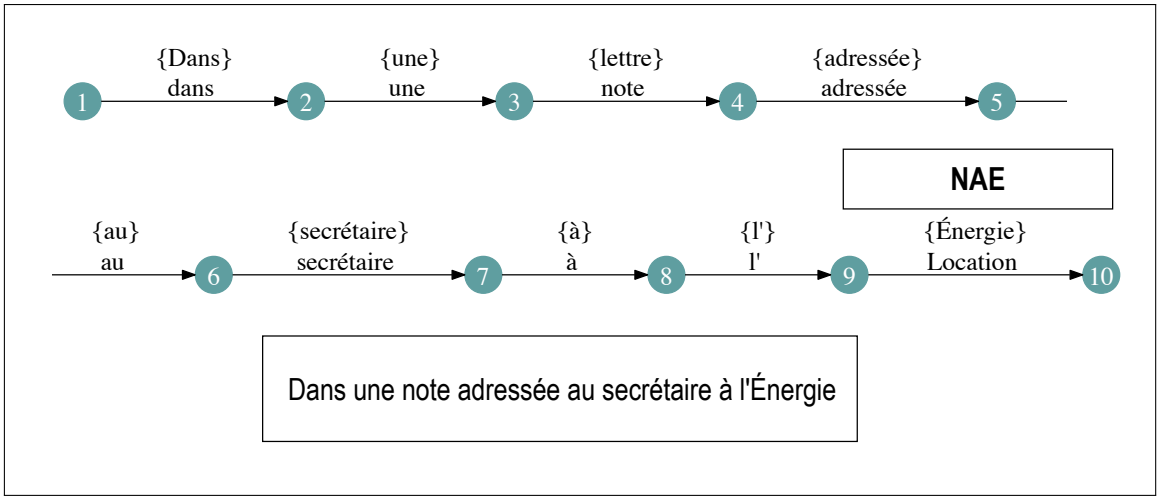


Figure 6.8 : Exemple de lectures obtenues dans le cas 3 avec élimination d'un segment non dénotationnel.

1 pour cette mention. L'alignement de la mention peut par ailleurs placer le cas NAE au rang 1, également avec un score. Pour un chemin présentant plus d'une transition correspondant à une mention, le score est la valeur minimale parmi les scores des alignements de rang 1 pour l'ensemble des mentions du chemin. Autrement dit, pour un chemin p présentant n mentions m_i avec $n \geq 1$, on calcule un score s_p tel que :

- chaque mention m_i est alignée sur une solution a_i, avec $a_i \in E_{ext}$ (entités issues de la BC, NIL ou NAE),
- chaque alignement a_i est muni d'un score s_i.
- $s_p = \min_{1 \leq i \leq n} s_i$. Le score d'un chemin est donc pénalisé par les scores faibles qui le composent, avec l'idée que de faibles probabilités d'alignement au niveau de certains segments réduisent d'autant plus la vraisemblance globale du chemin. Dans le cas où $n = 1$, on a $s_p = s_1$.

Au sein d'une zone d'ambiguïté, la lecture choisie correspond au chemin au score s_p le plus élevé, comme l'illustre la figure 6.9.

3 Expériences et évaluation

Les différents cas d'identification présentés à la section 1 donnent lieu à des expériences pour la conception du système Nomos et son évaluation :

Cas 1 REN Gold, candidat NAE non inclus ;

Cas 2 REN automatique déterministe de LIAne (Ll), candidat NAE non inclus ;

Cas 3 REN automatique déterministe de LIAne (Ll), candidat NAE inclus ;

Cas 4 REN automatique non déterministe de LIAne (LN), SxPipe/NP (SA), de leur union (LNSAU) ou de leur intersection (LNSAI), candidat NAE inclus.

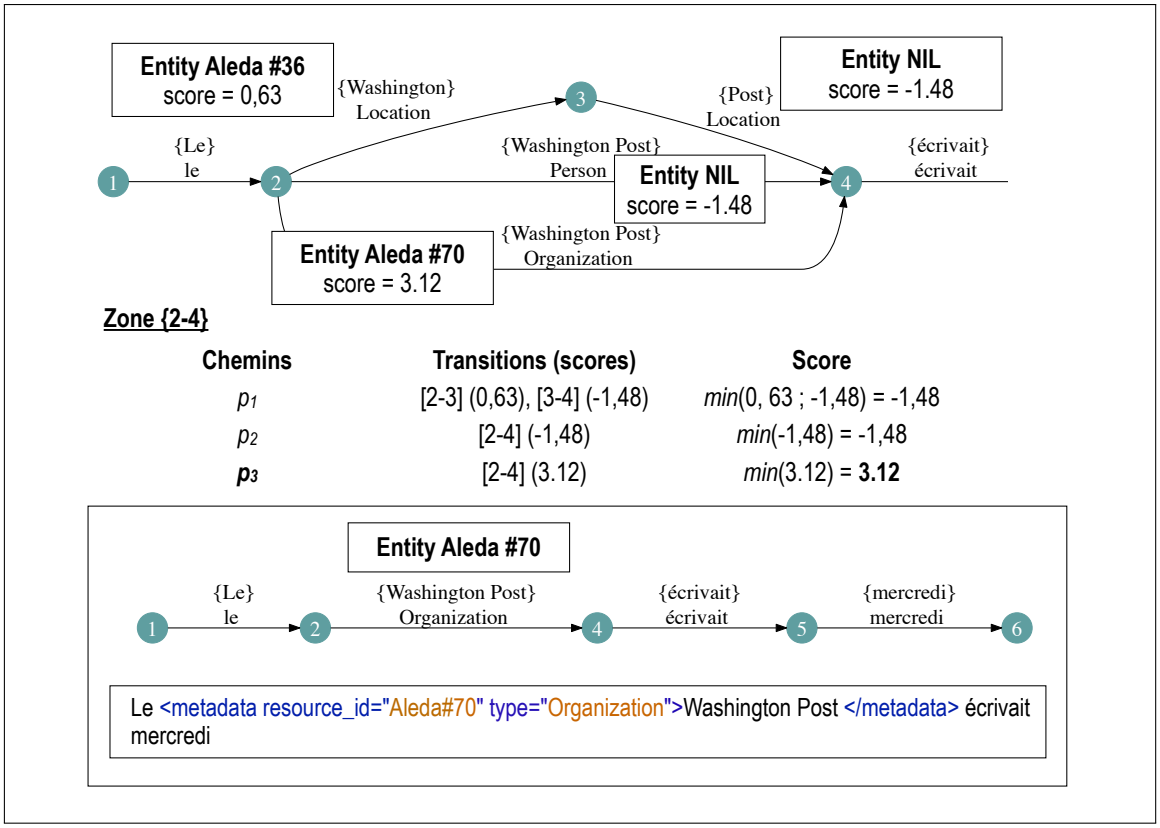


FIGURE 6.9 : Exemples de lectures obtenues dans le cas 4.

Comme cela a été évoqué, plusieurs modèles peuvent être développés pour le cas 2 afin d'évaluer l'efficacité de l'introduction du candidat NAE dans les cas 3 et 4. Il s'agit de comparer l'approche dite abstentionniste, dans laquelle le système peut ne retourner aucun alignement dans certains cas, aux modèles tenant compte de façon explicite des cas de mentions non dénotationnelles à l'aide du candidat NAE. Des expériences sont donc également menées dans le cas 2 avec la REN de L1, sans le candidat NAE, avec les modèles suivants :

- Application du meilleur modèle obtenu pour le cas 1 sur les analyses de REN de L1, sans NAE (L1a) ;
- Développement d'un modèle propre à L1 (L1b) selon la procédure commune à toutes les REN (cf. section suivante).

À partir de ces deux configurations, on peut dériver deux autres configurations par l'application d'un seuil de score égal à 0,5 ; les alignements retournés par chacun des modèles sur les données de test et dont le score est inférieur à ce seuil sont éliminés. Les deux modèles dérivés sont notés L1a′ et L1b′. On y interprète l'élimination comme le cas NAE, bien qu'il s'agisse davantage d'une abstention ne spécifiant pas la cause de l'absence d'alignement.

3.1 Conception des modèles

La conception des modèles d'apprentissage supervisé pour le composant de Liage repose sur les éléments suivants :

1. Corpus d'apprentissage et de test GAFP (cf. chapitre 5, section 2.1) : dans les expériences menées ici, les dépêches de ce corpus sont partagées en dix lots de taille égale afin de mener une évaluations par validation croisée, avec dix divisions distinctes des données pour l'apprentissage et le test. Chaque portion de données de test correspond donc à environ 10% du corpus soit 10 dépêches, et chaque portion de données d'entraînement à environ 90% soit 86 dépêches.

2. Mode d'obtention des mentions sur le corpus GAFP : mentions de l'annotation de référence (REN Gold) ou résultats d'un module de REN, qui correspond aux configurations L1, LN, SA, LNSAU ou LNSAI présentées ci-avant. Chaque configuration résulte en un ensemble différent de mentions. On cherche notamment à identifier la configuration de REN la plus efficace pour la tâche d'identification, autrement dit celle permettant de maximiser les performances de détection et d'alignement des mentions étant donné un modèle de Liage.

3. Sélection des candidats c issus de la BC pour chaque requête r à aligner. Il s'agit ici d'entités pour lesquelles un ensemble de variantes possibles est pré-établi de façon statique dans la base Aleda, retournées pour une requête si la mention est incluse dans cet ensemble. Comme évoqué précédemment, la qualité du Liage pour les entités issues de la BC dépend en partie de l'efficacité de cette sélection, formulé en tant que *rappel* des candidats et ne peut dépasser le taux de ce rappel.

4. Traits caractérisant chaque paire (r, c) sous la forme d'un vecteur. Plus précisément, il s'agit de déterminer parmi eux les traits permettant d'obtenir le meilleur modèle, tant du point de vue de l'identification des mentions d'entités que du repérage des faux positifs lorsque la configuration le prévoit.

5. Inclusion d'un candidat NAE pour la distinction des faux positifs, et dans une certaine mesure des correspondances partielles, possiblement retournés par la REN. La configuration de REN L1 est testée avec et sans NAE afin d'évaluer la pertinence de cette solution.

Les points 2 et 4 constituent des éléments variables dans la conception du modèle d'identification et donnent donc lieu à plusieurs configurations d'expériences. Celles-ci cherchent à déterminer le module de REN et le sous-ensemble de traits qui, combinés, donnent le modèle le plus efficace.

Un processus de sélection des traits les plus adaptés doit donc être menée pour chaque configuration de REN. L'ensemble proposé compte 39 traits (table 6.3 et section d'annexe B.1), ce qui correspond potentiellement à 2^{39} sous-ensembles à tester pour chacune des sept configurations de REN, incluant L1 sans NAE. Devant l'impossibilité pratique de procéder de façon exhaustive à la sélection du meilleur sous-ensemble, on mène une exploration heuristique reposant sur deux algorithmes dits *gloutons*. Ceux-ci consistent soit à ajouter les traits disponibles un à un, soit à retirer un à un les traits à partir de l'ensemble, tant qu'une addition ou une ablation, respectivement, fait progresser les performances du système.

Baselines On définit une baseline pour chaque type d'exploration :

- Sélection par addition : on considère le sous-ensemble de départ B^2 avec les traits c_wpw et c_geonw ([1] et [2]), qui modélisent la notoriété des entités de la BC. Cette caractéristique indique une approximation d'un *sens par défaut* de la mention à aligner, comme cela a été évoqué précédemment. On apprend un système baseline à l'aide de B^2 et de chaque configuration de REN.

- Sélection par ablation : on considère l'ensemble des 39 traits proposés, B^{all}, et on apprend un système baseline à l'aide de B^{all} et de chaque configuration de REN.

Les résultats de test avec B^2 et B^{all} pour chaque configuration sont donnés à la table 6.4, selon les mesures d'évaluation introduites précédemment [5]. Ces résultats de baseline montrent, pour le

B^2		P_{REN}	R_{REN}	F_{REN}	Acc	Acc_{BC}	Acc_{NIL}	P_{ALL}	R_{ALL}	F_{ALL}	P_{NAE}	R_{NAE}	F_{NAE}
Cas 1	**Gold**	100	100	100	81,01	81,15	80,41	81,01	81,01	**81,01**	-	-	-
Cas 2	**Lla**	80,83	87,71	84,13	81,84	82,15	80,42	66,15	71,78	**68,85**	-	0	-
	Lla′	96,41	31,47	47,45	93,80	93,80	-	90,44	29,52	**44,51**	*25,88*	*94,37*	*40,62*
	Llb	80,83	87,71	84,13	81,84	82,15	80,42	66,15	71,78	**68,85**	-	-	-
	Llb′	96,35	29,19	44,81	93,54	93,54	-	90,13	27,31	**41,92**	*25,19*	*94,69*	*39,79*
Cas 3	**Ll**	80,83	87,71	84,13	81,84	82,15	80,42	66,15	71,78	**68,85**	-	0	-
Cas 4	**LN**	81,13	88,04	84,44	81,91	82,48	79,15	66,45	72,11	**69,16**	-	0	-
	SA	67,98	96,36	79,72	68,42	68,82	63,21	46,51	65,93	**54,55**	-	0	-
	LNSAU	70,35	82,38	75,89	80,66	82,19	70,93	56,75	66,45	**61,22**	-	0	-
	LNSAI	92,78	72,69	81,52	88,19	88,99	80,91	81,83	64,11	**71,89**	-	0	-

B^{all}		P_{REN}	R_{REN}	F_{REN}	Acc	Acc_{BC}	Acc_{NIL}	P_{ALL}	R_{ALL}	F_{ALL}	P_{NAE}	R_{NAE}	F_{NAE}
Cas 1	**Gold**	100	100	100	82,90	81,40	89,35	82,90	82,90	**82,90**	-	-	-
Cas 2	**Lla**	80,83	87,71	84,13	83,40	81,88	90,42	67,41	73,15	**70,16**	-	-	-
	Lla′	95,00	39,53	55,83	95,39	96,19	0	90,62	37,71	**53,26**	*27,99*	*90,00*	*42,70*
	Llb	80,83	87,71	84,13	83,54	82,06	90,42	67,53	73,28	**70,28**	-	-	-
	Llb′	96,38	38,10	54,61	96,93	96,93	-	93,42	36,93	**52,94**	*28,09*	*93,13*	*43,16*
Cas 3	**Ll**	90,72	68,66	78,16	88,16	89,62	54,55	79,98	60,53	**68,91**	41,98	66,25	51,39
Cas 4	**LN**	90,32	69,18	78,35	88,72	90,12	54,76	80,14	61,38	**69,51**	42,16	64,49	50,99
	SA	69,80	80,10	74,60	65,67	67,37	15,00	45,84	52,60	**48,99**	42,99	25,35	31,89
	LNSAU	87,42	64,63	74,32	88,83	90,38	00,00	77,66	57,41	**66,02**	53,92	71,46	61,46
	LNSAI	94,35	68,47	79,35	90,03	91,41	68,25	84,95	61,64	**71,44**	24,72	25,88	25,29

TABLE 6.4 : Sélection de traits : mesures baseline.

cas 1 (Gold), un taux de correction de l'alignement de 81,01%. Dans la tâche de Liage de TAC-KBP, cette correction est de l'ordre de 73% à 86% pour les meilleurs systèmes participant à l'édition de 2011, comme le rapportent Ji et al. [JGD11]. Ces scores ne sont pas comparables, dans la mesure où les données de test ne sont pas les mêmes dans notre tâche et dans le cadre de TAC, et plus généralement dans la mesure où nous traitons des données en français, tandis que la tâche de Liage dans TAC porte sur l'anglais. Le corpus GAFP présente par ailleurs des caractéristiques permettant de considérer notre tâche comme plus aisée, dans la mesure où ce corpus présente peu de cas NIL (18%) et n'a pas été développé selon des critères de difficulté de Liage, comme cela est le cas pour les données de TAC. On peut néanmoins constater des performances proches du cadre de TAC-KBP, qu'il s'agit d'améliorer dans les expériences menées ici. Comme cela a été évoqué précédemment, les autres configurations (cas 2 à 4) ne sont pas comparables avec la tâche de Liage de TAC, puisqu'elles reposent sur des données issues de REN automatique.

Sélection des traits L'exploration des sous-ensembles de traits se présente ainsi :

- $\Phi = \{\phi_1, \phi_2, ..., \phi_{38}\}$ l'ensemble des 38 traits proposés ;
- Φ_0 le sous-ensemble des traits défini pour le modèle baseline B^2 : $\Phi_0 = \{\text{c\_wpw}, \text{c\_geonw}\}$ pour l'exploration par addition, $\Phi_0 = \Phi$ pour l'exploration par ablation ;
- s_{base} le score du modèle baseline avec Φ_0, autrement dit la mesure F_{ALL}, qu'il s'agit d'améliorer par la meilleure sélection de traits ;

5. Les mesures concernant les performances de la solution NAE dans le cas 2, indiquées en italique, sont davantage à interpréter comme des taux d'abstention des modèles concernés. L'application d'un seuil minimal aux scores de Liage correspond en effet à une abstention du système et non à une détection explicite de faux positifs

Pour l'exploration des sous-ensembles par addition :

- Φadd est l'ensemble des traits à obtenir ; $\Phi_{add} = \Phi_0$ au début de l'exploration.
- Pour chaque ϕ_i de $\Phi \backslash \Phi_0$, on apprend un modèle μ_i avec les traits de $\Phi_{add} \cup \{\phi_i\}$.
- On retient le trait ϕ_i ayant donné le modèle μ_i dont le score s_i est le plus grand parmi tous les ϕ_i et supérieur à s_{base}. On ajoute ce trait ϕ_i à Φ_{add} et on le retire à Φ. On a alors $s_{base} = s_i$.
- L'exploration termine dès lors que pour tout ϕ_i ajouté, aucun modèle n'obtient un score supérieur à s_{base}. Le modèle final utilise les traits de Φ_{add}.

Pour l'exploration des sous-ensembles par ablation :

- Φabl est l'ensemble des traits à obtenir ; $\Phi_{abl} = \Phi_0 = \Phi$ au début de l'exploration.
- Pour chaque ϕ_i de Φ, on apprend un modèle μ_i avec les traits de $\Phi_{abl} = \Phi_{abl} \backslash \phi_i$.
- On identifie le trait ϕ_i qui, retiré à Φ_{abl}, donne le modèle μ_i dont le score s_i est le plus grand parmi tous les ϕ_i et supérieur à s_{base}. On retire ce trait de Φ_{abl} et de Φ. On a alors $s_{base} = s_i$.
- L'exploration termine dès lors que pour tout ϕ_i retiré, aucun modèle n'obtient un score supérieur à s_{base}. Le modèle final utilise les traits de Φ_{abl}.

Chaque configuration de REN donne lieu à une exploration selon les deux méthodes, et donc à un ensemble de traits et des scores propres à chaque REN.

3.2 Résultats

Les résultats de la sélection de traits pour l'apprentissage d'un modèle avec chaque configuration de REN sont donnés à la table 6.5. Les scores correspondent à l'ensemble de traits (Φ_{add} ou Φ_{abl}) ayant donné le meilleur modèle pour chaque REN. Les appariements ainsi réalisés entre Φ_{add} ou Φ_{abl} et configurations de REN, qui correspondent aux différents modèles obtenus, sont présentés de façon exhaustive à l'annexe B (section 2, table B.1).

Ces résultats donnent lieu aux observations suivantes :

- L'exploration des sous-ensembles de traits par addition permet d'obtenir, à l'exception du cas 2 avec L1a′, les modèles les plus performants, avec une marge de progression de l'ordre de 5 points par rapport à la baseline B^2. Cette marge est de l'ordre de moins de 2 points pour les progressions à partir de B^{all}.
- L'addition de traits dans le cas 1 (Gold) aboutit à une correction de 84,72%, soit une progression de 3,7 points, inférieure de moins de 4 points au taux de rappel des candidats constatés sur les données de référence avec la base d'entités Aleda (88,01%, cf. *supra*, section 2.2.1).
- Les configurations 3 et 4 avec Φ_{add} donnent avec L1, LN et LNSAI les meilleurs résultats, dépassant d'environ 2,5 points les configurations du cas 2 sans seuil d'élimination (L1a et L1b), où la solution NAE n'est pas considérée.

Φ_{add}		P_{REN}	R_{REN}	F_{REN}	Acc	Acc_{BC}	Acc_{NIL}	P_{ALL}	R_{ALL}	F_{ALL}	P_{NAE}	R_{NAE}	F_{NAE}
Cas 1	**Gold**	100	100	100	84,72	83,64	89,35	84,72	84,72	**84,72**	-	-	-
Cas 2	**L1a**	80,83	87,71	84,13	85,92	84,76	91,25	69,44	75,36	**72,28**	-	-	-
	L1a′	96,99	37,71	54,31	97,41	97,92	0	94,48	36,74	**52,90**	*28,20*	*94,37*	*43,42*
	L1b	80,83	87,71	84,13	85,92	84,76	91,25	69,44	75,36	**72,28**	-	-	-
	L1b′	96,85	33,94	50,26	98,66	98,66	-	95,55	33,49	**49,59**	*26,81*	*94,69*	*41,79*
Cas 3	**L1**	84,98	86,48	85,72	86,62	85,37	92,37	73,61	74,90	**74,25**	81,73	26,56	40,09
Cas 4	**LN**	87,64	83,88	85,71	87,60	86,88	91,35	76,77	73,47	**75,08**	72,08	43,83	54,51
	SA	72,59	96,94	83,02	71,43	71,84	66,04	51,85	69,25	**59,30**	92,38	14,70	25,36
	LNSAU	81,75	81,27	81,51	86,88	86,55	89,17	71,03	70,61	**70,82**	81,25	44,20	57,25
	LNSAI	92,95	72,82	81,66	91,96	93,17	80,91	85,48	66,97	**75,10**	-	0	-

Φ_{abl}		P_{REN}	R_{REN}	F_{REN}	Acc	Acc_{BC}	Acc_{NIL}	P_{ALL}	R_{ALL}	F_{ALL}	P_{NAE}	R_{NAE}	F_{NAE}
Cas 1	**Gold**	100	100	100	84,59	83,64	88,66	84,59	84,59	**84,59**	-	-	-
Cas 2	**L1a**	80,83	87,71	84,13	85,25	84,13	90,42	68,90	74,77	**71,72**	-	-	-
	L1a′	80,83	87,71	84,13	85,25	84,13	90,42	68,90	74,77	**71,72**	-	-	-
	L1b	80,83	87,71	84,13	85,47	84,22	91,25	69,08	74,97	**71,91**	-	-	-
	L1b′	96,99	33,49	49,78	98,64	98,64	-	95,67	33,03	**49,11**	*26,71*	*95,00*	*41,70*
Cas 3	**L1**	90,81	68,08	77,81	91,21	92,81	55,56	82,83	62,09	**70,98**	41,47	66,87	51,20
Cas 4	**LN**	90,64	68,60	78,09	90,52	92,00	54,76	82,04	62,09	**70,69**	41,58	65,83	50,97
	SA	73,58	94,15	82,60	71,75	72,52	59,77	52,79	67,56	**59,27**	65,75	18,62	29,02
	LNSAU	87,96	68,86	77,24	90,56	91,71	74,29	79,65	62,35	**69,95**	57,96	70,47	63,60
	LNSAI	94,53	70,81	80,97	91,74	92,94	79,59	86,72	64,95	**74,28**	41,51	25,88	31,88

TABLE 6.5 : Résultats des différents modèles de Nomos après sélection de traits.

- Les configurations du cas 2 avec seuil d'élimination (L1a′ et L1b′) obtiennent des résultats montrant l'inefficacité de l'utilisation de ce seuil dans la tâche d'alignement et de repérage des faux positifs de REN. Ce critère d'élimination est en effet le même pour tous les cas à prédire et ne tient pas compte de la configuration particulière de chaque ordonnancement. Plusieurs entités candidates peuvent en effet se présenter avec un degré de similarité collective bas sans que la mention en question doive être considérée comme non dénotationnelle. L'apprentissage supervisé pour l'ordonnancement assigne par ailleurs la classe négative à la majorité des exemples, puisque, sur un ensemble de candidats pour une mention donnée, seule une solution (issue de la BC, NIL ou NAE) reçoit le label positif ; le modèle ainsi appris peut avoir tendance à assigner un score inférieur à 0,5 à la majorité des cas à prédire. Un trop grand nombre d'analyses est ainsi éliminé lorsque ce seuil est appliqué.

- Dans les configurations 3 et 4, la solution NAE permet d'éliminer des lectures non dénotationnelles dans un certain nombre de cas, de façon variable selon les modèles. Le rappel est dans tous les cas largement inférieur à la précision, indiquant que des critères de repérage de faux positifs manquent à l'ensemble des traits proposés pour cet aspect de la tâche. On peut observer à ce titre que si ce rappel présente une large marge d'amélioration, il est d'autant plus important de maintenir une précision P_{NAE} élevée, afin de minimiser l'introduction d'erreurs par cette solution dans l'effort mené pour éliminer les erreurs d'autres étapes du traitement.

- La configuration 4, où la REN est non-déterministe, présente avec LN et LNSAI des résultats supérieurs à la configuration 3 avec L1, déterministe. L'écart est cependant faible (moins de 1 point), et on peut rappeler que LNSAI ne présente quasiment aucune d'ambiguïté d'analyse. On note néanmoins que la précision (P_{REN} et P_{ALL}) est supérieure à L1 dans les configurations non-déterministes LN et LNSAI, indiquant que l'approche jointe est efficace quant à la reconnaissance de mentions non dénotationnelles, qui auraient éventuellement été conservées sans analyse alternative par une REN déterministe.

- Les configurations avec SA et LNSAU donnent des résultats décevants, en particulier SA

appliqué de façon isolée pour la REN. Ces scores moins encourageants que dans les autres configurations peuvent être interprétés comme un effet de la conception de SxPipe/NP qui, en tant que module de la chaîne d'analyse en cascade SxPipe, présente de façon inhérente des ambiguïtés touchant les mentions d'entités. SxPipe/NP est effet prévu pour fonctionner en amont d'un module de désambiguïsation, npNormalizer, et ce de façon étroitement liée. L'analyse des résultats retournés par NP est ainsi mal appréhendée par Nomos, tandis que le module npNormalizer, développé conjointement à NP, permet d'obtenir des résultats largement supérieurs (cf. chapitre 5, section 3.2 et *infra*). Une analyse montre qu'un nombre significatif d'erreurs dans la configuration SA correspond à des détections de mentions partielles, telles que :

- <*Banque Centrale*> *du* <*Mexique*>, où la mention de référence est <*Banque Centrale du Mexique*>,
- <*Didier*> <*Guillemot*> où la mention de référence est <*Didier Guillemot*>,
- <*Kansas City*> <*Star*> où la mention de référence est <*Kansas City Star*>.

- Dans toutes les configurations, des erreurs récurrentes d'alignement de mentions vers des entités issues de la BC sont à constater. Il s'agit principalement de cas où l'entité effectivement dénotée est absente d'Aleda et où l'une des autres entités candidates est sélectionnée, quand NIL devrait l'être. Ce type d'erreur correspond donc à la fois à un défaut de rappel des candidats et à la non identification du cas NIL. Les exemples les plus fréquents concernent des institutions existant dans plusieurs pays sous la même dénomination : *Cour suprême*, *Sénat*, *Chambre des représentants*... . La mention *Cour suprême* apparaît trois fois dans le corpus GAFP avec trois dénotations différentes :

  - la Cour suprême des États-Unis, dont l'article Wikipedia a l'URL `http://fr.wikipedia.org/wiki/Cour_suprême_des_États-Unis`
  - la Cour suprême justice colombienne, dont l'article Wikipedia a l'URL `http://fr.wikipedia.org/wiki/Cour_suprême_de_justice_(Colombie)`,
  - la Cour suprême russe, pour laquelle il n'existe à ce jour pas d'article Wikipedia.

  Ces trois entités ne figurent pas dans la base Aleda : les deux premières ont été manquées lors du processus d'importation à partir de Wikipedia, la troisième ne pouvait être importée puisqu'absente de Wikipedia. Mais plusieurs autres entités pouvant être dénotées par la variante *Cour suprême* sont en revanche présentes dans Aleda : la Cour suprême du Canada, celle du Royaume-Uni ou du Pakistan, notamment. Chacune de ces entités est donc candidate pour l'alignement de cette mention et, dans la majorité des cas, la Cour suprême britannique est sélectionnée pour l'alignement.

Comme évoqué précédemment, toute correspondance partielle est prise en compte dans l'évaluation de toutes les configurations comme un faux positif, affectant le score de précision ; la mention de référence partiellement reconnue est alors considérées comme non reconnue par l'évaluation et affecte ainsi le score de rappel. Ce mode d'évaluation s'oppose aux cas où les correspondances partielles reçoivent des scores inférieurs à ceux des résultats entièrement corrects mais contribuent néanmoins positivement au score final. Dans notre cas, les correspondances partielles contribuent négativement à ce score, et ce doublement.

Meilleurs modèles

LNSAI On constate avec la configuration LNSAI que l'amélioration apportée par les traits de Φ_{add} touche surtout les performances de l'alignement, les scores de REN étant très proches de

Cas 4	P_{REN}	R_{REN}	F_{REN}	Acc	Acc_{BC}	Acc_{NIL}	P_{ALL}	R_{ALL}	F_{ALL}	P_{NAE}	R_{NAE}	F_{NAE}
LN	87,64	83,88	85,71	87,60	86,88	91,35	76,77	73,47	**75,08**	72,08	43,83	54,51
LNSAI	92,95	72,82	81,66	91,96	93,17	80,91	85,48	66,97	**75,10**	-	0	-

TABLE 6.6 : Résultats des meilleurs modèles de Nomos.

la baseline correspondante. Celle-ci est déjà élevée au niveau de la précision (P_{REN}), ce qui découle directement de l'utilisation de l'intersection des résultats des deux systèmes LIANE et SxPipe/NP. Leur accord mutuel permet en effet d'écarter un certain nombre de lectures erronées ; le score de rappel (R_{REN}) montre cependant que cette élimination est trop drastique. Cette précision élevée s'accompagne de taux nuls au niveau de la précision P_{NAE} et du rappel R_{NAE}, signifiant que la solution NAE n'a jamais été appliquée. Ce fonctionnement est satisfaisant dans le cas présent, où très peu de faux positifs demeurent après l'application de l'étape de REN. La solution NAE, en quelque sorte inactive, n'introduit en effet pas de bruit dans les résultats. On note également que la précision P_{REN} s'accompagne d'une bonne précision P_{ALL}, obtenue avec une correction de l'alignement également élevée.

LN Les traits de Φ_{add} concernent ici à la fois les résultats de REN et ceux de la tâche globale. Pour la REN, le modèle obtenu avec LN gagne en précision (P_{REN}) de façon significative mais perd presque autant en rappel (R_{REN}). La précision P_{ALL} et le rappel R_{ALL} augmentent en revanche tous deux de 10 à 12 points. Les scores P_{NAE} et R_{NAE} montrent que la solution NAE s'est appliquée dans moins de la moitié des cas de faux positifs, avec une précision raisonnable. Si ce taux de rappel est relativement bas, on peut noter qu'il est à considérer en regard d'une configuration où aucun faux positif n'est repéré en tant que tel. On observe en effet une précision (P_{REN} et P_{ALL}) supérieure d'environ 7 points à celle obtenue avec un modèle sans élimination de faux positifs (L1b) et dont le rappel R_{ALL} est proche.

Le modèle avec LNSAI se distingue en particulier par sa bonne précision (P_{REN} et P_{ALL}). Celle du modèle obtenu avec LN lui est inférieure, mais les mesures de rappel et de précision y sont plus équilibrées. Ces deux modèles, dont les scores respectifs sur la tâche globale d'identification sont supérieurs à 75%, peuvent ainsi être employés dans les différentes applications envisagées selon le critère considéré comme le plus pertinent : pour des tâches entièrement automatisées, telles que l'annotation d'archives de dépêches en grandes quantité, la précision et donc le modèle avec LNSAI pourront être privilégiés ; dans le cadre d'un enrichissement de documents traités un à un, simultanément à leur production et avec une validation manuelle des résultats d'identification, le modèle avec LN pourra être plus utile, notamment du fait de son meilleur rappel.

Comparaison avec le système baseline

Une comparaison des résultats obtenus par les expériences menées pour le développement du système Nomos peut être établie avec le système NPNORMALIZER, présenté au chapitre 5 (section 3.2). NPNORMALIZER peut être vu comme un système baseline pour l'identification, dont les modalités de développement diffèrent fortement de l'approche adoptée pour Nomos. Les résultats d'évaluation de NPNORMALIZER sur le corpus GAFP, mis en regard des deux meilleurs modèles obtenus pour Nomos, sont reproduits à la table 6.7. Nomos permet de dépasser le score de NPNORMALIZER avec un écart inférieur à 1,5 point, représentant une réduction du taux d'erreur de 6%. On peut observer que les configurations LNSAI et LN fournissent une meilleure précision et un meilleur rappel, respectivement, que NPNORMALIZER, au niveau de la REN et du Liage, permettant ainsi de favoriser l'un ou l'autre alternativement à NPNORMALIZER pour des tâches spécifiques d'identification. Au niveau de la précision, nous avons pu observer que NPNORMALIZER détecte 25%

	P_{REN}	R_{REN}	F_{REN}	Acc	Acc_{BC}	Acc_{NIL}	P_{ALL}	R_{ALL}	F_{ALL}	P_{NAE}	R_{NAE}	F_{NAE}
npNormalizer	87,75	78,22	82,71	88,78	89,71	80,00	77,90	69,44	**73,43**	-	-	-
LN	87,64	83,88	85,71	87,60	86,88	91,35	76,77	73,47	**75,08**	72,08	43,83	54,51
LNSAI	92,95	72,82	81,66	91,96	93,17	80,91	85,48	66,97	**75,10**	-	0	-

Table 6.7 : Comparaison des meilleurs modèles de Nomos avec npNormalizer.

des faux positifs retournés par le module NP, soit un rappel R_{NAE} inférieur à celui de LN. L'apport des développements liés à Nomos réside également dans le processus de développement des modèles, qui a permis d'examiner de façon plus précise divers aspects du problème du Liage et du repérage de résultats de REN non dénotationnels, à l'aide de traits dérivés de connaissances explicites sur les entités et selon une procédure de sélection systématique.

Comparaison avec l'état de l'art

Autres expériences et évaluations

Comme indiqué précédemment, l'un ou l'autre des meilleurs modèles obtenus à l'issue des expériences présentées ici peut être utilisé dans les configurations applicatives de l'identification d'entités, autour de la tâche d'enrichissement de contenus, notamment si une précision élevée est privilégiée sur la performance générale. Le système Nomos fait ainsi l'objet d'usages et donc d'évaluations dans quelques unes de ces configurations applicatives, sur lesquelles porte le chapitre suivant. Ces évaluations portent également sur une comparaison entre Nomos et npNormalizer. Un point important de l'évaluation d'un système tel que Nomos est en effet la capacité de généralisation de ses résultats : ceux-ci peuvent être satisfaisants lors des expériences telles que celles présentées dans ce chapitre en raison d'un biais mal identifié ; les évaluations menées sur de nouvelles données et hors de la configuration usuelle permettront de vérifier si de tels problèmes existent.

D'autres expériences de développement et d'évaluation pourront également être menée sur le corpus GFTB, présenté à l'annexe C. Celui-ci est le résultat d'une annotation manuelle selon les mêmes modalités que dans le cas du corpus GAFP. Les articles du journal *Le Monde* constituant ce corpus datant des années 1990, il présente une couverture en termes d'entités différente du corpus GAFP, constituant ainsi un enjeu quant à l'adéquation des ressources employées dans la tâche d'identification réalisée par Nomos. Les expériences à venir avec le corpus GFTB seront également présentées à l'annexe C.

Chapitre 7

Applications : acquisition de métadonnées et enrichissement de dépêches AFP

Plusieurs aspects applicatifs ont été explorés autour de l'enrichissement de contenus textuels dans le cadre du Medialab de l'Agence France-Presse. Les travaux menés dans ce contexte industriel concernent principalement l'intégration de la tâche d'identification d'entités pour la production de métadonnées dites *sémantiques*, permettant la livraison de contenus enrichis. Afin d'ancrer ces contenus dans un modèle propre à l'AFP et de favoriser leur intégration dans le cadre du Web Sémantique, les métadonnées ainsi obtenues font l'objet d'une collecte vers une base dédiée. Cette base de référence est destinée à la représentation formelle des métadonnées ainsi qu'à leur mise à disposition en conformité avec les standards du Web Sémantique (section 1).

L'enrichissement des contenus par l'Annotation Sémantique donne lieu à une indexation à partir des métadonnées, ainsi qu'à des fonctionnalités de Recherche d'Information centrée sur les entités (section 2). Des applications plus ciblées d'accès à l'information enrichie en métadonnées peuvent également être conçues, comme c'est le cas pour le moteur de recherche de citations développé par le Medialab (section 3) où l'identification d'entités complète de façon cruciale l'analyse de contenus proposée.

1 Création et population d'un référentiel de métadonnées pour l'AFP

Comme cela a été abordé au chapitre 4 (section 3.3.3), la mise en place de l'enrichissement de contenus textuels à l'AFP s'accompagne du développement d'un modèle d'ancrage sémantique des métadonnées produites.

1.1 AFP Metadata Ontology (AMO)

Usages Le modèle ontologique développé pour l'AFP consiste en une spécification de la nature et du périmètre des métadonnées à retenir afin de former un ensemble de référence pour le métier de l'Agence. Le référentiel défini par ce modèle, sous le nom de *AFP Metadata Ontology* (AMO), occupe dans le processus d'enrichissement de contenus les fonctions de stockage des métadonnées associées aux documents traités, puis de mise à disposition de ces métadonnées pour des usages internes, liés à l'indexation, la consultation ou la documentation, et des usages externes : cet ensemble de référence est également destiné à intégrer la production dans le cadre du Web Sémantique, par le biais d'une formalisation conforme aux standards développés dans ce domaine.

Le référentiel AMO peut être vu comme une base de connaissances dont les entrées principales, les métadonnées, correspondent aux entités identifiées lors du processus d'enrichissement et au sujet desquelles un certain nombre d'informations sont rassemblées selon la représentation ontologique adoptée. AMO constitue ainsi le versant orienté vers les *données*, au sens des Linked Data, de l'enrichissement de contenus, tandis que l'Annotation Sémantique se place du côté des processus mis en œuvre sur les contenus eux-mêmes.

L'Annotation Sémantique fait intervenir des ressources issues des Linked Data (la base d'entités Aleda, reposant sur Wikipedia et GeoNames) à partir desquelles sont identifiées les entités mentionnées dans les contenus. Les métadonnées résultant de cette identification sont ainsi ancrées dans les ressources adoptées, dont la couverture englobe potentiellement l'ensemble de la production. Comme évoqué au chapitre 4 (section 3.3.3), AMO ne se substitue pas à ces ressources, mais en intègre une sous-partie, correspondant à la fois à des entités recensées dans ces ressources, mentionnées dans les contenus et jugées pertinentes pour la représentation référentielle propre à l'AFP. Le sous-ensemble d'entités commun aux ressources employées par l'identification et à AMO donne ainsi lieu à une double représentation, relevant d'un côté du schéma d'Aleda et de la base de connaissances associée au système Nomos (Nomos-kb), et de l'autre du schéma d'AMO. Ce dernier est directement lié à l'usage des entités représentées dans la production : il inclut donc des éléments de contexte, correspondant aux occurrences de chaque entité représentée dans des contenus de l'Agence.

Dans l'état actuel des travaux liés au référentiel AMO, seule la création d'instances à partir des entités annotées comme métadonnées dans les contenus de l'AFP est prise en charge. AMO a cependant vocation à constituer une base de connaissances plus complète, notamment par la spécification d'attributs caractérisant les instances d'entités et de relations entre entités, dans la mesure où ces attributs et relations constituent des connaissances jugées pertinentes pour les utilisateurs du référentiel. La conception du modèle tient compte de ces possibilités de spécifications plus avancées dans de futurs développements. On peut à cet égard noter que le référentiel AMO n'a pas vocation à modéliser de façon exhaustive les descriptions des entités qu'il recense, comme chercheraient à le faire des ressources encyclopédiques et publiques dans les Linked Data telles que DBpedia. Il s'agit au contraire de délimiter un ensemble référentiel constitué des entités les plus importantes pour la production de l'AFP et de n'en représenter que certaines caractéristiques, liées de façon pertinente à leur traitement dans la production. Cet aspect de la modélisation ne prive cependant pas AMO de connaissances supplémentaires, puisque sa conformité aux standards du Web Sémantique permet d'accéder de façon directe à tout élément informatif concernant les entités dans les ressources des Linked Data, qu'il s'agisse de DBpedia ou d'autres ensembles.

Administration La population du référentiel AMO est abordée en premier lieu de façon semi-automatique, par collecte des métadonnées ajoutées aux contenus lors de l'Annotation Sémantique à l'aide des systèmes Nomos ou npNormalizer, puis validation manuelle par les journalistes des entrées jugées pertinentes pour un stockage pérenne. Ce processus fait l'objet d'une description à la section 1.3. Une gestion manuelle d'AMO est également prévue, pour l'ajout de nouvelles entités indépendamment du processus de collecte automatique, mais également et de façon plus générale afin de permettre des fonctionnalités d'administration. Il s'agit principalement de la vérification de l'intégrité des données, de la correction d'erreurs issues de processus automatiques, de la mise à jour d'éléments informatifs sujets à évolution (attributs d'entités tels que le nom, dans le cas de changements de noms d'entreprise après une restructuration, par exemple, ou relations entre entités telles que la fonction d'une personne dans une entreprise...). La validation des entités présentées comme candidates pour une entrée au référentiel, en association avec des processus automatiques ou non, est également envisagée comme partie prenante de l'activité d'administration.

Le référentiel AMO est conçu selon les spécifications énoncées au chapitre 4 (section 3.3.3), qui concernent principalement l'élaboration de son modèle ontologique et son périmètre de population.

1.2 Modèle ontologique

Le modèle ontologique du référentiel AMO a été conçu manuellement et en collaboration directe avec le Medialab, dans l'objectif de fournir une modélisation consensuelle du domaine et adaptée aux usages envisagés. De façon similaire aux modèles généralement adoptés pour les ensembles de données des Linked Data, AMO privilégie la simplicité formelle : peu de classes et relations conceptuelles sont définies, afin de maintenir le consensus quant aux choix de représentation d'une part, et de permettre une maintenance efficace d'autre part. On peut observer à ce titre qu'une complexité formelle au niveau de la représentation des entités, qui constituent ici les objets principaux de la réflexion, ne paraît pas utile étant donné la nature des traitements envisagés, qui s'intéressent aux entités dans une perspective d'information générale. Une définition sophistiquée des entités, visant à une énumération exhaustive et à une description formelle complète, se situerait hors du champ du présent travail. L'accent est mis ici sur la population d'AMO, à laquelle la formalisation simple du modèle permet de donner un accès non entravé par une représentation complexe. Le langage OWL est adopté pour l'expression des axiomes et assertions, dans sa version OWL-DL (cf. chapitre 1, section 1.2.2). La figure 7.1 donne une vue générale des classes, relations et attributs spécifiés dans AMO et décrits ci-après. Les éléments pris en charge dans le cadre des travaux présentés ici sont indiqués par une astérisque.

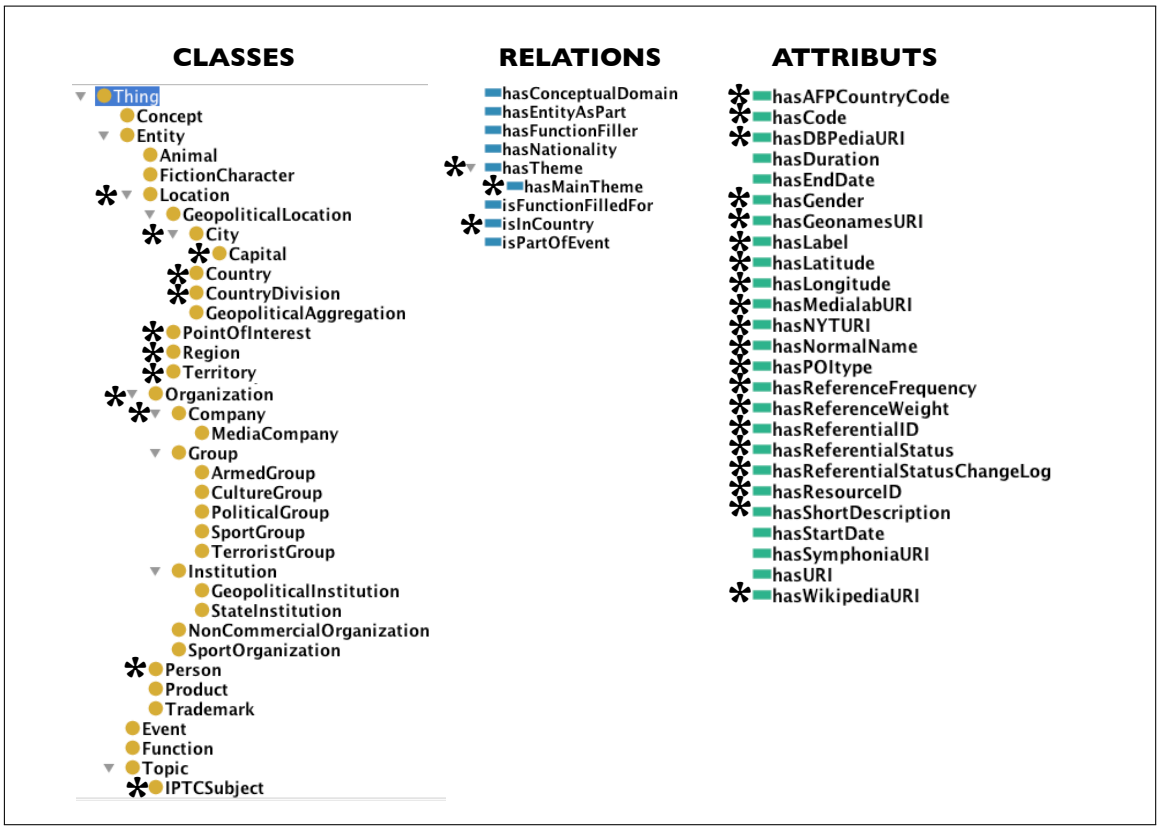


Figure 7.1 : Référentiel AMO : vue générale du modèle.

1.2.1 Entités et relations

Les entités constituent le principal ensemble d'objets destinés à être représenté dans AMO : les classes et relations conceptuelles qui y sont définies les concernent donc en premier lieu. Les classes conceptuelles définies pour les entités, modélisées comme sous-classes d'une classe générale *Entity*, correspondent d'abord de façon directe aux types d'entités intégrés aux ressources de l'Annotation Sémantique : *Person*, *Organization*, *Location* ; elles-mêmes renvoient par ailleurs aux types manipulés par les modules d'Extraction d'Information et de Reconnaissance d'Entités Nommées précédant le processus d'identification d'entités. Une distinction importante entre classe ontologique et type d'Extraction d'Information est introduite ici : la classe *Person* est déclarée comme disjointe des classes *Organization* et *Location*, empêchant l'appartenance d'une même instance à l'une et l'autre de ces classes ; *Organization* et *Location* ne sont en revanche pas déclarées comme disjointes, permettant ainsi qu'une instance d'entités puisse être définie par ces deux classes, dont les attributs et relations peuvent s'appliquer à cette instance même si elle a fait l'objet d'une déclaration explicite pour l'une seulement. Cette modélisation, permise par le formalisme ontologique, prend notamment en charge le problème de la métonymie pouvant toucher certaines entités, en particulier les lieux géopolitiques dont l'interprétation se ramène au type ORGANIZATION dans certains contextes.

Un raffinement de ces classes est proposé afin de représenter certains sous-types d'entités plus particulièrement pertinents pour les contenus de l'AFP ; d'autres types d'entités sont également intégrés au modèle, dans la perspective de traitements dépassant les types les plus usuels.

Au sein des classes d'entités, seules les classes *Person*, *Organization* et *Location* peuvent donner lieu à une instanciation directe à partir des métadonnées collectées sur les contenus annotés. Plus précisément :

- La classe *Person* correspond au type PERSON d'Aleda et aucune sous-classe n'est à ce jour spécifiée. Les entités identifiées par Nomos ou NPNORMALIZER associées au type PERSON donnent donc lieu de façon directe à des instances de la classe *Person*.

- La classe *Organization* correspond au type ORGANIZATION d'Aleda, qui présente également le type COMPANY, correspondant à la sous-classe *Company* de la classe *Organization*. Cette distinction n'était pas prise en compte dans l'évaluation de Nomos et NPNORMALIZER, les entreprises étant incluses dans le type ORGANIZATION, mais peut être faite dans le processus de collecte des métadonnées vers AMO. Les entités de type ORGANIZATION et COMPANY donnent ainsi lieu à des instances de *Organization* et *Company*, respectivement. Les autres sous-classes de *Organization* ne peuvent pas, en revanche, être directement instanciées à partir des entités identifiées dans Aleda. Un raffinement conceptuel des instances peut être envisagé dans des traitements automatiques ultérieurs, directement intégrés à l'Annotation Sémantique ou à la gestion manuelle du référentiel.

- La classe *Location* correspond au type LOCATION d'Aleda, qui spécifie par ailleurs pour les entités de ce type un sous-type extrait de GeoNames (cf. chapitre 3, section 1.3, table 3.6 page 95). Ce sous-type permet d'instancier directement les sous-classes *City*, *Capital*, *Country*, *CountryDivision*, *PointOfInterest*, *Region* et *Territory* à partir d'entités identifiées dans les contenus. Comme pour *Organization*, les autres sous-classes de *Location* ne peuvent pas être directement instanciées à partir des résultats de Nomos ou NPNORMALIZER. Aleda définit également pour chaque lieu un code ISO-3166-2 indiquant le pays d'appartenance du lieu (et le lieu lui-même lorsqu'il s'agit d'un pays). Cet attribut permet d'instancier directement la relation *isInCountry* entre les instances de *Location*, quelle que soit leur sous-classe dès lors que l'entrée d'Aleda indique ce code, et l'instance de *Country* correspondante.

Autour des entités, d'autres classes et relations sont spécifiées afin de modéliser les objets entrant en jeu dans leur représentation. Il s'agit principalement de la classe *Topic* incluant les sujets issus de la taxonomie de l'IPTC (cf. chapitre 4, section 2.2). La représentation des entités dans AMO dérive en effet de leurs occurrences dans les contenus de l'AFP, eux-mêmes catégorisés thématiquement selon cette taxonomie. Chaque entité est donc associée à ces sujets IPTC, qui la caractérisent indirectement. Une entité donnée peut être associée de façon plus significative à un sujet IPTC en fonction de la fréquence avec laquelle elle est mentionnée dans des contenus catégorisés par ce sujet ; cette association thématique est modélisée dans le référentiel AMO par le biais des relations conceptuelles (*Object Property*) *hasTheme* et *hasMainTheme*. Celles-ci sont par ailleurs les seules relations, avec la relation *isInCountry* évoquée précédemment, réalisées entre instances du référentiel AMO à ce jour.

1.2.2 Classes et relations satellites

Les autres classes d'AMO ainsi que des relations conceptuelles associées ont été incluses dans la conception initiale du modèle dans la perspective de traitements mais envisagés dans l'avenir par le Medialab. Il s'agit de la représentation des événements, concepts et fonctions, dont l'intégration aux côtés des entités suit les objectifs suivants :

Événements Les événements, tels que les sommets politiques internationaux, les compétitions sportives majeures ou les manifestations culturelles importantes, tiennent une place considérable dans le travail de l'Agence. Comme les entités, considérées ici dans une acception relative à la notion d'individu, ils constituent des éléments particulièrement structurant de l'information. Leur modélisation en tant que métadonnées de référence pour l'AFP est donc envisagée et prévue avec la classe *Event*. Considérés comme des entités nommées dans certains contextes d'Extraction d'Information, les événements peuvent faire l'objet d'une collecte à partir des contenus textuels par le biais de l'Annotation Sémantique, comme c'est le cas pour les entités. Il s'agirait alors de procéder à la reconnaissance de leurs mentions dans les contenus puis à leur identification en tant qu'instances uniques (*JO2012*, *Festival de Cannes* ou *G20*). Le rapport entre entités et événements devrait lui aussi donner lieu à une modélisation, puisque les premières interviennent comme actants dans les seconds : les entités représentées dans AMO peuvent alors associées aux instances d'événements lorsque de telles associations sont inférables à partir des contenus traités, ce qui est prévu avec les relations *hasEntityAsPart* et *isPartOfEvent*. Cette intégration dans AMO demande notamment, dans les travaux à venir, une réflexion sur le type d'événements à modéliser et sur la définition qui doit en être faite. On peut notamment s'interroger sur la granularité des événements à modéliser comme instances, en particulier pour les événements récurrents (*Festival de Cannes*, *G20*) dont il est difficile d'établir s'ils doivent donner lieu à une instance ou si chaque édition (*Festival de Cannes 2009*, *G20 2010*) constitue une instance distincte.

Concepts L'enrichissement de contenus textuels de l'AFP est présentement limité aux entités, mais peut être envisagé dans une perspective plus générale. L'Annotation Sémantique qui en est le moyen principal se définit en effet relativement à des contenus et à un modèle, dont les concepts eux-mêmes peuvent faire l'objet d'une annotation. Les concepts ou termes relatifs à des domaines particuliers (politique : *élections*, économique : *inflation*, judiciaire : *mise en examen*, ...) seraient ainsi repérés à travers leur réalisation textuelle puis associés par la classe *Concept* aux contenus traités, et par transitivité aux entités mentionnées dans ces mêmes contenus avec la relation *hasConceptualDomain*. On obtiendrait par ce biais un enrichissement de l'information disponible au sujet des entités. La réflexion principale à mener à cet égard concerne le périmètre et la profondeur de modélisation de tels concepts

dans AMO. Des ressources du type de WordNet[1], telles que le réseau lexical du français WOLF[2] peuvent être considérées afin d'enrichir de façon automatique la classe *Concept*.

Fonctions La représentation des entités dans AMO se limite présentement aux instances d'entités elles-mêmes ainsi qu'à leur association avec des sujets IPTC en fonction de leur usage dans les contenus catégorisés. Une extension de cette représentation est envisagée, en particulier en ce qui concerne les relations entretenues entre personnes et organisations, autrement dit les fonctions occupées par les premières dans les secondes. La notion de fonction peut être représentée de façon directe par un ensemble de relations conceptuelles (*Object Property*), de type *(Person, isDirectorOf, Organization)*, ce qui donnerait lieu à l'établissement d'autant de relations ontologiques que de fonctions possibles. Les fonctions sont représentées dans AMO non par des relations mais à travers une classe conceptuelle renvoyant à la notion de fonction elle-même. Cette disposition permet de modéliser aisément divers degrés de spécification par le biais de sous-classes (*membre* étant moins spécifique que *directeur*, par exemple), et ainsi d'instancier une fonction même en l'absence de précisions sur sa nature exacte. La mise en relation d'entités autour des instances de fonctions ainsi représentées peut alors se faire par l'instanciation de deux relations (*Object Property*), l'une spécifiant la personne occupant la fonction (*hasFunctionFiller*), l'autre indiquant quelle organisation est concernée (*isFunctionFilledFor*). Cette modélisation des fonctions permet également de laisser non-spécifié l'un de ces deux rôles, dans le cas où les processus d'Extraction d'Information à partir des contenus textuels échouent à identifier la totalité de la relation, par exemple. Les figures 7.2 et 7.3 illustrent ce choix de modélisation.

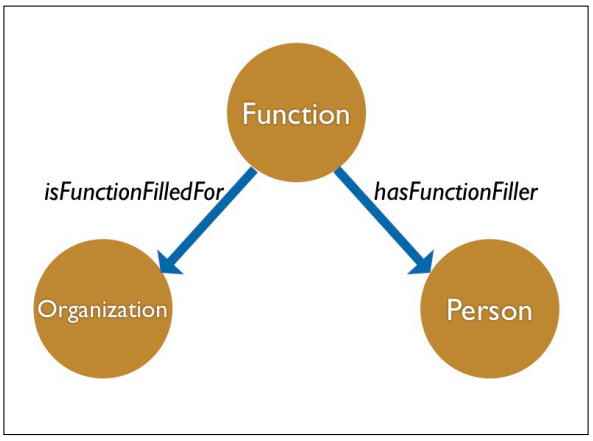


FIGURE 7.2 : Référentiel AMO : modélisation des fonctions.

1.2.3 Attributs

Les classes conceptuelles d'AMO sont associées à des attributs (*Data Property*) représentant les caractéristiques importantes des instances modélisées, aux niveaux de leur représentation conceptuelle mais également du fonctionnement des traitements dans lesquels le référentiel peut intervenir. Les attributs particulièrement importants à ce titre sont les différents identifiants et URI associés aux entités. Chacune d'elles est en effet marquée par un identifiant unique interne au

1. http://wordnet.princeton.edu/
2. http://atoll.inria.fr/~sagot/wolf.html

Figure 7.3 : Référentiel AMO : exemple de modélisation d'une fonction.

référentiel, mais également par son identifiant dans la base Aleda et l'URL Wikipedia ou GeoNames correspondante. Ces attributs permettent de maintenir une cohérence de représentation entre le référentiel et la ressource principale utilisée pour sa population, Aleda, lors des processus d'enrichissement, ainsi que le réseau des Linked Data. Ils peuvent ainsi être vus comme des liens de *synonymie*, représentés en OWL par le constructeur *owl:sameAs*. Les sens et usages des attributs définis dans AMO sont spécifiés à la table 7.1, qui précise les classes concernées par chacun d'entre eux.

1.3 Population

Une fois le modèle ontologique du référentiel AMO défini, sa population s'envisage autour de trois axes :

Production AFP Le critère de pertinence relativement au métier de l'AFP est déterminant dans le choix des entités représentées dans le référentiel. C'est en ce sens que la source principale d'entités pour sa population consiste en la production de contenus de l'AFP elle-même, à partir de laquelle les métadonnées issues du processus d'Annotation Sémantique sont collectées en vue d'une éventuelle intégration. Les entités mentionnées dans les fils de dépêches et autres contenus textuels — légendes photographiques, notamment — se présentent en effet comme les candidats les plus immédiatement susceptibles de répondre au critère de pertinence. L'identification automatique d'entités dans ces contenus, à l'aide des systèmes Nomos ou npNormalizer, est alors employée pour l'acquisition des métadonnées à même de peupler le référentiel.

Intégration des référentiels existants Plusieurs référentiels de données existent dans les différents silos de l'AFP (cf. chapitre 4, section 2.3), caractérisés par une hétérogénéité de schéma et une faible spécification sémantique. La conception du référentiel AMO se place précisément dans la perspective d'une refonte de ces ressources. L'intégration des référentiels existants dans ce référentiel unifié est indispensable afin de maintenir une cohérence de gestion documentaire au travers des traitements effectués à des époques différentes sur les contenus. Il s'agit alors de réaliser un alignement entre les schémas des ressources existantes et le modèle ontologique d'AMO afin de déterminer les règles d'importation de données à partir des premières à destination du second.

Intégration de données externes Un certain nombre de ressources externes, principalement dans le réseau des Linked Data, font l'objet une importation directe de données vers le référentiel AMO. Il s'agit alors de ressources définies relativement à des domaines et intérêts proches de ceux de l'AFP et dont le schéma conceptuel peut être mis en correspondance

Attribut	Description	Classes
hasNormalName	Nom normalisé ou canonique d'une instance	*toutes*
hasLabel	Variantes lexicales rencontrées en corpus ou indiquées dans les ressources (Aleda)	*toutes*
hasReferenceFrequency	Fréquence dans le corpus d'acquisition	*Person*, *Organization*, *Location*
hasReferenceWeight	Poids / popularité de l'entité dans Aleda	*Person*, *Organization*, *Location*
hasReferentialID	Identifiant dans la ressource d'origine (Aleda)	*Person*, *Organization*, *Location*
hasResourceID	Identifiant interne à AMO	*toutes*
hasURI	URI	*toutes*
hasWikipediaURI	URL Wikipedia indiquée pour les PERSON et ORGANIZATION dans Aleda	*Person*, *Organization*, *Location*
hasShortDescription	Brève description extraite du résumé de l'article Wikipedia	*Person*, *Organization*
hasCode	Code de catégorisation, notamment sujets IPTC	*toutes*
hasGender	Genre (masculin / féminin), indiqué pour les PERSON dans Aleda	*Person*
hasGeonamesURI	URL GeoNames, indiquée pour les LOCATION dans Aleda	
hasLatitude	Latitude, indiquée pour les LOCATION dans Aleda	*Location*
hasLongitude	Longitude, indiquée pour les LOCATION dans Aleda	*Location*
hasAFPCountryCode	Code pays ISO-3166-3 indiqué pour les lieux des référentiels AFP	*Location*
hasPOItype	Type des instances de la classe *POI* (aéroport, musée, ...)	*POI*
hasNYTURI	URI NYT, par résolution avec les données NYT	*Person*, *Organization*, *Location*
hasDBPediaURI	URI DBpedia (par résolution avec les données NYT)	*Person*, *Organization*, *Location*
hasStartDate	Date de début	*Event*
hasEndDate	Date de fin	*Event*
hasDuration	Durée	*Event*
hasReferentialStatus	Statut dans AMO (validé, obsolète, en attente de validation, ...)	*toutes*
hasReferentialStatusChangeLog	Log associé aux changements de statut dans AMO	*toutes*
hasMedialabURI	Identifiant de l'instance attribué dans AMO par le Medialab, pour information lors de l'intégration avec d'autres référentiels AFP	*toutes*

TABLE 7.1 : Description et classes des attributs d'instances dans AMO.

avec celui d'AMO. Le choix des ressources visées pour cette importation repose sur le critère de pertinence évoqué précédemment. Dans le cas de ressources recouvrant en partie la population existante d'AMO, un processus de résolution d'entités, tel que défini au chapitre 2 (section 3.3.4), doit être prévu afin de maintenir l'unicité de représentation de chaque instance d'entité.

1.3.1 Population initiale

La population initiale d'AMO vise à doter le référentiel d'un ensemble d'entités formant un noyau descriptif des contenus de l'AFP. Ce noyau peut être caractérisé en termes de notoriété des entités considérées — on cherche à représenter au moins les personnes et organisation les plus notables dans le traitement de l'actualité, y compris pour des périodes passées et au niveau français comme international. Pour les lieux, le périmètre des entités à représenter au stade initial est davantage défini par énumération : l'ensemble visé est celui des États du monde, de leurs capitales, des principales métropoles mondiales et des continents. Ces types de lieux constituent en effet un ensemble informatif transversal, susceptible d'apparaître dans tout contexte et en association avec toute catégorie thématique ; ces entités sont donc pertinentes à ce titre. Certains lieux non inclus dans cet ensemble peuvent par ailleurs faire l'objet d'une instanciation selon un critère de pertinence plus proche de la notoriété : on cherchera notamment à représenter les lieux où se sont produits des événements importants, ces lieux donnant d'ailleurs souvent leur nom aux événements en question. Il s'agit par exemple des villes de Tchernobyl en Ukraine ou Fukushima au Japon.

Les trois axes présentés ci-avant sont envisagés pour la population initiale du référentiel.

- L'acquisition d'entités à partir des métadonnées de contenus permet d'obtenir une approximation de leur notoriété. On peut en effet considérer qu'une entité mentionnée avec une fréquence élevée dans les fils d'actualité joue un rôle important dans l'information. Selon que cette fréquence est constatée sur une période limitée ou de façon plus constante, cette notoriété peut être vue comme liée à des événements particuliers — on évoquera très fréquemment certains candidats à la présidence de la République en période de campagne électorale, qui ne seront que peu ou plus mentionnés par la suite, du moins pendant quelques années — ou propre à une entité dont la célébrité est acquise sur une période longue — anciens chefs d'État, artistes les plus populaires… Les lieux constituant l'ensemble de base évoquée précédemment (pays, capitales, etc.) sont également susceptibles d'apparaître avec une fréquence élevée dans les fils de l'AFP. Dans les deux cas, la fréquence d'occurrence des entités peut être utilisée comme mesure de pertinence dès lors qu'un corpus de taille raisonnable est considéré pour ce mode d'acquisition.

- Les ressources externes donnant lieu à une importation de données vers AMO sont, d'une part, la liste des pays membres des Nations Unies [3], et d'autre part l'ensemble de données publiées par le *New York Times* (NYT) sur le réseau des Linked Data [4]. Les modalités de résolution des entités ainsi collectées avec les métadonnées acquises à partir de corpus de l'AFP sont décrites ci-après (section 1.3.2). La figure 7.4 présente un extrait des données publiées par le NYT sur les Linked Data.

- La population initiale comprend également l'intégration des référentiels existants : chacun d'eux voit son schéma mis en correspondance avec la conceptualisation adoptée dans AMO. Les éléments de chaque ensemble sont ensuite considérés comme des entités uniques et font l'objet d'une instanciation dans le référentiel. Cette intégration a été réalisée à partir des données géographiques des référentiels utilisés dans le silo texte de l'AFP (cf. chapitre 4, section 2.3). Les villes également obtenues par acquisition de métadonnées sur corpus et importation de ressources externes donnent lieu à une résolution d'entités décrite ci-après. Les autres référentiels existants doivent en revanche faire l'objet d'une validation manuelle actuellement en cours à l'AFP, en particulier la liste de noms de personnes établie par le silo photo : sur les quelques 1 300 000 labels recensés, les 277 400 labels en

3. `http://www.un.org/en/members/`
4. `http://data.nytimes.com/`

français [5] ont été présentés en tant que requêtes sur la base Aleda afin de déterminer pour chacune un ensemble d'entités candidates ; la validation en cours consiste ainsi à choisir l'entité adéquate pour chaque label, si elle est présente parmi les résultats de requêtes [6]. Cette validation s'appuie notamment sur la consultation des pages Wikipedia des entités candidates, dont les URL, fournies par Aleda, sont indiquées avec chaque résultat de requête.

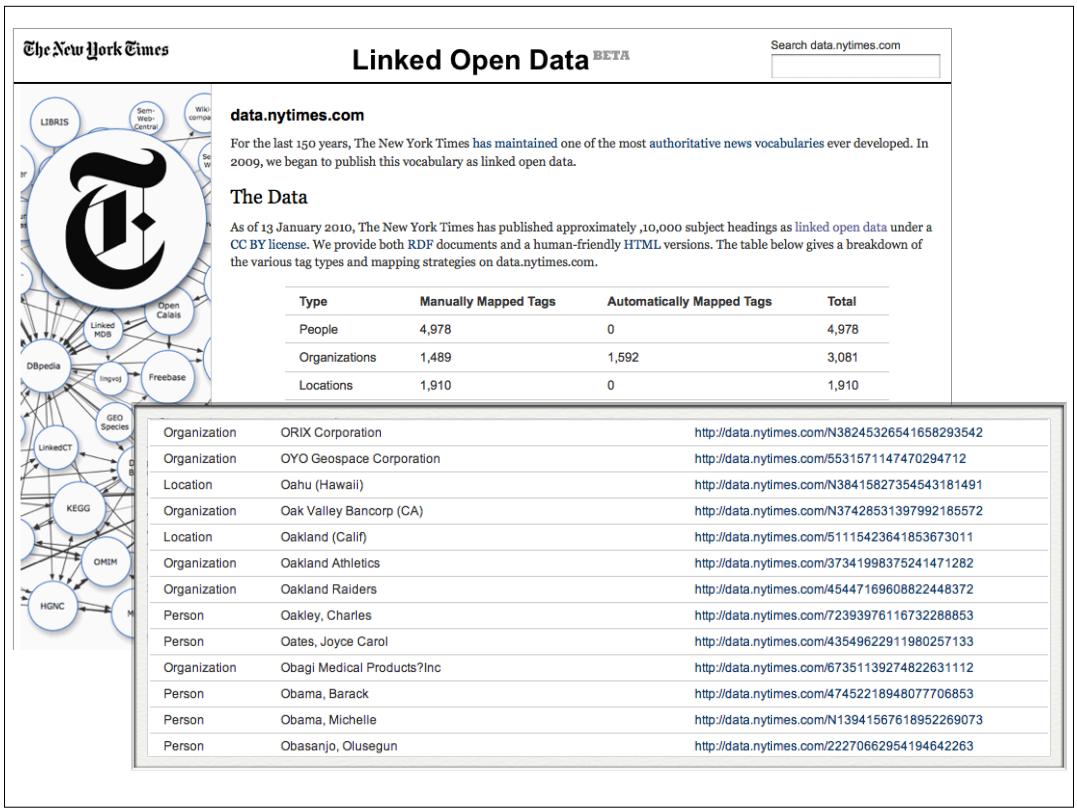


FIGURE 7.4 : Extrait des données publiées par le NYT sur les Linked Data et intégrées à AMO.

On peut noter que, dans ces trois axes méthodologiques de population, il s'agit principalement d'effectuer, entre des données sources (corpus, ressources externes ou référentiels pré-existants) et le référentiel cible, une série d'*alignements* de schémas, comme nous l'avons exposé dans [SS12b]. Chaque source de données est en effet caractérisée par un schéma qui lui est propre, pouvant être mis en correspondance avec le schéma cible afin d'y intégrer les données sources. Les types d'entités manipulés par le module de Reconnaissance d'Entités Nommées correspondent ainsi aux types de représentation dans la base de données Aleda, eux-mêmes alignés sur les classes ontologiques équivalentes d'AMO. Dans les référentiels AFP existants, les types d'entités représentées — villes et personnes — peuvent être alignées sur les classes ontologiques dont la sémantique est adéquate — *City* et *Person*. De même, dans les ressources externes, l'alignement peut intervenir entre les pays des Nations Unies et la classe ontologique *Country*, et entre les types de données du NYT — PERSON, ORGANIZATION et LOCATION — qui sont directement alignés sur les classes ontologiques de mêmes noms. L'alignement de schéma peut constituer une tâche

5. Chaque entrée du référentiel du silo photo se présente en effet sous la forme d'un label correspondant à un nom de personne ; plusieurs labels du même nom dans différentes langues sont listés avec l'indication de la langue.

6. Un même label peut être apparié à plusieurs entités si cela est considéré comme pertinent dans le processus de validation.

complexe dans des configurations d'intégration de sources de données multiples et hétérogènes mais demeure simple dans notre cas, où l'ensemble des ressources considérées forment des représentations d'entités selon des modèles peu complexes et consensuels.

1.3.2 Acquisition de métadonnées à partir de contenus textuels et validation

À partir du fil de dépêche généraliste en français de l'AFP, couvrant une période allant de début 2010 à fin 2011, soit environ 400 000 dépêches, un ensemble de métadonnées a été extrait et considéré pour une instanciation des entités correspondantes dans AMO. Ces métadonnées résultent d'une Annotation Sémantique réalisée par le système npNormalizer (cf. chapitre 5, section 3.2) dont la bonne précision était appropriée pour cette tâche, notamment en comparaison avec le système Nomos alors en cours de développement. L'extraction réalisée aboutit à une liste d'environ 240 000 entités [7], ramenée à environ 10 700 en ne retenant que les entités de type person et organization [8] apparaissant au moins 25 fois dans le corpus d'acquisition et alignées sur une entités de la BC. Les lieux donnent lieu à un traitement distinct, comme exposé ci-après. Les cas de Liage sur NIL sont écartés dans ce processus de population, puisqu'ils correspondent à des entités non encore identifiées. Le seuil minimal fixé à 25 réduit drastiquement le nombre d'entités considérées et indique qu'une majorité d'entités sont mentionnées peu souvent dans la production, tandis d'un nombre restreint d'entités constitue la majorité des mentions. Ces entités les plus couramment mentionnées forment *a priori* le corps de la population initiale visée ici. Chaque entité est soumise à la validation avec les renseignements suivants :

- Nombre d'occurrences total sur le corpus d'acquisition
- Nom canonique et variantes lexicales de leurs occurrences
- Type (person ou organization)
- URL de la page Wikipedia ou GeoNames

La validation, réalisée pour cette population initiale par les journalistes du Medialab, consiste dans cette configuration à accepter ou refuser une entité en tant qu'instance du référentiel AMO. Les entités refusées correspondent à des erreurs issues de l'identification (fausses mentions ou identification erronée) ou à des entités dont la représentation est jugée non pertinente pour le métier de l'Agence. Le nombre d'entités instanciées dans AMO à l'issue de ce processus s'élève à 3 800 personnes et 1 600 organisations, dont un exemple est reproduit à la figure 7.5 [9]. La table 7.2 présente les entités instanciées les plus fréquentes dans le corpus d'acquisition.

L'intégration des données du NYT à ce premier ensemble d'instances se traduit par d'éventuels doublons, qu'il s'agit d'identifier par un processus de résolution. Celui-ci correspond à la tâche également appelée *entity linkage*, évoquée au chapitre 2 (section 3.3.4) Pour ce faire, les entrées de chacun des deux ensembles ont été comparées une à une au niveau de l'URL Wikipedia, spécifié à la fois dans AMO à partir d'Aleda et dans les données du NYT. Cette procédure simple montre l'intérêt de la formalisation des données selon les standards du Web Sémantique et des Linked Data, en particulier la spécification de liens de synonymie entre ensembles de

7. Il s'agit bien ici d'entités uniques et non de mentions textuelles d'entités. Celles-ci s'élèvent, en nombre d'occurrences dans le corpus d'acquisition, à plus de 8 millions, parmi environ 120 millions de tokens au total. Ces résultats sont issus d'une annotation automatique et impliquent un certain taux d'erreur mais indiquent néanmoins des ordres de grandeur caractérisant les entités dans la production de l'AFP.

8. Les entités de type person et organization constituent environ 80% du total des entités extraites.

9. Les exemples extraits du référentiel AMO sont reproduits ici à l'aide du logiciel d'édition d'ontologies en OWL Protégé (`http://protege.stanford.edu/`).

données qui peuvent, comme dans ce cas, jouer le rôle de pivots de représentation et ce sans intervention manuelle. Les données du NYT présentent par ailleurs un grand nombre de liens de synonymie établis pour chaque entité recensée vers d'autres ensembles des Linked Data, en particulier GeoNames, Wikipedia et DBpedia ; les entités d'AMO résultant de la résolution sont donc également liées à DBpedia lorsque le NYT indique une telle synonymie.

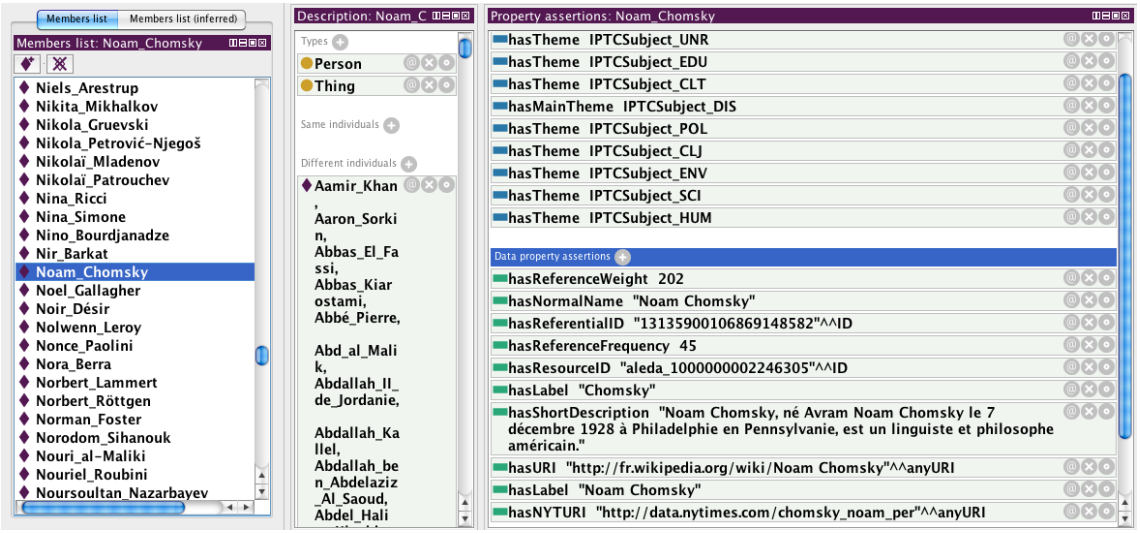


FIGURE 7.5 : Référentiel AMO : exemple d'instance (classe *Person*) issue de la validation et de la résolution d'entités après collecte sur corpus.

Les lieux ne sont pas concernés par la validation à l'issue de la collecte, dans la mesure où le périmètre pertinent repose à ce niveau sur des critères non associés à la notoriété affectant les personnes et organisations : l'ensemble des entités de type LOCATION retournés par le processus d'extraction de métadonnées a été filtré pour ne retenir que les pays et leurs capitales. Ce sous-ensemble a ensuite donné lieu à un autre processus de résolution d'entités à l'égard des lieux issus des référentiels AFP (villes incluant des capitales), des pays des Nations Unies et des lieux recensés dans les données du NYT. Pour ces dernières, une comparaison simple au niveau de l'URL GeoNames, spécifiée pour les instances d'AMO à partir d'Aleda ainsi que pour les données du NYT, permet de résoudre les entités identiques, de façon similaire à la méthode employée pour les personnes et organisations. Les pays des Nations Unies ainsi que les villes issues des référentiels AFP ont quant à eux été mis en correspondance avec les entités extraites du corpus d'acquisition sur la base de leur nom canonique ; cette correspondance approximative a ensuite été vérifiée manuellement par les journalistes du Medialab. La figure 7.6 donne un exemple d'instance de la classe *Location* ainsi représentée dans AMO.

1.3.3 Enrichissement de la population

La population du référentiel AMO est destinée à être enrichie afin de compléter le noyau d'entités initial et de prendre en compte les entités importantes émergeant au fil de l'actualité. Cette augmentation de la population est surtout envisagée selon l'axe d'acquisition de métadonnées à partir des contenus, ce qui peut donner lieu à deux types d'enrichissement :

- Par lot : comme pour la population initiale, un corpus de dépêches est constitué et l'Annotation Sémantique permet d'en extraire les entités. Celles-ci sont comparées à l'ensemble d'instances existant dans le référentiel, afin de déterminer les entités pouvant donner lieu

URL Wikipedia	fréquence
fr.wikipedia.org/wiki/Agence_France-Presse	173033
fr.wikipedia.org/wiki/Union_pour_un_mouvement_populaire	71697
fr.wikipedia.org/wiki/Barack_Obama	48158
fr.wikipedia.org/wiki/Nicolas_Sarkozy	44025
fr.wikipedia.org/wiki/Organisation_des_Nations_unies	41730
fr.wikipedia.org/wiki/Organisation_du_traité_de_l'Atlantique_Nord	20899
fr.wikipedia.org/wiki/Éric_Woerth	14718
fr.wikipedia.org/wiki/Parti_communiste_français	14505
fr.wikipedia.org/wiki/Commission_européenne	13261
fr.wikipedia.org/wiki/Brice_Hortefeux	11360
fr.wikipedia.org/wiki/Groupe_des_20	9566
fr.wikipedia.org/wiki/Électricité_de_France	9372
fr.wikipedia.org/wiki/Mahmoud_Ahmadinejad	8959
fr.wikipedia.org/wiki/Christine_Lagarde	8486
fr.wikipedia.org/wiki/Angela_Merkel	8124
fr.wikipedia.org/wiki/Hamid_Karzai	7791
fr.wikipedia.org/wiki/Silvio_Berlusconi	7687
fr.wikipedia.org/wiki/Éric_Besson	7510
fr.wikipedia.org/wiki/Liliane_Bettencourt	7311
fr.wikipedia.org/wiki/Airbus	7108
fr.wikipedia.org/wiki/Luc_Chatel	6964
fr.wikipedia.org/wiki/Benoît_XVI	6783
fr.wikipedia.org/wiki/Vladimir_Poutine	6633
fr.wikipedia.org/wiki/Ségolène_Royal	6453
fr.wikipedia.org/wiki/Hillary_Rodham_Clinton	6144
fr.wikipedia.org/wiki/Régie_autonome_des_transports_parisiens	6138
fr.wikipedia.org/wiki/Valérie_Pécresse	6093
fr.wikipedia.org/wiki/Google	6076
fr.wikipedia.org/wiki/Mahmoud_Abbas	6062
fr.wikipedia.org/wiki/Euskadi_ta_Askatasuna	5882
fr.wikipedia.org/wiki/Parlement_européen	5777
fr.wikipedia.org/wiki/Banque_centrale_européenne	5606
fr.wikipedia.org/wiki/Jacques_Chirac	5585
fr.wikipedia.org/wiki/Gordon_Brown	5309
fr.wikipedia.org/wiki/Fonds_monétaire_international	5248
fr.wikipedia.org/wiki/François_Bayrou	5222
fr.wikipedia.org/wiki/Jean-Paul_Huchon	5143
fr.wikipedia.org/wiki/Christian_Estrosi	5102
fr.wikipedia.org/wiki/Areva	5074
fr.wikipedia.org/wiki/Laurent_Gbagbo	4996
fr.wikipedia.org/wiki/Hezbollah	4981
fr.wikipedia.org/wiki/Conseil_de_sécurité_des_Nations_unies	4838
fr.wikipedia.org/wiki/Sénat	4827
fr.wikipedia.org/wiki/Xavier_Bertrand	4609
fr.wikipedia.org/wiki/Bertrand_Delanoë	4522
fr.wikipedia.org/wiki/Organisation_de_coopération_et_de_développement_économiques	4491
fr.wikipedia.org/wiki/Bernard_Accoyer	4443
fr.wikipedia.org/wiki/Organisation_mondiale_du_commerce	4305
fr.wikipedia.org/wiki/Aung_San_Suu_Kyi	3910
fr.wikipedia.org/wiki/Nouveau_Centre	3906

Table 7.2 : Référentiel AMO : entités instanciées les plus fréquentes (classes *Person* et *Organization*) dans le corpus d'acquisition.

à de nouvelles instances, par opposition à celles faisant déjà l'objet d'une modélisation. Les entités candidates sont ensuite sélectionnées par un processus de validation identique à la population initiale, reposant notamment sur le critère de pertinence.

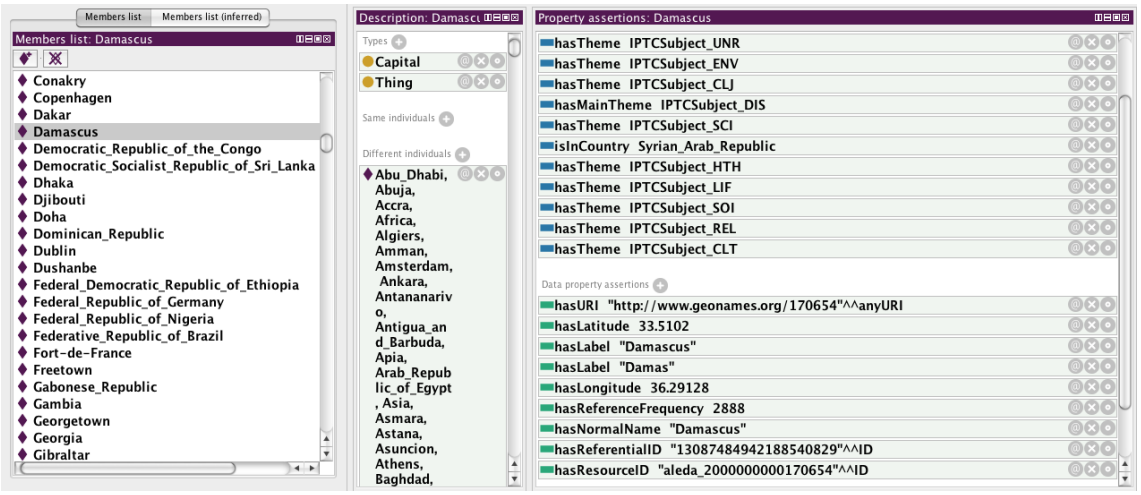


FIGURE 7.6 : Référentiel AMO : exemple d'instance (classe *Location*) issue de la résolution d'entités après collecte sur corpus et filtrage.

- Au fil de la production : dans la perspective d'une intégration du processus d'enrichissement à celui de la production quotidienne de contenus, il est envisagé que les résultats de l'Annotation Sémantique menée sur chaque document soient communiqués de façon continue à un système d'enrichissement du référentiel, administré de façon semi-automatique. Les entités ainsi proposées à un rythme quotidien ou hebdomadaire peuvent alors faire l'objet d'une validation et d'une instanciation individuelles dans le référentiel.

Un enrichissement statique à partir d'archives récentes, postérieures au corpus utilisé pour la population initiale, a été mené par le Medialab, selon une procédure de validation par les journalistes identique à la première. L'extraction d'entités a cette fois concerné uniquement les personnes, mentionnées dans le fil général en français de l'AFP de janvier à mai 2012. À partir des 38 000 personnes ainsi identifiées, seules les entités ne faisant pas déjà l'objet d'une instance dans AMO et apparaissant au moins 25 fois dans le corpus d'acquisition ont été présentées à la validation manuelle. L'exclusion des entités déjà instanciées dans AMO est triviale : toutes les entités extraites sont identifiées relativement à la base Aleda, dont les instances du référentiel indiquent l'identifiant ; il suffit donc de ne conserver que les entités munies d'un identifiant absent du référentiel. Après validation, 435 nouvelles entités ont été sélectionnées et instanciées dans la classe *Person* d'AMO. Les plus fréquentes sont listées à la table 7.3.

Ce processus d'enrichissement sur archives récentes peut être répété de façon régulière afin notamment de maintenir à jour le périmètre des entités représentées. Les nouvelles entités à examiner devraient *a priori* se présenter en nombre d'autant plus réduit à chaque passe dès lors que le processus est répété à intervalles rapprochés, la majeure partie des entités mentionnées et pertinentes devant déjà avoir été instanciées dans le référentiel.

L'enrichissement du référentiel au fil de la production est quant à lui envisagé dans de futurs travaux au sein du Medialab, notamment à l'issue du projet de recherche EDyLex[10] auquel participe l'AFP. Le développement principal à réaliser dans cet objectif consiste en une synchronisation des outils d'enrichissement de la production, utilisés à partir des consoles de rédaction, avec le

10. Projet ANR financé par le programme STIC CONTINT 2009 (projet ANR-09-CORD-008). Novembre 2009 - Juin 2013.

URL Wikipedia	fréquence
fr.wikipedia.org/wiki/Philippe_Poutou	1342
fr.wikipedia.org/wiki/Ali_Abdallah_Saleh	1143
fr.wikipedia.org/wiki/Anders_Behring_Breivik	1058
fr.wikipedia.org/wiki/Olivier_Besancenot	1030
fr.wikipedia.org/wiki/Macky_Sall	1009
fr.wikipedia.org/wiki/Marie-George_Buffet	940
fr.wikipedia.org/wiki/Philippe_de_Villiers	890
fr.wikipedia.org/wiki/Martine_Aubry	808
fr.wikipedia.org/wiki/Dominique_de_Villepin	705
fr.wikipedia.org/wiki/Jean-Louis_Borloo	645
fr.wikipedia.org/wiki/Alassane_Ouattara	622
fr.wikipedia.org/wiki/Dmitri_Medvedev	576
fr.wikipedia.org/wiki/Zine_el-Abidine_Ben_Ali	572
fr.wikipedia.org/wiki/Patrice_de_Maistre	439
fr.wikipedia.org/wiki/Abdallah_Senoussi	376
fr.wikipedia.org/wiki/Anne_Lauvergeon	360
fr.wikipedia.org/wiki/Dioncounda_Traoré	354
fr.wikipedia.org/wiki/Burhan_Ghalioun	342
fr.wikipedia.org/wiki/Kim_Jong-eun	331
fr.wikipedia.org/wiki/Carlos_Gomes_Júnior	284
fr.wikipedia.org/wiki/Nafissatou_Niang_Diallo	282
fr.wikipedia.org/wiki/Saddam_Hussein	279
fr.wikipedia.org/wiki/Ziad_Takieddine	271
fr.wikipedia.org/wiki/Ahmed_Ben_Bella	270
fr.wikipedia.org/wiki/Ahmet_Davutoğlu	250
fr.wikipedia.org/wiki/Valérie_Trierweiler	250
fr.wikipedia.org/wiki/Christine_Boutin	237
fr.wikipedia.org/wiki/Rick_Perry	233

Table 7.3 : Référentiel AMO : entités instanciées les plus fréquentes (classe *Person*) dans le corpus d'acquisition (seconde population).

référentiel et son administration. Au-delà de l'ajout d'entités identifiées à partir de la base Aleda, il s'agit également d'identifier de nouvelles entités : le processus d'identification d'entités permet en effet de détecter les mentions référant à des entités absentes des ressources employées (cas NIL). L'enrichissement peut alors prévoir des traitements visant à identifier ces entités par des moyens semi-automatiques, impliquant la recherche et la consultation de ressources externes. Les nouvelles entités résultant de ce processus de *découverte* d'entités émergentes peuvent alors être intégrées à la base Aleda d'une part, et proposées comme instances candidates du référentiel AMO d'autre part.

2 Enrichissement de dépêches et Recherche d'Information

2.1 Enrichissement de dépêches

L'enrichissement de contenus à l'aide de métadonnées constitue l'application première de l'identification d'entités dans le cadre du présent travail. Le processus d'identification d'entités réalisé sur les contenus textuels, ici les dépêches de l'AFP, à l'aide d'un système tel que Nomos ou NP-Normalizer, accomplit une Annotation Sémantique qui aboutit à la production de métadonnées à partir des mentions d'entités au sein des documents. L'ancrage des métadonnées dans la ressource d'entités associée à ces systèmes, Aleda, est le point central du caractère sémantique de l'enrichissement des contenus.

L'enrichissement par Annotation Sémantique peut être réalisé par lots, sur des corpus de

dépêches archivées, ou au fil de la production, parallèlement au processus rédactionnel. Dans le second cas, l'Annotation Sémantique est déclenchée par les rédacteurs au sein de la console, qui joue le rôle de *content management system* (CMS), et peut intégrer une phase de validation manuelle avant que les annotations ne soient associées aux documents transmis sur les fils (production en mode *push*). Dans le premier, les processus d'enrichissement peuvent être initié par un département de l'AFP tel que celui en charge de la documentation, dont le CMS peut intégrer des fonctionnalités plus étendues : en plus de l'Annotation Sémantique des entités mentionnées dans les contenus, des traitements d'Extraction d'Information plus complexes tels que l'extraction de citations, qui sera abordée à la section suivante (3), sont envisagés. Les archives de dépêches ainsi traitées peuvent ensuite être mises à disposition sur les serveurs dédiés à la production en mode *pull*, par lequel les clients de l'AFP adressent des requêtes afin d'obtenir des contenus selon des paramètres particuliers. Le mode *pull* peut également accéder à la production enrichie lors des processus réalisés pour la livraison directe en mode *push*, dont les résultats sont eux aussi archivés. La figure 7.7 donne une vue générale de l'intégration potentielle des processus et outils d'enrichissement dans la chaîne de production de l'AFP. L'Annotation Sémantique des entités est associée, dans ce schéma, à un annotateur s'appuyant sur le système Nomos. Les autres outils d'annotation accomplissant d'autres traitements relatifs à l'Extraction d'Information y sont également représentés. Ces deux types de système sont mis en relation avec le référentiel AMO présenté ci-avant. Les fonctionnalités d'administration du référentiel y sont illustrées par le CMS 3, qui peut faire appel aux différents systèmes d'annotation par exemple pour mener des itérations d'enrichissement d'AMO. Cet enrichissement est également prévu, comme cela a été exposé précédemment, par le biais de candidats détectés lors de l'enrichissement de la production pour la livraison et présentés à l'administration du référentiel.

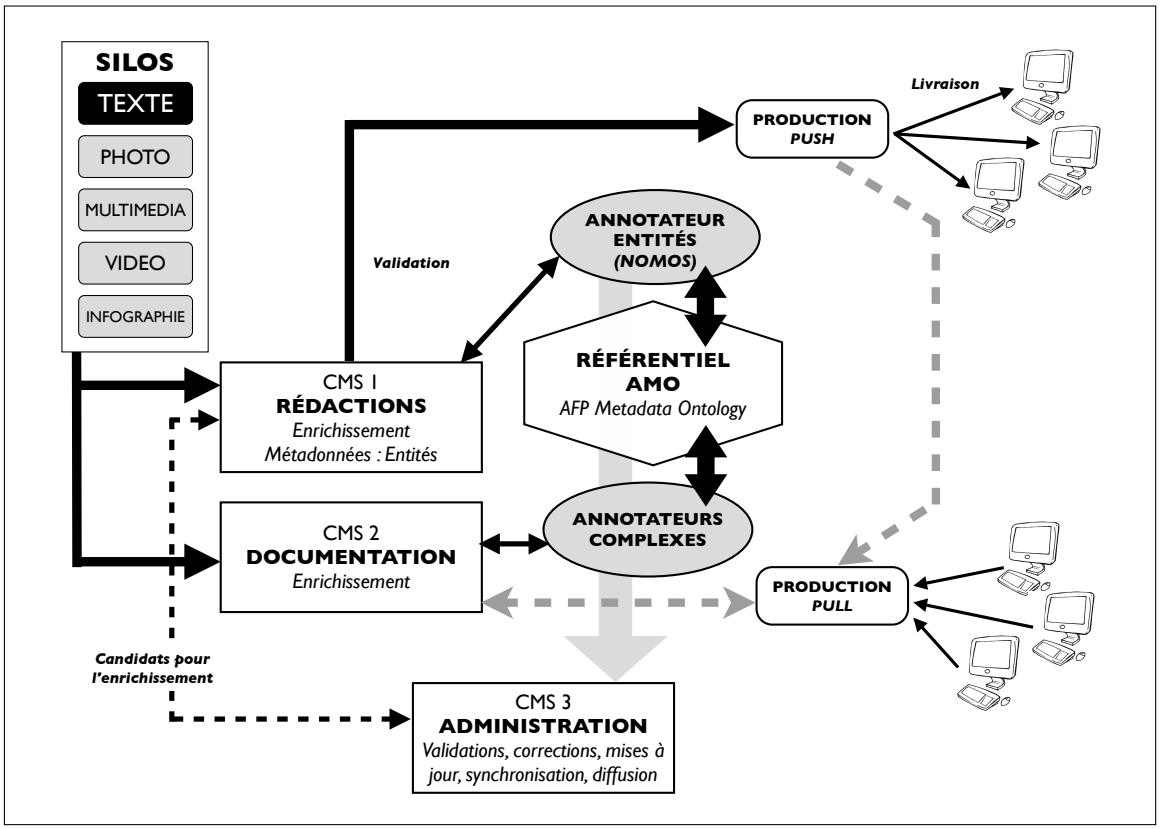


FIGURE 7.7 : Intégration de l'enrichissement à la chaîne de production : vue générale.

L'enrichissement des contenus vise à rendre possible un certain nombre d'applications utilisatrices de ces contenus reposant sur les métadonnées associées ; cet objectif s'inscrit dans la perspective générale du Web Sémantique, étudiée au début du présent travail. Il s'agit notamment d'orienter la RI, étroitement liée au Web mais également cruciale dans l'activité d'une organisation telle que l'AFP, vers un paradigme *sémantique*, dans lequel les métadonnées jouent le rôle de descripteurs formels et qui sera abordé ci-après (section 2.2) dans le cadre des travaux menés par le Medialab. L'enrichissement se présente cependant également comme un objectif pour lui-même : les contenus ainsi diffusés et mis à disposition par l'AFP, à destination de ses clients mais également en interne, bénéficient d'une valorisation notable relativement aux contenus bruts. Cette valorisation se traduit par différents facteurs, en particulier :

- Une consultation et une manipulation plus riches du point de vue de l'utilisateur : les documents sont munis de liens hypertexte ancrés sur les métadonnées, pointant vers diverses ressources et donnant ainsi accès à un espace informatif plus large que celui des contenus eux-mêmes. Dans le cas présent, ces liens peuvent mener aux articles encyclopédiques de Wikipedia, à la base géographique de GeoNames, aux Linked Data de DBpedia ou aux données du NYT, et chacune de ces ressources présente elle-même d'autres liens vers d'autres ensembles d'information. Ce mode de livraison des contenus favorise la cohérence et l'exhaustivité de l'information, mais ouvre également un espace d'exploration où l'utilisateur est actif vis-à-vis des contenus.

- La potentialité de nouveaux usages des contenus en aval à leur production : l'adoption des standards du Web Sémantique dans les processus d'enrichissement permet la diffusion de contenus sous une forme augmentée sans qu'il soit nécessaire d'établir au préalable l'usage qui pourra être fait des métadonnées par les utilisateurs. Il s'agit ainsi de processus dont la visée est applicative dans un cadre générique et non figé. Des applications clients pourront notamment mettre en correspondance les métadonnées mises à disposition par l'AFP avec leurs propres ressources référentielles, dans des perspectives d'indexation ou de filtrage de la production de l'Agence. En interne, les contenus enrichis lors de la production peuvent être réutilisés ultérieurement, une fois archivés, par des applications développées notamment par le Medialab et s'appuyant sur les métadonnées produites.

Formats L'intégration des processus d'enrichissement à la production suppose l'adoption de formats de représentation pour les métadonnées ajoutées aux contenus. Il s'agit dans le cadre de l'AFP d'intégrer ces métadonnées au format NewsML (cf. chapitre 4, section 2.1), reposant sur le langage XML. Les outils d'Annotation Sémantique, en particulier Nomos, produisent des annotations au format XML (cf. chapitre 6, section 2.3) ; le format NewsML y est par ailleurs pris en charge, une dépêche dans ce format pouvant être donnée en entrée au système et retournée en sortie avec sa structure d'origine. Les annotations réalisées y sont insérées sous forme de balises XML dont les noms et attributs sont définis par Nomos ou tout autre système intervenant dans les traitements. Ces balises peuvent être directement converties dans d'autres format pour l'affichage, la diffusion et l'archivage, qu'il s'agisse de RDFa ou rNews, par exemple (cf. chapitre 4, section 3.3). L'annexe A présente une dépêche enrichie en métadonnées, aux formats HTML (métadonnées en RDFa) et NewsML (métadonnées en XML) (figures A.18 à A.24). Les informations ajoutées aux balises RDFa sont obtenues à partir des métadonnées en XML et d'un accès au référentiel AMO qui centralise les informations et liens relatifs aux entités ainsi représentées. Dans le cas d'une entité absente du référentiel, les informations peuvent être obtenues à partir de la base Aleda.

Qualité de l'enrichissement Lors de la consultation de dépêches enrichies, les performances du module d'identification d'entités sont immédiatement visibles : les segments textuels anno-

tés ne correspondant pas à des mentions d'entités et les mentions dont les frontières détectées sont erronées introduisent un premier niveau de bruit ; les liens établis sur de telles mentions forment un second niveau d'erreurs d'autant plus notables qu'ils sont sans objet dans les cas de faux positifs ; enfin, les liens établis sur des mentions correctes mais dont la référence est fausse constituent un troisième niveau nuisant à la qualité générale de l'enrichissement. Parallèlement à ces erreurs relevant de la précision de l'identification, les annotations manquées peuvent également contribuer à dégrader la qualité de l'enrichissement, en particulier lorsque des entités jugées comme importantes sont absentes des résultats. La notoriété d'une entité est en effet perçue par les utilisateurs comme un facteur devant permettre à un système de la reconnaître sans aucune difficulté ; la méthodologie de l'identification intègre d'ailleurs l'idée d'un « sens par défaut » des mentions, dont la sélection consiste à accorder un poids d'autant plus grand à une entité que sa notoriété est grande, en fonction d'indications disponibles dans les ressources utilisées (cf. chapitre 6, section 2.2.3).

L'estimation de la qualité d'un système d'identification automatique d'entités tel que Nomos peut donc être réalisée non seulement par une évaluation reposant sur les métriques usuellement adoptées dans les tâches de TAL et d'Extraction d'Information, comme cela a été fait au chapitre précédent, mais également relativement à un contexte d'utilisation concret comme celui de la RI, abordé dans la section suivante. Les erreurs retournées par un système peuvent en effet se présenter en nombre réduit tout en dégradant fortement la qualité des résultats telle qu'elle est perçue lors de l'utilisation, en particulier lorsqu'il s'agit d'erreurs difficilement justifiables aux yeux des non-spécialistes et récurrentes.

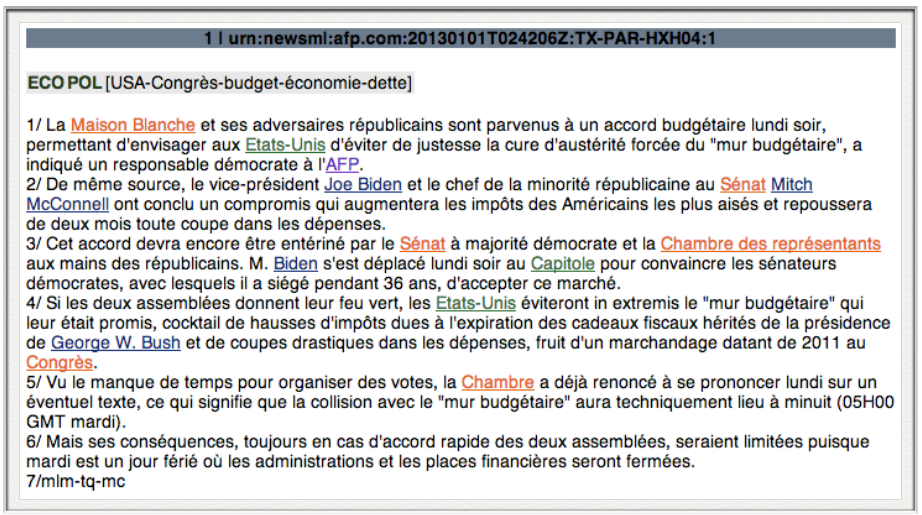
2.2 Recherche d'Information orientée entités

L'intégration des métadonnées de documents à la tâche de RI constitue l'une des applications les plus directes de l'enrichissement de contenus, qui vise à dépasser le seul niveau de l'accès par mots-clés à l'information. Une orientation de la RI vers les entités est ainsi envisagée dans le contexte de travail de l'AFP, où l'indexation des contenus sur les métadonnées produites par l'Annotation Sémantique permet non seulement de paramétrer la diffusion des contenus à destination des clients, mais également de créer un espace de recherche et d'exploration de la production pour des usages internes, tels que la documentation et la consultation d'archives. Ces besoins fonctionnels trouvent dans l'indexation par entités une approche renouvelée, dans laquelle les contenus sont d'une part caractérisés de façon explicite et précise et d'autre part accessibles *via* plusieurs points d'entrée. La RI orientée entités proposée ici introduit en effet un processus d'indexation des contenus sur les entités indiquées par les métadonnées de documents, en plus de l'indexation classique : plein-texte, slugs, sujets IPTC, date, pays, fil.

L'intégration des entités dans la formulation des requêtes peut donner lieu à un champ de recherche dédié, par lequel une entité identifiée de façon unique et explicite est spécifiée comme descripteur recherché. On peut ainsi formuler une requête en vue d'obtenir des dépêches au sujet du président américain actuel, Barack Obama ou du chanteur américain Michael Jackson, sans que ces descripteurs puissent être confondus avec l'épouse du premier, Michelle Obama, ou l'écrivain britannique homonyme du second. Une telle requête peut également être enrichie par la spécification de champs multiples et de sémantiques différentes ; la dépêche présentée à la figure 7.8 peut ainsi être incluse dans les résultats d'une requête portant sur le vice-président américain Joe Biden, le chef de la minorité au Sénat Mitch McConnell, dans des documents datés de janvier 2013, comportant le slug *budget* et catégorisés par le sujet IPTC ECO (économie).

Alternativement, un champ similaire peut ne spécifier qu'un nom ou label d'entité, tel que *Obama* ; dans ce cas, les résultats de recherche sont regroupés en fonction des différents descripteurs ou entités pouvant correspondre à ce label — dépêches concernant le président américain

d'un côté, son épouse de l'autre [11]. Cette distinction portant sur le niveau d'intégration de la spécification sémantique — requête et résultats ou résultats seulement — se retrouve dans la typologie des différentes approches de la RI sémantique telle que la présente notamment Mangold [Man07]. L'indexation par entités, dont un sous-ensemble correspond aux instances du référentiel ontologique de l'AFP, AMO, permet par ailleurs d'intégrer les différents champs informatifs relatifs à ces entités dans le système de requêtes, formant ainsi un ensemble de *facettes* de recherche. Des champs ou facettes de recherche peuvent ainsi indiquer la classe (*Person*, *Organization* ou *Location*) des entités spécifiées comme descripteurs recherchés ; dans la perspective d'une population plus étendue et riche du référentiel, des connaissances telles que les fonctions occupées par des personnes au sein d'organisations ou les dates d'événements pourront également fair l'objet de champs de recherche. Les différentes facettes spécifiées permettent ainsi de contraindre les résultats en fonction de critères précis et dont la sémantique est spécifiée et accessible notamment *via* le référentiel AMO.



1 | urn:newsml:afp.com:20130101T024206Z:TX-PAR-HXH04:1

ECO POL [USA-Congrès-budget-économie-dette]

1/ La Maison Blanche et ses adversaires républicains sont parvenus à un accord budgétaire lundi soir, permettant d'envisager aux Etats-Unis d'éviter de justesse la cure d'austérité forcée du "mur budgétaire", a indiqué un responsable démocrate à l'AFP.
2/ De même source, le vice-président Joe Biden et le chef de la minorité républicaine au Sénat Mitch McConnell ont conclu un compromis qui augmentera les impôts des Américains les plus aisés et repoussera de deux mois toute coupe dans les dépenses.
3/ Cet accord devra encore être entériné par le Sénat à majorité démocrate et la Chambre des représentants aux mains des républicains. M. Biden s'est déplacé lundi soir au Capitole pour convaincre les sénateurs démocrates, avec lesquels il a siégé pendant 36 ans, d'accepter ce marché.
4/ Si les deux assemblées donnent leur feu vert, les Etats-Unis éviteront in extremis le "mur budgétaire" qui leur était promis, cocktail de hausses d'impôts dues à l'expiration des cadeaux fiscaux hérités de la présidence de George W. Bush et de coupes drastiques dans les dépenses, fruit d'un marchandage datant de 2011 au Congrès.
5/ Vu le manque de temps pour organiser des votes, la Chambre a déjà renoncé à se prononcer lundi sur un éventuel texte, ce qui signifie que la collision avec le "mur budgétaire" aura techniquement lieu à minuit (05H00 GMT mardi).
6/ Mais ses conséquences, toujours en cas d'accord rapide des deux assemblées, seraient limitées puisque mardi est un jour férié où les administrations et les places financières seront fermées.
7/mlm-tq-mc

Entités-descripteurs

- **AFP** [Aleda #1000000000055197]
- **Joe Biden** [Aleda #1000000000255711]
- **George W. Bush** [Aleda #1000000000680078]
- **Chambre des Représentants des États-Unis** [Aleda #1000000000089459]
- **Congrès des États-Unis** [Aleda #1000000000088935]
- **Maison Blanche** [Aleda #1000000002872401]
- **Mitch McConnel** [Aleda #1000000000263731]
- **Sénat des États-Unis** [Aleda #1000000000089073]
- **United States** [Aleda #2000000006252001]
- **United States Capitol** [Aleda #2000000004140827]

Figure 7.8 : Dépêche et entités-descripteurs pour l'indexation et la RI.

Évaluation de l'identification dans la RI Comme cela a été évoqué précédemment, l'évaluation d'un système d'identification tel que Nomos peut être étendue au-delà des scores mesurés dans le cadre d'expériences isolées des contextes de son utilisation. La RI sémantique constitue à ce titre une application pour laquelle la qualité de l'identification joue un rôle décisif : l'identification des entités est en effet le moyen d'obtention des descripteurs permettant d'indexer puis de retourner les documents en fonction des requêtes formulées. Sa qualité n'est alors plus seulement évaluée au

11. Les résultats d'une telle requête peuvent par ailleurs présenter une intersection de résultats si les deux entités-descripteurs apparaissent conjointement dans les mêmes documents

niveau local des annotations insérées aux contenus textuels, mais également au niveau global des documents eux-mêmes. Une entité constitue en effet un descripteur de document dès lors qu'elle y est mentionnée[12] et ce quelle que soit les variantes lexicales de ses mentions. La figure 7.8 illustre cette relation entre document et entité-descripteur à l'aide d'une dépêche issue du corpus d'évaluation.

L'évaluation proposée ici a été menée sur un corpus de 70 dépêches du fill généraliste de l'AFP en français, datées de janvier 2013. Les données de référence ont été constituées de façon semi-automatique, de façon similaire à la procédure d'annotation du corpus de référence pour l'évaluation générale du système Nomos (cf. chapitre 5) : les dépêches ont d'abord été annotées automatiquement à l'aide du système npNormalizer, puis les résultats d'identification ont été revus manuellement par les journalistes. Cette revue manuelle n'a cependant pas donné lieu, comme pour le corpus GAFP, à une correction systématique des annotations de mentions au sein des contenus textuels. Il s'agissait en effet pour cette tâche de RI de déterminer, pour chaque dépêche, l'ensemble des entités constituant des descripteurs. Les corrections ont donc consisté à examiner la liste d'entités-descripteurs retournée par npNormalizer, puis à en éliminer les descripteurs faux — entités non mentionnées dans la dépêche — et à y ajouter les descripteurs manquants. On peut observer que cette procédure de correction manuelle permet non seulement de disposer de données de référence pour l'évaluation de Nomos, mais également d'évaluer npNormalizer et ainsi de comparer les performances des deux systèmes sur cette tâche.

La table 7.4 indique les caractéristiques du corpus d'évaluation. Les tables 7.5 et 7.6 présentent les résultats des deux systèmes, npNormalizer et Nomos, pour cette tâche de RI. Plus précisément, l'évaluation mesure d'une part la qualité de l'indexation des dépêches en termes d'entités, et d'autre part les performances de RI où chaque entité est considérée en tant que requête. Dans le premier cas, les entités non alignées sur une entrée de la base Aleda (cas NIL) sont incluses dans le compte des descripteurs correctement associés aux dépêches. Elles sont en revanche ignorées dans le second cas, où on ne considère comme requêtes que les entités identifiées. Les résultats sont donnés selon deux métriques : la version *micro* accorde autant de poids à chaque document dans l'évaluation de l'indexation et à chaque entité en tant que descripteur de document dans l'évaluation de la qualité des requêtes ; on a donc des valeurs de précision et de rappel pour chaque document, qui sont moyennées pour l'obtention des valeurs globales. La version *macro* calcule les valeurs globales pour chaque couple dépêche / entité, chacun d'eux ayant donc le même poids dans cette évaluation.

Dépêches	70
Entités uniques identifiées	465
Entités uniques NIL (nom unique)	71
Nombre moyen d'entités uniques par dépêche	18
Nombre moyen d'entités uniques par dépêche hors NIL	16

Table 7.4 : Caractéristiques du corpus d'évaluation pour la tâche de RI.

Cette évaluation montre des résultats de bonne qualité pour npNormalizer, mais décevants et largement inférieurs pour Nomos, dont l'évaluation présentée au chapitre précédent avait pourtant donné des scores légèrement supérieurs à npNormalizer. On peut à cet égard souligner que :

- le développement de Nomos sur le corpus GAFP pâtit sans doute de données d'apprentissage et d'évaluation insuffisantes, ainsi qu'un biais lié à ces données ne permettant pas de

12. Il peut cependant être envisagé de ne considérer une entité, ou tout autre type de descripteur dérivé du contenu documentaire, qu'à partir d'un seuil minimal d'occurrences ou d'un autre critère (poids tfidf ou test statistique tel que le test *t*).

	npNormalizer		Nomos LN		Nomos LNSAI	
	micro	**macro**	**micro**	**macro**	**micro**	**macro**
Précision	77,18	76,40	56,12	55,31	70,61	69,39
Rappel	78,80	79,25	61,69	60,22	69,89	68,73
F_1	**77,98**	**77,80**	**58,77**	**57,66**	**70,25**	**69,06**

Table 7.5 : Résultats d'évaluation des systèmes npNormalizer et Nomos sur la tâche de RI orientée entités : indexation.

	npNormalizer		Nomos LN		Nomos LNSAI	
	micro	**macro**	**micro**	**macro**	**micro**	**macro**
Précision	78,87	78,06	76,85	71,61	75,73	72,12
Rappel	90,75	88,57	75,92	69,57	83,96	78,70
F_1	**84,39**	**82,98**	**76,38**	**70,58**	**82,23**	**75,27**

Table 7.6 : Résultats d'évaluation des systèmes npNormalizer et Nomos sur la tâche de RI orientée entités : requêtes.

reproduire les performances constatées sur de nouvelles données ; il sera donc nécessaire de mener des expériences et un développement plus approfondis de Nomos à l'aide de nouvelles données afin de relever ces performances et d'atteindre une bonne généralisation des modèles obtenus ;

- le module npNormalizer associé à la chaîne SxPipe a bénéficié d'un temps et d'une concentration de développement supérieure à Nomos au cours de notre travail au Medialab de l'AFP : son intégration au processus d'enrichissement dès les premiers mois des travaux menés lui a conféré une place indispensable dans la collecte de métadonnées et le développement des prototypes initiaux d'exploitation ; il a ainsi été nécessaire de travailler à son amélioration en parallèle au développement du système Nomos. Ces améliorations ont essentiellement été menées par le signalement et la correction systématiques d'erreurs, par intervention manuelle dans les règles de Liage et constitution de listes (*white lists* pour forcer certaines reconnaissances, *black lists* pour en empêcher d'autres). Ce constat sur les modalités de notre contribution au besoin d'enrichissement des contenus de l'AFP nous mène à espérer que de futurs travaux consacrés de façon plus spécifique à Nomos permettront de renverser cette tendance.

3 Détection automatique de citations et attribution d'auteurs

La détection de citations se présente comme une application concrète de l'enrichissement de la production de l'AFP, dont les objectifs sont d'ordre plus spécifiques que les contextes d'usage présentés ci-avant. Il s'agit d'une tâche d'Extraction d'Information dont les cibles ne sont pas uniquement des éléments atomiques tels que les entités nommées, mais également des formes et structures informatives complexes. Les citations constituent en effet une part importante de l'activité de l'AFP qu'il est intéressant d'identifier dans la perspective de traitements et d'usages relevant de la Recherche d'Information et plus généralement de la valorisation des contenus dans leur exploitation interne et externe. Elles ont donné lieu, au cours de la thèse présentée ici, à l'élaboration d'un système de détection automatique. L'identification d'entités y joue un rôle crucial, dans la mesure où les citations à isoler au sein des contenus revêtent un caractère informatif complet lorsqu'elles sont associées au locuteur, souvent une personnalité liée à l'actualité, dont les

propos sont rapportés. Le système proposé introduit une prise en charge partielle mais efficace de la coréférence afin d'étendre les résultats de détection aux citations dont les auteurs sont mentionnés par le biais de formes anaphoriques.

La place centrale des citations dans la production de l'AFP ainsi que les modalités de leur définition dans le présent contexte sont d'abord discutées afin de déterminer le périmètre de l'application réalisée, tant au niveau des besoins dits *métiers* à son origine que sur le plan de l'analyse linguistique et de l'automatisation (section 3.1). La chaîne de traitement développée est ensuite présentée (section 3.2), en particulier les composants spécifiques à la prise en charge des différentes configurations citationnelles et à l'attribution d'auteurs, y compris anaphoriques. La section 3.3 rapporte des résultats quantitatifs sur un cas applicatif, avant d'illustrer l'intégration de la chaîne dans un moteur d'indexation et de recherche ainsi que dans des outils de visualisation et de navigation.

3.1 Périmètre de l'application

3.1.1 Usages des citations et objectifs

La production de l'AFP est caractérisée par une grande quantité de citations : le corpus GAFP, présenté au chapitre 5 (section 2.1), présente 136 citations pour 96 dépêches, soit près de 1,5 par dépêche. Les citations constituent en effet un élément essentiel du travail de l'Agence, qui se situe en amont de la chaîne d'information. Ce rôle repose de façon cruciale sur la capacité des agenciers à garantir la fiabilité et l'exactitude des faits rapportés, fonction assurée par la désignation explicite de *sources* d'information et, dans de nombreux cas, par la reproduction de propos, qu'ils soient privés et tenus par ces sources, ou publics et relevant de personnalités liées à l'actualité, comme le souligne notamment Tétu [Tét02]. L'appui fondamental de l'information sur des paroles prononcées explique que les citations soient désignées dans le langage métier de l'AFP par le terme *verbatims*. La citation dans le cadre de la production textuelle de l'AFP est en effet indissociable de la présence de guillemets, indiquant de façon systématique les segments relevant d'un locuteur autre que le journaliste et venant ainsi fonder l'événement rapporté. À cette relation précise de propos tenus se joint la désignation explicite de l'entité à l'origine de ces propos, sans laquelle la fiabilité et l'exactitude ne seraient pas assurées.

Les citations occupent une place centrale dans le mode de diffusion de l'information par l'AFP non seulement comme éléments structurant et garants de sa fiabilité, mais également en tant qu'informations elles-mêmes : dans divers domaines de l'actualité, dont la vie politique est le prototype, les déclarations de tous bords et sur l'ensemble des sujets traités constituent à eux seuls des informations susceptibles d'être relatés, puis éventuellement analysés, au niveau journalistique. Ce phénomène relève tant de déclarations d'opinions, faites par diverses personnalités sur l'actualité et la vie politique :

> *« Dans sa campagne démagogique, Nicolas Sarkozy est prêt à raconter n'importe quoi, surtout le contraire de ce qu'il disait hier », dénonce* ***l'eurodéputé*** *sur son blog.*

que de discours performatifs par lesquels ces personnalités s'engagent, promettent ou déclenchent des événements :

> *« La Chine et les Etats-Unis doivent ensemble prendre des mesures pour parvenir à l'objectif de la dénucléarisation de la péninsule coréenne », a acquiescé* ***M. Kerry****.*

L'importance des citations dans la production de l'AFP se traduit par un besoin d'accès immédiat et systématique aux citations, notamment dans le cadre de la recherche documentaire essentielle au travail journalistique. Il est en effet utile voire nécessaire, pour l'analyse de nouveaux événements, de connaître les éléments antérieurs à leur développement, parmi lesquels les

différentes déclarations des personnalités en jeu figurent parmi les plus pertinents. Dans cette perspective, le TAL et en particulier l'Extraction d'Information donnent les moyens de réaliser une collecte des citations au travers des fils et corpus de dépêches publiées par l'AFP, puis de les rendre accessibles sous une forme structurée et exploitable automatiquement. L'identification automatique d'entités joue dans ce cadre un rôle crucial, notamment au niveau de la structuration : l'interprétation et l'exploitation des citations en tant qu'éléments informatifs sont en effet indissociables de leurs auteurs, qui doivent être identifiés de façon précise et univoque comme, de façon générale, dans toute application où les entités sont ciblées en tant qu'individus extra-linguistiques. La détection de citation repose, dans le système proposé, sur une analyse linguistique surfacique intégrant les principaux éléments du phénomène de discours rapporté (DR), adaptée à la définition des citations couverte par le terme *verbatim* propre à l'Agence. L'identification des auteurs est quant à elle permise par le déploiement des systèmes présentés dans les chapitres précédents, npNormalizer et Nomos.

Le développement d'un système de détection automatique de citations pour les contenus de l'AFP repose donc sur les spécifications suivantes :

- Reconnaissance des citations telles que définies par le métier de l'Agence : propos rapportés, placés entre guillemets et attribués à un auteur. Les éléments identifiables au niveau linguistique comme relevant du DR mais sans guillemets ne sont pas considérés dans l'application.

- Attribution d'auteur (i) : l'application restreint les citations, selon des critères liés au caractère informatif des éléments détectés, à celles dont l'auteur est une personne identifiable ; elle exclut ainsi les auteurs non individuels, tels que les organisations ou les medias :

  – *Le 19 novembre, l'inspectrice du travail avait rejeté ces licenciements « se fondant notamment sur le défaut de démonstration du motif économique », mais elle avait ensuite dû « étudier des éléments d'information fournis en dernière minute » par la direction de l'usine, rapporte* ***la CGT*** *qui a toujours contesté la justification économique de la fermeture.*
  – *« La liberté de la presse doit avoir des limites », a affirmé dans un éditorial le quotidien* ***Global Times****.*

  anonymes :

  > *« Il n'avait aucune motivation politique », selon* ***la source*** *proche du dossier.*

  et indéfinis :

  > *Si, selon* ***un membre*** *de son comité de soutien jeudi, M. Abdallah se disait confiant et avait « commencé à faire ses bagages », il risque de devoir faire preuve d'encore un peu de patience.*

  L'indexation des citations détectées est ainsi réalisée au niveau de leurs auteurs, qui doivent être des entités de type personne, désignées de façon univoque au sein des ressources utilisées par les systèmes d'identification d'entités (npNormalizer et Nomos).

- Attribution d'auteur (ii) : les auteurs associés aux citations détectées incluent les noms propres, les formes nominales et pronominales personnelles, dans la mesure où les restrictions introduites précédemment sont respectées. Dans les deux derniers cas, il est nécessaire de déterminer quels sont les référents dénotés par ces formes afin d'indexer les citations sur des entités identifiées par un processus de résolution d'anaphores.

Le cas applicatif traité ici est celui de la recherche de citations dans le cadre de la campagne présidentielle française de 2012. Un tel événement de la vie politique est en effet à l'origine de nombreux échanges, déclarations et réactions verbales dans l'espace public, et donc d'une large production de dépêches en faisant état. Il s'agit également d'un sujet attirant une attention particulière, tant du public que des professionnels de l'information, pour lequel une présentation structurée et systématisée des contenus constitue donc un outil pertinent et efficace d'analyse, de comparaison et de recoupements.

3.1.2 Citations journalistiques et analyse linguistique

La détection de citations dans les contenus textuels de l'AFP constitue une tâche particulière d'Extraction d'Information, dont les éléments cibles sont les verbatims rapportés par les agenciers, associés aux personnes identifiées comme leurs auteurs. Une telle tâche semble pouvoir s'appuyer de façon directe sur l'analyse linguistique générale du discours rapporté (DR), étudiée sous plusieurs de ses aspects notamment par Rosier [Ros08] ou Van Raemdonck [VR02]. Les verbatims constituent cependant une forme particulière de citations, où la présence de guillemets est une contrainte nécessaire à l'interprétation citationnelle des segments de texte ciblés. Le DR ne repose pas de façon exclusive sur une telle contrainte : les différentes configurations identifiées comme relevant du DR, qui peuvent impliquer la présence de guillemets, relèvent de structures propositionnelles particulières et de l'emploi de prédicats sémantiquement liés à la parole. On considère ainsi généralement les configurations de DR suivantes :

Discours rapporté direct (DD) Les propos rapportés sont clairement identifiés comme tels, le plus souvent par des guillemets, indiquant que le discours a été prononcé en les termes exacts, et par une frontière propositionnelle marquée (virgule ou deux points). Le placement du prédicat est dans cette configuration :

- à l'initiale (**DD**$_{ini}$) :

  *Interrogée par LCI sur le refus opposé par le député béarnais à la proposition de désistement mutuel formulée par Eva Joly, Mme **Duflot** a répondu : « Je ne suis pas tellement surprise. [...] Il sort de sa boîte au moment de l'élection présidentielle, puis il retourne à l'intérieur. »*

- ou en incise (**DD**$_{inc}$) :

  *« François Bayrou se refuse à participer à un débat collectif », a-t-**elle** accusé.*

Discours rapporté indirect (DI) Les propos rapportés ne le sont pas nécessairement en termes exacts et sont exprimés dans une proposition conjonctive, subordonnée au prédicat verbal de parole :

> ***Le député PS** a déclaré qu'il trouvait ces réformes justes et nécessaires.*

Dans ces différentes configurations, on analyse généralement les propos rapportés comme l'objet du verbe de parole, qui est en effet typiquement transitif (*dire, déclarer, affirmer, avouer...*). On observe également la possibilité d'employer un verbe intransitif dans la construction en incise :

> *« Ces réformes sont justes et nécessaires », a monologué **le premier ministre** devant l'Assemblée.*

que Lamiroy et Charolles [LC08] expliquent par une échelle de transitivité des verbes révélée par l'incise, permettant à certains d'exprimer par *fusion*, selon l'expression de Gross [Gro81], à la fois les propos tenus et la manière dont ils ont été prononcés. Dans l'exemple ci-dessus, le premier

ministre tient son propos *en monologuant*. On ajoute à ces configurations les citations reposant sur un groupe prépositionnel apposé aux propos rapporté, qui s'apparente à l'incise de discours direct :

> *Selon **Martine Aubry**, « on se rend compte que finalement Nicolas Sarkozy ne réunit les partenaires sociaux que lorsqu'il peut les utiliser pour sa propre communication personnelle. »*

Les verbatims caractéristiques des corpus de dépêches dépassent toutefois les frontières de ces configurations en marquant les propos rapportés de façon dite *libre* : les guillemets peuvent n'entourer que certains segments de propositions, voire uniquement certains mots, et plusieurs segments discontinus peuvent ainsi être marqués au sein d'une même phrase ou proposition ; inversement, les guillemets ne respectent pas nécessairement les frontières de phrase, en pouvant couvrir plusieurs phrases, dont certaines en partie seulement :

- *Tous ces obstacles à l'accès à internet relèvent « d'une stratégie » visant à pousser les blogueurs « à se rendre dans les ambassades pour utiliser leur réseau afin de pouvoir les accuser ensuite d'être soutenus par des gouvernements étrangers », estime pour sa part **Ivan Diaz**.*

- *La courbe de l'épidémie au Mexique est toujours « descendante », a souligné **le ministre** dans un communiqué, soulignant que le nombre de décès représentait 1,9% du total de cas confirmés.*

- *Les Français « observent », a poursuivi **M. Longuet** pour qui la campagne électorale est « un choc de personnalités. »*

- *« C'est tout le problème. Les seuls représentants de la jeunesse sont bien souvent des étudiants », laissant de côté les apprentis, élèves de l'enseignement professionnel et jeunes chômeurs, soit 2,5 millions de personnes, note **M. Gille**.*

- *Cette indépendance « était la seule option viable pour la stabilité de la région. La réussite d'un Kosovo indépendant est une priorité de notre administration et de notre pays », a-t-**il** ajouté.*

On observe également que les guillemets peuvent apparaître dans la configuration de DI, mêlant ainsi discours direct et discours indirect, avec un effet d'*hyperréalisme* propre au style journalistique, selon l'analyse proposée par Rosier [Ros02] :

> *Le lancement de la série d'une cinquantaine de concerts de Michael Jackson, prévus à Londres à partir de cet été, a été repoussé de quelques jours, a annoncé mercredi **Randy Phillips**, assurant que cela n'avait « rien à voir » avec la santé de la star.*

Plus généralement, le placement de segments textuels entre guillemets ne se limite pas aux configurations usuelles du DR, et peuvent donner lieu à des configurations *mixtes*, où l'idée d'un rapport de discours est présente, portée par un prédicat impliquant une forme de parole, mais où l'on ne retrouve pas la structure mettant en jeu un prédicat de parole complété par un objet correspondant aux propos tenus :

- ***Elle** a fustigé mardi le « repli dogmatique » du patronat à deux jours du dernier round de la négociation sur la sécurisation de l'emploi, l'appelant à des « efforts » pour aboutir à un « accord novateur ».*

- ***François Hollande** (PS) a préconisé vendredi de « revoir le mode d'affectation » dans l'éducation.*

Cet emploi mixte des guillemets est notamment abordé par des travaux en sémantique formelle, tels que ceux de Maier [Mai07]. Sans entrer ici dans un examen plus précis de la relation entre usage des guillemets et DR, il semble nécessaire de prendre en charge ce phénomène dans le système développé ici afin de rendre compte de son importance quantitative au sein des contenus à traiter. Cette prise en charge repose notamment sur le recensement des verbes pouvant donner lieu, dans leur environnement immédiat, à de telles citations, et à identifier au moins les motifs récurrents qui les caractérisent.

Exploration de corpus L'exploration semi-automatique d'un corpus de 5 000 dépêches nous a permis de repérer un certain nombre de ces motifs, ainsi que d'estimer la répartition des autres configurations de citations (DI, DD_{ini} ou DD_{inc}). Nous nous sommes d'abord appuyés sur une liste de 110 verbes, constituée préalablement dans le cadre du développement de l'analyseur syntaxique FRMG [LC05], et sur le lexique morphosyntaxique du français Le*fff* [Sag10]. Les phrases présentant ces verbes sous leur forme fléchie de troisième personne du singulier au présent et au passé composé, caractéristique de leur usage citationnel et fournie par le Le*fff*, ont été automatiquement extraites du corpus puis filtrées par l'application de motifs surfaciques correspondant aux différentes configurations. Ces motifs relativement simples permettent d'isoler les cas de DI lorsque le verbe est suivi de la conjonction *que*, les cas de DD_{ini} lorsque le verbe est suivi de deux points (« : ») et de guillemets ouvrants, et les cas de DD_{inc} lorsque que le verbe suit des guillemets fermants suivis d'une virgule [13]. En ne retenant que les phrases concernées par ces motifs, on observe que :

- parmi les verbes de cette liste apparaissant dans ce corpus en configuration citationnelle (91 sur 110), une majorité (70 sur 91) apparaît en DD_{inc} ;
- 20 de ces 70 verbes apparaissent uniquement en DD_{inc} ; on remarque que certains d'entre eux (*temporiser*, par exemple) ne sont pas usuellement recensés parmi les verbes de parole ;
- la configuration DD_{inc} est majoritaire : 66% des exemples, contre 21% de DI et 3% de DD_{ini}.

Ces trois observations nous mènent à considérer la configuration DD_{inc} comme particulièrement productive et à explorer de nouveau le corpus afin d'y trouver d'éventuels nouveaux verbes ainsi employés. À partir du motif conçu pour le repérage de DD_{inc} avec un verbe de la liste initiale, nous élargissons l'extraction à toute phrase du corpus présentant des guillemets fermants suivis d'une virgule. Cette méthode relativement grossière et imprécise a permis de repérer avec rapidité et peu de recherche manuelle des exemples de DD_{inc} faisant intervenir des verbes non encore listés, au nombre de 121, tels que *fustiger, ironiser, commenter, motiver* ou *énumérer*. Enfin, cette liste étendue de verbes a pu faire l'objet d'une classification selon leur distribution sur les trois configurations au travers du corpus :

> **Classe 1** verbes transitifs usuellement recensés comme verbes de parole (*dire, répondre*), apparaissant en DI, DD_{ini} et DD_{inc} (50% des verbes) ;
>
> **Classe 2** verbes intransitifs apparaissant uniquement en DD_{inc} (*ricaner, fulminer, ironiser*) (19% de verbes) ;
>
> **Classe 3** verbes transitifs apparaissant uniquement en DD_{inc} (*commenter, fustiger, analyser, tancer*) (31% des verbes) ;

13. Dans le cas de formes verbales au passé composé, les motifs intègrent la possible présence d'un pronom clitique entre l'auxiliaire et le participe (*a-t-il déclaré*). Ces motifs simples ne visent ni une représentation linguistique précise ni l'exhaustivité, mais permettent, dans le processus d'exploration, de repérer efficacement les modalités de distribution générale des éléments en jeu dans la citation au sein des contenus à traiter.

Une analyse linguistique menée à la suite de ces observations, portant sur les caractéristiques syntaxiques et discursives de la configuration DD_{inc}, rapportée dans Danlos et al. [DSS10] et Sagot et al. [SDS10], permet notamment de constater que l'interprétation des propos rapportés comme objet du verbe de citation ne peut s'appliquer dans le cas des verbes de la classe 3. Ceux-ci présentent en effet, en dehors de leur usage en incise citationnelle, un objet direct ne relevant pas de la parole (par exemple *commenter une situation*), contrairement aux verbes de la classe 1 qui, sur le modèle de *dire*, peuvent substituer un prédicat nominal de parole aux propos rapportés. Plus généralement, notre étude propose d'analyser la configuration DD_{inc} comme relevant du niveau sémantico-discursif et non syntaxico-phrastique.

Au-delà des conclusions d'ordre linguistique proposées par notre étude, que nous ne détaillons pas davantage ici, l'exploration de corpus a donc permis d'aboutir à une ressource verbale riche et en adéquation avec les contenus à traiter, disponible pour le développement du système de détection automatique. Cette ressource a par ailleurs été progressivement enrichie au cours des travaux portant sur les dépêches de l'AFP. L'association des 258 verbes ainsi obtenus et classifiés à l'analyse de leur comportement en incise citationnelle a par ailleurs fait l'objet d'une intégration dans le Le*fff*.

3.2 Chaîne de traitement

3.2.1 Vue générale et composants

La chaîne de traitement réalisée pour la détection automatique de citations dans les corpus de dépêches de l'AFP est illustrée par la figure 7.9. Elle s'appuie sur la chaîne d'analyse linguistique surfacique SxPipe, introduite au chapitre 5 (section 3.1.1), qui permet la conception et l'application modulaire de grammaires locales.

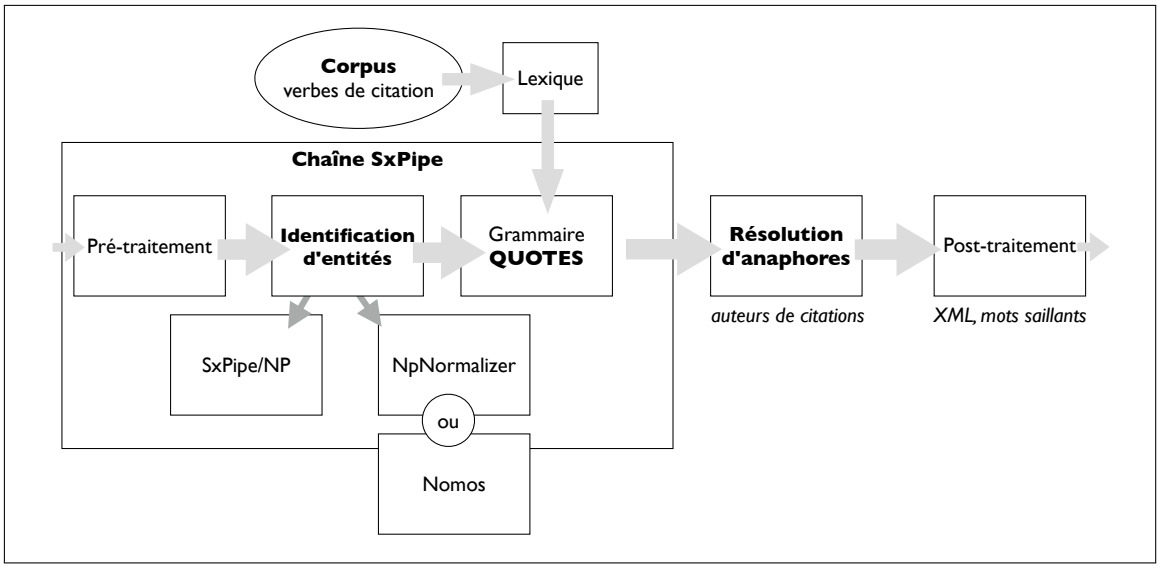


Figure 7.9 : Vue générale et composants de la chaîne de détection de citations.

quotes La grammaire non contextuelle quotes a été développée dans ce cadre et intègre la reconnaissance :

- des configurations usuelles de DR (DI, DD_{ini} et DD_{inc}), dès lors que des guillemets apparaissent à un moins un endroit de la phrase, y compris lorsque le prédicat citationnel est

prépositionnel (*selon, d'après...*) ;

- des configurations dites mixtes, où les verbatims ne font pas à proprement parler partie d'une proposition analysable comme l'objet d'un prédicat verbal de citation ;

- des auteurs de ces citations ; les citations reconnues pour lesquelles les auteurs ne correspondent pas aux critères informatifs introduits précédemment (auteurs collectifs, anonymes, indéfinis) sont exclues des résultats retournés par QUOTES.

La grammaire QUOTES est précédée de traitement usuels (segmentation en phrases[14], en tokens et en formes, reconnaissance de formes spéciales telles que les URL...) et d'un processus d'identification d'entités tel que décrit dans le présent travail.

Les règles de QUOTES permettent la reconnaissance simultanée des verbatims (segments entre guillemets), des prédicats de citations (verbaux ou prépositionnels) ou des occurrences de verbes indiquant la présence d'une citation dite mixte, ainsi que des auteurs associés à ces composants de citations. Il s'agit des mentions entités et des GN marqués dans les étapes précédentes du traitement (cf. *infra*), ainsi que des pronoms personnels sujets. La reconnaissances des prédicats de citation verbaux repose sur leur présence dans un lexique associé à la grammaire, comprenant les formes fléchies utiles[15] des verbes obtenus selon la méthode expliquée précédemment. La classe associée à chacun de ces verbes détermine le type de configuration dans lesquelles il peut apparaître. QUOTES définit ainsi par exemple les règles suivantes (notation simplifiée) :

```
<Quote> = (<DR_incise> | <DR_indirect> | ...)
<DR_incise> = " <token>+ " virgule? <verbe_inc_présent> (<Auteur_pers> | <Auteur_GN>)
<DR_incise> = " <token>+ " virgule? aux_avoir <verbe_inc_partpassé> (<Auteur_pers> |
<Auteur_GN>)
<DR_incise> = " <token>+ " virgule? aux_avoir t-tiret <Auteur_pro> <verbe_inc_partpassé>
<DR_indirect> = <AUT> <verbe_ind> que (<token>* " <token>+ " <token>*)+
```

Identification d'entités L'étape de Reconnaissance d'Entités Nommées est réalisée de façon non-déterministe par le module NP de SxPipe, puis la désambiguïsation et l'alignement des mentions reconnues sur des entités recensées dans la base Aleda sont accomplis par le module NPNORMALIZER, introduit au chapitre 5 (section 3.2) et intégrée à la chaîne SxPipe. Ce processus d'identification pourra être pris en charge par le système Nomos dans de futurs travaux[16]. Le texte donné en entrée à QUOTES est donc muni d'annotations marquant les mentions d'entités ainsi que les entités dénotées.

Groupes nominaux Afin de permettre la détection de citations dont l'auteur est un groupe nominal (GN) référant dans une dépêche donnée à l'une de ces entités, le marquage des GN est ajouté au texte avant son analyse par QUOTES. Ce repérage se restreint ici aux GN considérés comme pertinents au niveau informatif pour l'utilisation de l'application par l'AFP. Nous avons concrètement procédé à la sélection et au marquage des GN de la façon suivante :

14. La segmentation tient compte dans ce contexte des frontières de phrase situées à l'intérieur de segments entourés par des guillemets. Le module QUOTES traite en effet le texte phrase par phrase : les phrases au sein de citations ne doivent donc pas être séparées afin de pouvoir être correctement analysées par QUOTES.

15. Les formes fléchies considérées sont celles de troisième personne du singulier du présent, de l'imparfait, du passé simple et du passé composé, ainsi que les formes gérondives permettant d'inclure les citations telles que « [...] », *ajoutant* que « [...] ». On ajoute également les formes utiles du verbe auxiliaire de passé composé *avoir* Ces formes sont obtenues à partir du lexique Le*fff*.

16. L'application de détection de citations a en effet été conçue parallèlement au développement du système Nomos et devait être exploitable par l'AFP comme par des applications tierces avant que Nomos n'atteigne des performances suffisantes pour un tel usage.

- À partir d'un corpus de dépêches, constitué pour l'apprentissage et l'évaluation de la résolution d'anaphores des auteurs de citations et décrit ci-après, les GN ont été extraits par application d'un segmenteur en constituants ou *chunker*[17]
- Les têtes nominales des GN extraits du corpus ont ensuite été listées et ordonnées selon leur nombre total d'occurrences ; parmi les plus fréquentes (seuil fixé à un minimum de 10 occurrences), les noms considérés comme informatifs dans le cadre de la présente application ont été sélectionnés. Il s'agit notamment de *maire, député, candidat, centriste, porte-parole*. La même opération a été répétée pour les adjectifs apparaissant dans les GN extraits (*adjoint, communiste*).
- Une grammaire locale simple a été écrite et ajoutée à la chaîne SxPipe pour le repérage des GN, reposant sur la liste des noms et adjectifs obtenues à partir du corpus, convertie en un lexique adapté aux grammaires locales de SxPipe. Cette grammaire permet ainsi de marquer les GN tels que *le député-maire de Caen*, *le ministre* ou *le candidat socialiste*.
- Cette grammaire locale est appliquée avant le module QUOTES et permet ainsi de reconnaître et de marquer de façon simple la plupart des GN pertinents pour notre tâche[18].

Pronoms Le repérage des pronoms personnels sujets (*il* ou *elle*) pouvant représenter l'auteur des citations à détecter est intégré aux règles de reconnaissance de la grammaire QUOTES. Ils sont ainsi marqués et doivent ensuite être mis en relation avec les entités auxquelles ils réfèrent au sein des dépêches traitées. On peut noter que, dans le cadre général de la résolution d'anaphores pronominales, une étape de discrimination des pronoms sujets à l'anaphore est accomplie, afin de ne pas considérer le pronom *il* dans *Il se trouve que…*, par exemple, comme anaphorique. Cette difficulté est évitée grâce au contexte citationnel qui seul déclenche la détection de pronoms personnels pertinents pour la tâche.

Résolution d'anaphores Une fois les citations reconnues et leurs auteurs marqués par QUOTES, seuls les résultats où l'auteur est une mention d'entité peuvent donner lieu à une interprétation directe et complète. Les GN et pronoms marqués comme auteurs sont donc soumis à un processus de résolution d'anaphores, reposant sur un système développé dans le cadre du Medialab. Il s'agit d'un système de résolution d'anaphores spécialisé pour ce cas d'anaphores en position d'auteurs de citations, qui tient donc compte des configurations discursives particulières induites par cette configuration. Le développement de ce système a suivi les étapes suivantes :

- Sélection d'un corpus d'exploration, d'apprentissage et d'évaluation : environ 1 200 dépêches correspondant au cas applicatif (campagne présidentielle française de 2012[19]) ont été sélectionnées puis annotées par la chaîne de traitement SxPipe intégrant le module QUOTES.
- Parmi ces dépêches, environ 800 ne présentaient pas ou très peu d'erreurs et ont été retenues pour la suite de la tâche. Les corrections nécessaires ont été faites et une exploration des différentes configurations de la chaîne de coréférence impliquant les entités, les GN et les pronoms auteurs de citations a été menée.

17. Ce chunker a été développé au sein de l'équipe-projet Alpage par Benoît Crabbé (communication personnelle). Il repose sur le modèle des CRF [LMP01] et est entraîné sur le corpus FTB [ACT03].

18. L'intégration d'une grammaire simple et spécialisée pour les GN a été préférée à l'application directe du chunker sur les contenus à analyser afin de conserver des temps de traitement courts. L'application du chunker nécessite en effet un temps de chargement non négligeable et demande de quitter la chaîne SxPipe puis d'y revenir pour l'application de QUOTES. Des développements plus complets permettraient de contourner ce problème. Le repérage des GN par cette méthode donne cependant des résultats satisfaisants et demeure simple à réaliser.

19. La sélection des dépêches selon ce critère est permise par le slug *France2012*, spécialement introduit dans la production de l'AFP pour la couverture de cet événement et systématiquement ajouté aux dépêches qui en font état.

- L'exploration de corpus a donné lieu à la conception d'un système d'apprentissage, dont les traits sont dérivés des différentes observations sur les configurations de la chaîne de coréférence. Il s'agit notamment de la prise en compte de verbes de citations impliquant l'identité de l'auteur avec la citation précédente (*continue-t-il, ajoute-t-il,* de l'observation de la structure rédactionnelle (citation et entité nommée dans le titre de la dépêche, par exemple), et de traits usuels dans la tâche de résolution d'anaphores (distance en phrases et en mots entre l'anaphore et le candidat antécédent, par exemple).

- Le corpus constitué pour cette tâche a donné lieu à une annotation manuelle des chaînes de coréférence impliquant les auteurs GN et pronoms marqués par QUOTES, afin de fournir au système de résolution des données de référence pour l'apprentissage et l'évaluation. Une interface graphique a été développée afin de permettre aux journalistes du Medialab de procéder à cette annotation. Chaque auteur anaphorique devait y être relié à un antécédent, au moins l'un des antécédents d'une même chaîne coréférentielle devant être une entité nommée. Dans le cas contraire, la dépêche concernée est exclue du corpus de référence.

- L'apprentissage du système à partir de ces données de référence a permis d'aboutir, après une évaluation croisée par division du corpus en 10 lots, à des résultats de résolution très satisfaisants : une précision de 97% et un rappel de 91%, donnant un F-score (F1) de 94%.

On peut observer que ces performances sont obtenues sur des données de référence où les erreurs des étapes précédentes du traitement ne figurent plus, ce qui ne serait pas le cas dans la configuration applicative. On note cependant que le périmètre restreint et fortement spécifié du problème à traiter permet d'obtenir une précision élevée, constatée lors de l'exploitation du système dans le cas applicatif lui-même.

Post-traitement Après l'application de la grammaire QUOTES, on obtient des dépêches annotées à la fois en entités (entités nommées et références vers la base Aleda) et en citations. Ces annotations sont converties dans un format XML spécifié pour l'utilisation des données dans l'application d'indexation et de recherche. On extrait également de chaque dépêche les mots obtenant les scores de saillance [20] les plus élevés. Ces mots sont fournis à l'application pour y constituer des termes d'indexation thématique utiles au processus de Recherche d'Information, tels que *nucléaire* ou *retraites*.

3.3 Résultats et démonstration

La chaîne de traitement présentée ici a été appliquée à un ensemble d'environ 20 000 dépêches en relation avec la campagne présidentielle française de 2012. Le nombre de citations détectées et associées à des auteurs s'élève dans ce corpus à près de 95 000, dont 45% avec des auteurs anaphoriques. Cette distribution par type d'auteurs montre l'importance de l'application du module de résolution d'anaphores, qui permet d'obtenir un rappel de citations élevé.

La qualité de la chaîne de traitement a été évaluée par les journalistes du Medialab, qui ont constaté de bonnes performances générales. Une évaluation plus approfondie et reposant sur les métriques adéquates (précision et rappel) demeure à mener, mais les résultats obtenus ont été considérés comme bons et de nature à permettre une exploitation directe. Celle-ci comporte plusieurs volets :

- Indexation des dépêches sur les auteurs de citations et les mots saillants détectés en post-traitement, les slugs et catégories IPTC associés aux documents, et indexation plein-texte.

20. Ces scores de saillance sont dérivés d'un test t relatif au corpus de dépêches sur le sujet de la campagne présidentielle décrit précédemment.

- Stockage des dépêches et de des index dans un serveur dédié.
- Intégration dans un moteur de recherche Solr [21] (cf. figure 7.10).
- Intégration du moteur de recherche dans une interface graphique de recherche de dépêches conçue par l'AFP : les champs de recherche permettent de spécifier des requêtes par date, thème (indiqué par les mots saillants) et auteur de citation (cf. figure 7.11).
- Intégration du moteur de recherche dans une application cliente : le quotidien français *Libération* a proposé sur son site, durant la campagne présidentielle, une plateforme d'accès au moteur de recherche de citations développé au Medialab à partir de la chaîne de traitement présentée ici. (Plateforme intitulée *Le match des mots*, cf. figure 7.12).

Les trois premiers points de cette intégration peuvent être répétés sur d'autres corpus de dépêches selon les besoins applicatifs identifiés par le Medialab.

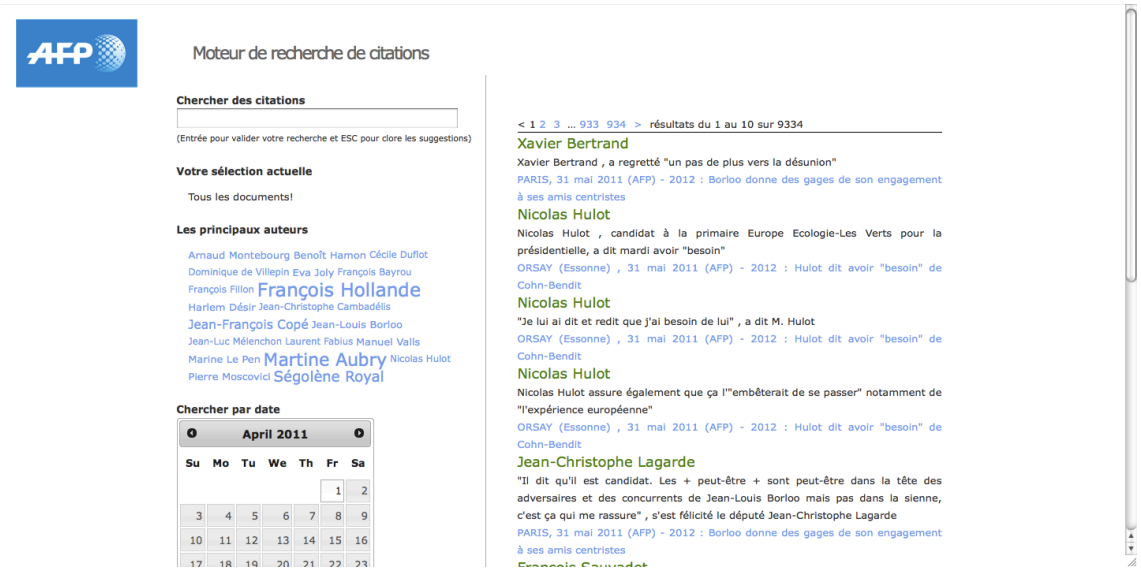


Figure 7.10 : Intégration de la détection de citations : moteur de recherche Solr.

21. `http://lucene.apache.org/solr/`

FIGURE 7.11 : Interface graphique de recherche de citations dans les contenus AFP.

Figure 7.12 : Plateforme de recherche de citations dans les contenus AFP *Le match des mots* (*Libération*).

Conclusion

Bilan

Les travaux présentés ont proposé une approche de l'enrichissement de contenus textuels, besoin formulé dans un cadre industriel par l'AFP. Son traitement a nécessité de délimiter le domaine dont relève ce sujet, tant d'un point de vue technologique que méthodologique. C'est dans cette perspective que nous avons donné une description générale du Web Sémantique, qui se présente entre autres comme le paradigme de publication de contenus textuels auquel se rattachent en nombre croissant des acteurs de la diffusion et du traitement de l'information. Parallèlement à ce cadre de définition pratique, nous avons envisagé l'enrichissement de contenus textuels comme un objectif applicatif pour lequel il s'agit de déterminer la méthodologie adéquate, en identifiant notamment les disciplines et champs de recherche pertinents pour sa mise en œuvre.

L'AFP étant le lieu d'une production massive de données textuelles, l'emploi de techniques de TAL apparaît comme une nécessité afin de mener à bien l'enrichissement des contenus considérés. En examinant l'historique et les objectifs poursuivis par la tâche d'Extraction d'Information (EI), on observe par ailleurs que les visées du Web Sémantique en termes de traitement et d'exploitation de l'information, en grande partie disponible sous forme textuelle, se placent dans la lignée de la recherche de compréhension automatique du langage naturel caractérisant le TAL et l'intelligence artificielle. Notre présentation du Web Sémantique insiste sur le rôle structurant joué par l'Annotation Sémantique (AS) dans ce paradigme, en tant que moyen *sine qua none* de réalisation de l'interprétabilité automatique. Nous avons proposé de considérer à ce niveau une seconde relation essentielle entretenue entre EI et Web Sémantique : l'EI est en effet une composante indispensable à la mise en œuvre de l'AS dès lors qu'il s'agit de larges contenus textuels à traiter. Dans le contexte applicatif de l'AFP, où les entités forment la cible principale des métadonnées visées pour l'enrichissement, la Reconnaissance d'Entités Nommées (REN) constitue plus particulièrement la méthode appropriée afin de faire émerger des contenus les éléments informatifs pertinents pour cette tâche.

Notre étude a cependant cherché à énoncer les limitations inhérentes à l'EI et à la REN quant à la sémantique requise dans le cadre de l'AS et du Web Sémantique : celle-ci relève en effet d'une formalisation, reposant principalement sur la conceptualisation ontologique, qui la distingue des modèles classificatoires et typologiques usuellement employés en EI. Les modèles sous-jacents à l'AS correspondent en effet à des spécifications explicites des éléments informatifs représentés et mentionnés au sein des contenus ; ces spécifications s'appuient sur les langages et pratiques standardisées proposées par le Web Sémantique : les langages RDF et OWL et surtout les Linked Data, qui réalisent en grande partie ses objectifs de mise en réseau des connaissances.

Nous avons ainsi proposé de concevoir la mise en œuvre de l'enrichissement de contenus textuels de l'AFP comme une procédure d'Annotation Sémantique, reposant sur des ressources conformes aux standards requis, et intégrant pleinement un composant d'EI sous forme de REN afin de traiter en particulier les entités. Celles-ci posent plus spécifiquement la nature de la

sémantique véhiculée par les annotations produites : elles intéressent en effet le processus d'enrichissement en tant qu'individus, au cœur des événements structurants de l'actualité. Ce statut trouve une représentation dans les instances ontologiques des ressources d'AS, au-delà des types usuels — PERSONNE, ORGANISATION, LIEU — ciblés par l'EI. La sémantique ainsi exprimée par les annotations et portée par les métadonnées d'enrichissement est de nature référentielle : elle met en relation de façon univoque et explicite les mentions textuelles et les entités dénotées au travers de leur représentation ontologique.

Cette mise en œuvre de l'enrichissement par l'Annotation Sémantique s'inscrit ainsi de façon centrale dans la problématique de la dénotation et du moyen d'établir automatiquement cette relation entre mention et entité. Nous avons porté notre recherche méthodologique à ce sujet sur les travaux menés depuis 2009 dans le cadre de la campagne TAC-KBP, où la tâche de *Population de bases de connaissances* intègre de façon explicite et généralisée le problème du *Liage* des mentions textuelles rencontrées au sein de documents aux entités qu'elles dénotent, celles-ci étant préalablement rassemblées en un inventaire de descriptions. Le cadre méthodologique du Liage d'Entités nous a permis de compléter l'approche proposée par une prise en charge systématique de la dénotation, incluant la détection des cas de mentions dénotant des entités non recensées par les ressources employées. Le terme d'*identification d'entités* qualifie ainsi dans notre travail le processus par lequel la dénotation est explicitée par la mise en relation de mentions textuelles et d'individus représentés sous la forme de descriptions. Le Liage, qui repose dans la plupart des systèmes sur des mesures de similarité contextuelle entre occurrences de mentions et informations rassemblées au sujet des entités cibles, est enrichi dans notre implémentation par la prise en compte des métadonnées de documents caractérisant les dépêches de l'AFP, notamment la catégorisation thématique selon la taxonomie de l'IPTC.

La mise en œuvre de l'enrichissement dans le contexte particulier de notre tâche devait également tenir compte du caractère brut des contenus à traiter, dont le Liage dans le cadre de TAC-KBP ne tient pas compte. Cet aspect des données soumises à l'annotation demandait que soit prise en compte la propagation d'erreurs possible dans une configuration modulaire et séquentielle telle que celle que nous proposons : les systèmes de REN intégrés en amont de l'identification à proprement parler peuvent introduire un certain nombre d'erreurs, en particulier des faux positifs et des correspondances partielles, de nature à compromettre la qualité de l'enrichissement obtenu à l'issue du Liage.

Les mentions erronées retournées par la REN sont en effet sans objet quant à l'identification ; les métadonnées résultant de mises en relation entre des segments non dénotationnels ou correspondant à des mentions incorrectes et des descriptions d'entités sont alors un facteur de bruit qu'il s'agit d'éviter, notamment dans le cadre applicatif de la production de l'AFP. Nous avons ainsi proposé une modification de la méthode générale de Liage en introduisant une fonctionnalité d'évaluation des mentions présentées à l'identification, associée à la possibilité de leur élimination en cas de dénotation jugée peu probable. En appui de cette fonctionnalité, l'étape de Liage a été envisagée sur des analyses de REN non-déterministes, dans l'idée que les possibilités de dénotation étant donné un contexte permettent de sélectionner ou d'écarter certaines lectures au niveau surfacique. L'approche proposée ici est ainsi qualifiée de modulaire et jointe, dans la mesure où les deux tâches de REN et de Liage sont accomplies en séquence par des modules distincts, mais où la seconde réalise l'analyse finale concernant les deux niveaux.

Cette mise en œuvre de l'identification à partir de contenus textuels bruts a fait l'objet du développement du système Nomos, prévu pour intégrer tout système de REN disponible. Les expériences menées dans le cadre du présent travail se sont appuyées sur les systèmes SxPipe/NP et

LIAne, issus de la recherche académique. La conception de Nomos a fait suite à un premier module d'identification, npNormalizer, destiné à répondre de façon immédiate au besoin d'enrichissement dans une première phase de développement destiné à l'AFP. npNormalizer, qui repose sur des heuristiques simples, est directement intégré à la chaîne SxPipe à la suite de la REN fournie par NP. Ce module devait tenir lieu de baseline en termes de performances d'identification ainsi que de repère quant aux aspects de fonctionnement à améliorer avec Nomos. Largement bénéficiaire de corrections ciblées et constantes par retour des utilisateurs, il s'est néanmoins révélé difficile à concurrencer par l'approche caractérisant le système Nomos. Celui-ci, reposant sur une méthode d'apprentissage supervisé et une sélection systématique de traits parmi un ensemble de critères de description contextuelle liée à la dénotation, obtient des résultats encourageants dans le cadre d'une évaluation sur des données spécialement constituées pour notre tâche. Les performances de Nomos, supérieures à celles de npNormalizer au cours de cette évaluation, ont pu confirmer la pertinence de l'approche jointe quant au problème de l'identification. Des évaluations menées dans le cadre de tâches spécifiques telles que la Recherche d'Information, où les entités identifiées dans les contenus jouent le rôle de descripteurs pour l'indexation, ont cependant montré une efficacité moindre de Nomos en regard de npNormalizer.

Parallèlement au développement du système Nomos, nous avons travaillé à l'intégration de l'identification d'entités dans les chaînes de traitement de dépêches de l'AFP avec deux objectifs principaux. Il s'est agi d'une part de la création et de la population initiale d'un référentiel de métadonnées propre à l'AFP, selon des critères de pertinence relatifs au périmètre des entités concernées par la couverture de l'Agence. Ce référentiel, AMO, présente une structure ontologique à la fois simple et adaptable, et comprend une population d'entités munies d'informations quant à leur utilisation dans la production de l'AFP. Ces entités sont intégrées au réseau des Linked Data par le biais de liens de synonymie définis vers des ensembles de données tels que ceux du *New York Times*. Nous avons d'autre part contribué à la production d'une application ciblée reposant de façon cruciale sur l'identification d'entités : un système de détection automatique de citations avec attribution d'auteurs a en effet été employé dans la conception d'une application de Recherche d'Information durant la campagne présidentielle française de 2012, et utilisé dans une plateforme en ligne par le journal *Libération* à la même occasion.

Perspectives

Les perspectives immédiates faisant suite aux travaux que nous avons présentés ici concernent en premier lieu le développement et l'amélioration du système Nomos, qu'il s'agit de rendre utilisable dans le cadre de traitement de contenus de l'AFP avec un niveau de qualité équivalent voire supérieur à celui du module npNormalizer. De nouvelles expériences portant sur l'intégration de traits et des paramétrages de l'apprentissage supervisé figurent parmi ces perspectives à court terme.

L'identification dans le système Nomos reposant en grande partie sur l'examen des contextes respectifs des mentions textuelles et des entités disponibles pour la tâche, il paraît également pertinent d'intégrer à son fonctionnement une synchronisation avec le référentiel AMO : celui-ci reflète en effet, au niveau des entités qu'il recense, les contextes d'emploi dans les corpus de l'AFP ; ces informations pourraient ainsi venir enrichir les descriptions constituées au sein de la base de connaissances de Nomos et améliorer l'évaluation des similarités contextuelles lors de la procédure d'alignement.

Plus généralement, deux lignes importantes de développement concernant l'identification sont à considérer afin de compléter les travaux réalisés à ce jour : au niveau du problème dénotationnel

lui-même, d'une part, l'identification doit tenir compte du phénomène de la métonymie, qui participe de la non-univocité entre mentions textuelles et entités. À un niveau que l'on peut caractériser de fonctionnel d'autre part, une architecture reposant sur l'identification automatique doit intégrer l'aspect dynamique du périmètre référentiel formé par les entités, autrement dit le problème de l'apparition de nouvelles entités dans l'actualité et de leur prise en compte dans la tâche d'identification.

Identification d'entités et métonymie

Afin de mener à bien la résolution explicite du phénomène dénotationnel à l'œuvre entre mentions textuelles et entités en tant qu'individus extra-linguistiques, la tâche d'identification d'entités repose en grande partie, dans nos travaux comme dans le cadre général de TAC-KBP, sur les possibilités de dénotation d'une entité donnée, autrement dit sur l'établissement des mentions pouvant la dénoter. Cette procédure tient en général compte des problèmes de variation surfacique et encyclopédique touchant la dénotation des entités. La métonymie constitue à cet égard un facteur de variation qu'il s'agit de considérer au même titre que les autres : elle intervient en effet dans le phénomène dénotationnel et l'ignorer peut mener à des cas d'identification erronée.

La métonymie consiste, au niveau des entités, à employer une mention normalement associée à une entité E_1 pour dénoter une entité distincte E_2, cette opération étant possible lorsque E_1 et E_2 présentent les caractéristiques usuelles de la métonymie lexicale, telles que la proximité ou l'ingrédience. Ainsi, dans l'exemple suivant :

(26) Comme prévu, Barcelone a eu la possession du ballon mardi soir face au Bayern (0-4).

la mention *Barcelone* réfère à l'équipe de football de la ville catalane, dont le nom canonique est « FC Barcelone », et non à la ville elle-même. Sans prise en compte de la métonymie illustrée ici, un système d'identification tel que Nomos pourra lier la mention à la ville, qui est la seule entité recensée dans Aleda avec cette variante, ou retourner le cas NIL si la similarité entre les connaissances disponibles sur la ville et le contexte d'occurrence de cet exemple est jugée trop faible.

Métonymie et Entités Nommées

Dans son étude des Entités Nommées (EN) et de leur traitement en TAL, Ehrmann [Ehr08] aborde la question de la métonymie et de sa détection automatique au niveau des EN. Elle rappelle la teneur des travaux fondateurs en la matière de Markert, Nissim et Hahn [MH02], qui ont souligné le manque d'analyses et de ressources concernant la métonymie dans le cadre du TAL et non seulement de la description linguistique et lexicale. Leur examen de corpus [MN03 ; MN06] révèle à la fois une régularité du phénomène, son caractère productif et sa fréquence relativement élevée (entre 17% et 30% d'occurrences métonymiques d'EN selon les types). La régularité de la métonymie à l'égard des EN paraît concerner essentiellement les lieux et les organisations, pour lesquelles Markert et Nissim proposent des *patrons* explicitant le glissement des classes sémantiques des EN concernées, tels que `organisation-for-product` (*acheter une BMW*) ou `place-for-event` (*un vétéran du Vietnam*).

À l'occasion de la conférence SemEval de 2007, Markert et Nissim [MN07a ; MN07b] proposent une tâche de détection automatique des emplois métonymiques d'EN de types PERSON et ORGANIZATION, par opposition à leurs emplois littéraux ; le cas des interprétations mixtes est également considéré (métonymies multiples à des niveaux différents, interprétations littérale et métonymique simultanées). Leur approche, comparée à la tâche de désambiguïsation lexicale (*Word Sense Disambiguation*), repose sur l'apprentissage supervisé, où la cible de la tâche de classification est une classe sémantique plutôt qu'un mot ou un sens.

On peut observer que cette prise en charge de la métonymie au niveau des EN s'inscrit de façon pleine dans le paradigme de classification usuel de l'Extraction d'Information et de la REN tel que nous l'avons abordé au chapitre 2, où les EN sont considérées en fonction de leur classe sémantique. La distinction des emplois littéraux et métonymiques correspond à l'attribution de la classe correcte étant donné une occurrence d'EN contextualisée, par assignation du patron adéquat, tel que ceux évoqués ci-dessus : `organisation-for-product` pour *acheter une BMW* ou `place-for-event` pour *un vétéran du Vietnam*.

Métonymie et sémantique référentielle

L'approche de la métonymie dans le contexte de la tâche d'identification d'entités entretient des relations avec les travaux portant sur les EN mais, comme nous l'avons souligné au cours de ce travail, l'identification se distingue de la REN notamment par la recherche d'une sémantique référentielle et non seulement typologique. Il s'agit ainsi non pas de classifier de façon plus ou moins fine et précise une mention d'entité mais d'indiquer de façon explicite le référent qu'elle dénote, étant donné un ensemble d'entités disponibles.

Les cas de métonymie dans le cadre d'une telle tâche peuvent être vus comme relevant directement de la définition de ce phénomène intervenant au niveau lexical : la métonymie permet en effet de désigner un objet du monde par l'emploi d'un terme en désignant normalement un autre ; dans le cas des entités, un individu est alors dénoté par une mention associée de façon usuelle à un autre. La résolution de métonymie en REN s'attache à détecter le phénomène lui-même et à spécifier la classe sémantique résultant du glissement ainsi réalisé, tandis que l'identification doit permettre de savoir quelle entité est effectivement dénotée, y compris par le biais d'une mention qui y réfère par métonymie et non en tant que simple variante.

La distinction entre traitement de la métonymie pour l'identification et résolution de métonymie conjointe à la REN ne relève pas uniquement du niveau d'analyse, classificatoire ou référentiel. Le périmètre des mentions concernées est également différent, dans la mesure où certains glissements typologiques n'induisent pas nécessairement un changement de référent. Ainsi, dans un cas correspondant au patron `organisation-for-members`, tel que

(27) IBM annonce une avancée de ses actions en février

il ne semble pas incorrect d'identifier l'entreprise informatique IBM, même si l'approche typologique rencontre un emploi jugé métonymique, où ce sont des personnes, les membres de l'organisation, qui sont en fait évoquées, ce que supporte le verbe *annoncer*. Une identification correspondant exactement à cette détection de métonymie reviendrait à disposer d'une description d'entité renvoyant à ces membres ; des descriptions de ce type ne sont généralement pas incluses dans les ressources des Linked Data ou les bases de connaissances telles que celles de la tâche de Liage de TAC-KBP, dont les entités en tant qu'individus précisément identifiables sont la population essentielle.

Plus généralement, les métonymies les plus courantes et dont la résolution paraît la plus pertinente pour la tâche d'identification semblent concerner les entités de type ORGANISATION régulièrement dénotées à l'aide de mentions d'entités de type LIEU, comme dans l'exemple 26. Une étude reposant sur des corpus de dépêches de l'AFP devrait être menée afin de confirmer la proéminence de cette classe de métonymie et ainsi l'utilité de sa prise en charge dans la tâche d'identification.

Méthodologie

L'intégration des dénotations métonymiques à la tâche d'identification est envisagée dans les prochains développements du système Nomos ainsi que de la ressource Aleda. Nous proposons

en effet de faire reposer la résolution des métonymies entre mentions d'entités de type LIEU et entités de type ORGANISATION sur la régularité du phénomène et sur la possibilité de disposer préalablement des couples mentions / entités qu'il concerne. Il s'agirait concrètement de mener une collecte sur corpus des mentions métonymiques correspondant à ce cas de figure, puis de dériver de cette exploration un ensemble de règles indiquant, pour une entité donnée, les entités entretenant une relation métonymique régulière avec elle et dont les mentions peuvent alors être ajoutées à l'ensemble des variantes déjà recensées pour cette entité. On peut par exemple s'attendre à une règle indiquant que les équipes de football européennes peuvent être dénotées par les noms des villes auxquelles elles sont rattachées. Ainsi, pour le club de football « FC Barcelone », la variante *Barcelone* pourra être ajoutée aux variantes déjà recensées dans Aleda : *Barcelona Atlètic, Barça, Barça Atlètic, FC Barcelone, FC Barcelone Atletic*. Lors de la procédure d'alignement, la mention *Barcelone* comptera alors parmi les entités candidates possibles la ville catalane et le club de football, et l'entité adéquate pourra être sélectionnée en fonction des similarités contextuelles établies par le système.

Cette proposition de méthodologie requiert un moyen de mise en relation systématique des entités pouvant entretenir une relation métonymique — dans notre exemple, il s'agit de savoir que l'entité « FC Barcelone » peut être désignée par la ville catalane en vertu de sa localisation géographique. Cette mise en relation peut être accomplie de façon automatique notamment par l'exploitation des descriptions au format RDF disponibles dans des ressources telles que DBpedia. Ainsi, la description DBpedia du club de football cité ici le catégorise à l'aide du concept `dbpedia-owl:SportsTeam` et lui attribue une propriété `dbpedia-owl:headquarter` dont la valeur est la description DBpedia de la ville de Barcelone en Espagne. Cette représentation systématique d'un large nombre d'organisations constitue une source d'information que nous envisageons d'explorer afin d'enrichir la base Aleda et de tenir compte d'un certain nombre de dénotations métonymiques.

Identification d'entités nouvelles

Comme nous l'avons expliqué au cours de notre travail, la tâche d'identification requiert la mise à disposition d'un ensemble d'entités, rassemblées sous la forme de descriptions dans des ressources adaptées — bases de connaissances dans le cadre de TAC-KBP, base Aleda pour le système Nomos. Nous avons souligné que ces ressources sont nécessairement incomplètes au regard de l'ensemble des entités pouvant être mentionnées dans les corpus traitées : leur construction automatique peut être lacunaire d'une part, et leur périmètre concerne généralement les entités présentant une certaine notoriété publique d'autre part. Enfin et surtout, ces ressources sont constituées de façon statique à un moment de l'actualité, quand celle-ci fait apparaître de façon constante et régulière de nouvelles personnalités et autres entités dans les événements relatés. Ce dernier facteur d'incomplétude des ressources peut se révéler particulièrement problématique pour l'enrichissement des contenus de l'AFP, synchronisée de façon inhérente avec l'actualité et pour laquelle la précision sur les derniers éléments de l'information est cruciale.

Parallèlement à l'accomplissement de la tâche d'identification, il paraît donc nécessaire de développer les moyens d'une prise en charge systématique des entités émergeant au fil de l'actualité. Il est dans un premier temps envisageable de procéder à la reconstruction de la base Aleda avec une fréquence élevée ; cette reconstruction est cependant dépendante des dépôts de l'encyclopédie Wikipedia, effectué toutes les quelques semaines par l'organisation chargée de sa maintenance. La nouveauté qu'il s'agit ici d'intégrer aux traitements s'entend cependant selon une temporalité bien plus resserrée, de l'ordre de quelques jours voire quelques heures.

On peut néanmoins considérer Wikipedia dans sa version en ligne comme une source d'information rapidement mise à jour et permettant ainsi de fournir les descriptions nécessaires à l'identification de nouvelles entités. La prise en charge des entités émergentes est ainsi envisagée

comme un processus dynamique, partie prenante de la tâche d'enrichissement des contenus. À partir d'entités alignées sur le cas NIL par un système tel que Nomos au cours d'une période donnée — un jour ou une semaine par exemple —, il est en effet possible de constituer un ensemble d'entités candidates, pour lesquelles deux procédures systématiques peuvent être définies. Le système d'enrichissement peut d'une part constituer à partir du nom correspondant à ces entités candidates une requête dirigée vers le Web, et plus particulièrement vers Wikipedia ; cette requête peut interroger le moteur de recherche de Wikipedia dans ses différentes versions linguistiques et ainsi rapporter un certain nombre d'éléments utiles pour l'identification, tels qu'un ou plusieurs articles dont le nom donné est le titre. D'autre part, les entités candidates peuvent être présentées à l'administration du référentiel de métadonnées de l'AFP, AMO ; il s'agit alors, pour l'administrateur, de déterminer si ces candidats doivent ou non acquérir le statut d'entités dans AMO, ainsi que de collecter les éléments nécessaires à l'établissement de ce statut, en premier lieu une URI. Ces deux procédures peuvent par ailleurs être couplées afin de fournir à l'administration du référentiel AMO des indications utiles à la décision. Au terme de cette identification dynamique, les nouvelles entités recensées au niveau de l'AFP peuvent être présentées à la base Aleda et ainsi permettre sa mise à jour.

La problématique de l'émergence de nouvelles entités et de leur traitement pour l'enrichissement des ressources associées à l'identification fait l'objet de travaux en cours dans le cadre du projet de recherche EDyLex (projet ANR-09-CORD-008), dans lesquels l'AFP et l'équipe-projet Alpage sont partenaires. Ces travaux devraient donner lieu prochainement à des propositions d'intégration ainsi qu'à la production de prototypes d'enrichissement des ressources référentielles AMO et Aleda.

Annexes

Annexe A

Agence France-Presse et taxonomie IPTC

1 Catégorisation IPTC et slugs

Table des mots clefs (fr) à utiliser au [13 11 2009]

Sujet	Sujets et rubriques IPTC : nom des rubriques	Synonymes : nom des rubriques dans les systèmes AFP	Keywords : Mots de slugs
CLT	Arts, culture, et spectacles (01000000)	[Arts, culture et spectacles]	
CLT	Archéologie (01001000)	[Archéologie]	[Archéologie]
CLT	Architecture (01002000)	[Architecture]	[Architecture]
CLT	Tauromachie (01003000)	[Tauromachie]	[Tauromachie]
CLT	Festivals et commémorations (01004000)	[Carnaval] + [Festival] + [festivités]	[Carnaval] + [Exposition] + [Festival] + [commémoration] + [festivités]
CLT	Cinéma (01005000)	[Cinéma]	[Cinéma]
CLT	Festival de cinéma (01005001)	[cinéma festival]	[cinéma-festival]
CLT	Danse (01006000)	[Danse]	[Danse]
CLT	Mode (01007000)	[Mode]	[Mode]
CLT	Langue (01008000)	[langage]	[langage]
CLT	Bibliothèque et musée (01009000)	[musées]	[musée] + [musées]
CLT	Littérature (01010000)	[Littérature]	[Littérature] + [livre] + [livres]
CLT	Musique (01011000)	[Musique]	[Musique] + [opéra]
CLT	Peinture (01012000)	[Peinture]	[Peinture]
CLT	Photographie (01013000)	[Photo] + [Photographie]	[Photo] + [Photographie]
CLT	Radio (01014000)	[radio]	[radio]
CLT	Sculpture (01015000)	[Sculpture]	[Sculpture]
CLT	Télévision (01016000)	[Télévision]	[Télévision] + [audiovisuel] + [tv]
CLT	Théâtre (01017000)	[Théâtre]	[Théâtre]
CLT	Patrimoine (01018000)	[Patrimoine]	[Patrimoine]
CLT	Coutumes et traditions (01019000)	[tradition]	[tradition]
CLT	Arts (général) (01020000)	[Art] + [Arts]	[Art] + [Arts]
CLT	Divertissement (01021000)	[Spectacle] + [Spectacles]	[Spectacle] + [Spectacles]
CLT	Culture (général) (01022000)	[Culture]	[Culture] + [science-fiction]
CLT	Bande dessinée (01024000)	[BD]	[BD]
CLT	Dessin animé (01025000)	[cinéma animation]	[cinéma-animation]
CLT	Média (01026000)	[Média] + [Médias]	[Média] + [Médias]
CLT	Internet (01027000)	[Internet]	[Internet]
CLJ	Police et justice (02000000)	[Police et justice]	
CLJ	Criminalité (02001000)	[Criminalité]	[Criminalité] + [piraterie] + [pédophilie]
CLJ	Homicide (02001001)	[Homicide]	[Homicide] + [meurtre] + [meurtres]
CLJ	Vol (02001003)	[vol]	[Vol]
CLJ	Trafic de drogue (02001004)	[trafic drogues]	[Drogue]
CLJ	Viol (02001005)	[viol]	[viol]
CLJ	Agression (02001006)	[agression]	[agression] + [agressions]
CLJ	Enlèvement (02001007)	[Enlèvement]	[enlèvement] + [otage] + [otages]
CLJ	Incendie criminel (02001008)	[pyromane]	[pyromane]
CLJ	Système judiciaire (02002000)	[Justice]	[Justice]
CLJ	Police (02003000)	[Police]	[Gendarmerie] + [Police]
CLJ	Enquête (02003002)	[enquête]	[enquête]
CLJ	Arrestation (02003003)	[arrestation]	[arrestation]
CLJ	Peines (02004000)	[condamnation]	[condamnation]
CLJ	Amende (02004001)	[amende] + [amendes]	[amende] + [amendes]
CLJ	Prison (02005000)	[Prison]	[Prison]
CLJ	Droits (02007000)	[Droits]	[Droits]
CLJ	Procès (02008000)	[Procès]	[Procès] + [assises]
CLJ	Accusation (02009000)	[accusation]	[accusation]
CLJ	Crime organisé (02010000)	[mafia]	[mafia]
CLJ	Droit International (02011000)	[DroitInternational]	[DroitInternational]
CLJ	Cour ou tribunal international (02011001)	[TPIR]	[CIJ] + [CPI] + [TPI] + [TPIR] + [TSL] + [TSSL]
CLJ	Extradition (02011002)	[extradition]	[extradition]
CLJ	Fraude (02012001)		[contrefaçon] + [escroquerie]
CLJ	Détournements (02012002)	[Blanchiment]	[Blanchiment]
CLJ	Corruption (02012006)	[Corruption]	[Corruption]
DIS	Désastres et accidents (03000000)	[Désastres et accidents]	
DIS	Sécheresse (03001000)	[Sécheresse]	[Sécheresse]
DIS	Tremblement de terre (03002000)	[séisme]	[séisme]
DIS	Famine (03003000)	[Famine]	[Famine]
DIS	Incendie (03004000)	[Incendie]	[Incendie] + [incendies]
DIS	Inondation (03005000)	[Inondation]	[Inondation] + [inondations] + [tsunami]
DIS	Désastre météorologique (03007000)	[Intempéries]	[Intempéries] + [canicule]
DIS	Accident nucléaire (03008000)		[accident-nucléaire]
DIS	Pollution (03009000)	[Pollution]	[Pollution]
DIS	Accident de la route (03010001)	[accident route]	[accident-route] + [route-accident]
DIS	Accident de chemin de fer (03010002)	[accident train]	[accident-train]
DIS	Accident dans l'air et l'espace (03010003)	[accident avion]	[accident-air] + [accident-aviation] + [accident-avion] + [aviation-accident]
DIS	Accidents maritimes (03010004)	[accident mer] + [naufrage]	[accident-mer] + [naufrage]
DIS	éruption volcanique (03011000)	[Volcan]	[Volcan]
DIS	Secours d'urgence (03012000)	[secours]	[secours]
DIS	Accident (général) (03013000)	[Accident] + [Accidents]	[Accident] + [Accidents]
DIS	Situation d'urgence (03014000)		[explosion]
DIS	Désastre (général) (03015000)	[Catastrophe] + [Catastrophes] + [Désastres]	[Catastrophe] + [Catastrophes] + [Désastres]
DIS	Catastrophe naturelle (03015001)	[Catastrophe naturelle]	[cyclone] + [ouragan] + [typhon]
DIS	Avalanche/Glissement de terrain (03015002)	[avalanche]	[avalanche]
DIS	Prévention et organisation des secours (03016000)	[Plans urgence] + [catastrophe plan]	
ECO	Economie et finances (04000000)	[Economie et finances]	
ECO	Agriculture (04001000)	[Agriculture] + [agr]	[Agriculture] + [agr]
ECO	Pêche (04001002)	[Pêche]	[Pêche] + [agr-mer] + [ali-mer] + [mer-agr] + [mer-ali]
ECO	élevage (04001004)	[élevage]	[élevage]
ECO	Viticulture (04001005)	[Viticulture]	[Viticulture] + [vins]
ECO	Chimie (04002000)	[Chimie] + [chm]	[Chimie] + [chm]
ECO	Biotechnologie (04002001)	[bio] + [biotechnologies]	[bio] + [biotechnologie] + [biotechnologies]
ECO	Hygiène et Cosmétiques (04002003)	[cos]	[cos] + [cosmétiques]
ECO	Pharmacie (04002006)	[pha]	[pha] + [pharmacie]
ECO	Informatique et technologie de l'information (04003000)	[Informatique] + [tec] + [technologies]	[Informatique] + [tec] + [technologies]
ECO	Réseaux et télécommunications (04003002)	[net]	[net]
ECO	Logiciels (04003005)	[logiciel] + [logiciels]	[logiciel] + [logiciels]
ECO	équipement de télécommunication (04003006)	[technologies télécoms]	[technologies-télécoms] + [tle]
ECO	Services télécoms (04003007)	[tel] + [télécommunications] + [télécoms]	[tel] + [télécommunications] + [télécoms]
ECO	équipement et immobilier (04004000)	[batiment] + [construction]	[batiment] + [construction]
ECO	Travaux Publics (04004001)	[BTP]	[btp]
ECO	Immobilier (04004003)	[imm] + [immobilier]	[imm] + [immobilier]
ECO	énergie et ressources (04005000)	[Energie] + [gaz] + [pétrole]	[Energie] + [eee] + [gaz] + [pétrole]
ECO	énergies alternatives (04005001)	[énergie solaire] + [éolien] + [éolienne]	[énergie-solaire] + [éolien] + [éolienne]
ECO	Charbon (04005002)	[Charbon] + [min]	[Charbon] + [min]
ECO	Pétrole et gaz (aval) (04005003)	[dis oil] + [oil dis]	[dis-oil] + [oil-dis]
ECO	Pétrole et gaz (amont) (04005004)	[oil]	[oil]
ECO	énergie nucléaire (04005005)	[nuc] + [énergie nucléaire]	[nuc] + [énergie-nucléaire]
ECO	électricité (04005006)	[Electricité]	[Electricité]
ECO	Traitement des déchets (04005007)	[eee env] + [env eee]	[eee-env] + [env-eee]
ECO	Alimentation en eau (04005008)	[approvisionnement eau]	
ECO	Ressources naturelles (04005009)	[min]	
ECO	Services financiers (04006000)	[audit] + [comptabilité] + [courtage] + [finance] + [ser] + [services]	[audit] + [comptabilité] + [courtage] + [finance] + [ser] + [services]
ECO	Banque (04006002)	[ban] + [banque]	[ban] + [banque]
ECO	Services de santé (04006005)	[san sfi] + [services santé]	[san-sfi] + [services-santé]
ECO	Assurances (04006006)	[ass] + [assurances]	[ass] + [assurance] + [assurances]
ECO	Sociétés de transport (04006012)	[Logistique] + [poste]	[Logistique] + [log] + [poste]
ECO	Ventes aux enchères (04006020)	[Enchères]	[enchères]

FIGURE A.1 : Table des slugs : catégories IPTC 01 à 04

ECO	Distribution (04007000)	[bcc] + [dis] + [distribution]	[bcc] + [dis] + [distribution]
ECO	Habillement (04007001)	[dis tex] + [tex dis]	[dis-tex] + [tex-dis]
ECO	Alimentation (04007003)	[ali] + [ali dis] + [dis ali]	[ali] + [ali-dis] + [dis-ali]
ECO	Vente par correspondance (04007004)	[VPC]	[VPC]
ECO	Commerce électronique (04007009)	[dis net] + [distribution internet] + [net dis]	[dis-net] + [distribution-internet] + [net-dis]
ECO	Commerce de luxe (04007010)	[Luxe]	[luxe]
ECO	Macro économie (04008000)	[croissance] + [économie]	[croissance] + [économie]
ECO	Banques centrales (04008001)	[BCE] + [Fed] + [bac] + [banque centrale]	[BCE] + [Fed] + [bac] + [banque-centrale]
ECO	Consommation (04008002)	[consommation] + [cso]	[consommation] + [cso]
ECO	Marché obligataire (04008003)	[marchés obligations] + [obl]	[marchés-obligations] + [obl]
ECO	Indicateurs économiques (04008004)	[idc] + [indicateur]	[idc] + [indicateur]
ECO	Dette des pays émergents (04008005)	[det]	[det]
ECO	Marché des changes (04008006)	[cha] + [changes]	[cha] + [changes] + [devises]
ECO	Dette publique (04008008)	[bud det] + [det bud] + [dette]	[bud-det] + [det-bud] + [dette]
ECO	Taux d'intérêts (04008009)	[tau] + [taux]	[tau] + [taux]
ECO	Institutions économiques internationales (04008010)	[BAD] + [BM] + [BRI] + [Berd] + [FMI] + [OMC] + [int]	[BAD] + [BM] + [BRI] + [Berd] + [FMI] + [OMC] + [int]
ECO	Commerce International (04008011)	[commerce] + [ech]	[commerce] + [ech]
ECO	Organisations économiques (04008013)	[orp]	[orp]
ECO	Inflation et déflation (04008016)	[inflation]	[inflation]
ECO	Marchés des matières premières (04008024)	[matières premières]	[mat] + [matières-premières]
ECO	Marchés (04009000)	[fut] + [marchés]	[fut] + [marchés]
ECO	Marché de l'énergie (04009001)	[marchés pétrole] + [mat]	[eee-mat] + [marchés-pétrole] + [mat-eee] + [mat-oil] + [oil-mat]
ECO	Métaux (04009002)	[aur mat] + [marchés métaux] + [mat aur] + [mat met] + [met mat]	[aur-mat] + [marchés-métaux] + [mat-aur] + [mat-met] + [met-mat]
ECO	Bourse (04009003)	[bourse]	[bou] + [bourse]
ECO	Denrées (04009004)	[marchés agriculture] + [mat agr]	[agr-mat] + [marchés-agriculture] + [mat-agr]
ECO	Médias (04010000)	[med] + [médias entreprise] + [médias entreprises]	[med] + [médias-entreprise] + [médias-entreprises]
ECO	Publicité (04010001)	[Publicité]	[Publicité]
ECO	édition (04010002)	[édition]	[Edition]
ECO	Services en ligne (04010006)	[internet entreprise] + [internet entreprises] + [med net] + [net med]	[internet-entreprise] + [internet-entreprises] + [med-net] + [net-med]
ECO	Industrie musicale (04010011)	[musique entreprise] + [musique entreprises]	[musique-entreprise] + [musique-entreprises]
ECO	Industries métallurgiques et mécaniques (04011000)	[mec] + [mécanique] + [métallurgie]	[mec] + [mécanique] + [métallurgie]
ECO	Industrie aéronautique et de l' espace (04011001)	[Aéronautique] + [aer] + [satellite]	[Aéronautique] + [aer] + [satellite]
ECO	Industrie automobile (04011002)	[Automobile] + [aut]	[Automobile] + [aut]
ECO	Industrie de défense (04011003)	[def] + [industrie défense]	[def] + [industrie-défense]
ECO	équipement électrique (04011004)	[ele] + [équipement électrique]	[ele] + [équipement-électrique]
ECO	Construction navale (04011008)	[chantier naval] + [mec mec] + [mer mec]	[chantier-naval] + [mec-mec] + [mer-mec]
ECO	Métaux et minéraux (04012000)	[met] + [métaux]	[met] + [métaux]
ECO	Matériaux de construction (04012001)	[ciment] + [matériaux]	[ciment] + [matériaux]
ECO	Or et métaux précieux (04012002)	[aur] + [diamant] + [métaux précieux]	[aur] + [diamant] + [métaux-précieux]
ECO	Sidérurgie (04012003)	[Sidérurgie]	[Sidérurgie]
ECO	Mines (04012005)	[Mine] + [Mines]	[Mine] + [Mines]
ECO	Industries de transformation (04013000)	[industrie]	[industrie]
ECO	Alimentation (04013002)	[Alimentation]	[Alimentation]
ECO	Papier et emballage (04013004)	[emballage] + [industrie papier] + [pap]	[emballage] + [industrie-papier] + [pap]
ECO	Textile (04013007)	[tex] + [textile]	[tex] + [textile]
ECO	Tabac (04013008)	[industrie tabac]	[industrie-tabac]
ECO	Tourisme et loisirs (04014000)	[Tourisme] + [tou]	[Tourisme] + [tou]
ECO	Casinos et jeux (04014001)	[casino]	[Courses-jeux] + [casino]
ECO	Hotellerie (04014002)	[hôtellerie]	[hôtellerie]
ECO	Restauration (04014004)	[Restauration]	[Restauration]
ECO	Industrie de transport (04015000)	[Transport] + [Transports] + [trn]	[Transport] + [Transports] + [trn]
ECO	Transport aérien (04015001)	[air trn] + [transport aviation] + [trn air]	[air-trn] + [transport-aviation] + [trn-air]
ECO	Transport ferroviaire (04015002)	[transport SNCF] + [transport rail]	[transport-SNCF] + [transport-rail]
ECO	Transport routier (04015003)	[route transport] + [transport route]	[route-transport] + [transport-route]
ECO	Transport fluvial et maritime (04015004)	[port] + [transport mer] + [trn mer]	[mer-trn] + [port] + [transport-mer] + [trn-mer]
ECO	Vie des sociétés (04016000)	[Entreprise] + [Entreprises]	[Entreprise] + [Entreprises]
ECO	Assemblées générales (04016002)	[ago]	[ago]
ECO	Problèmes de concurrence (04016004)	[concurrence]	[concurrence]
ECO	Fusions-acquisitions (04016005)	[acquisition] + [fus] + [fusion]	[acquisition] + [fus] + [fusion]
ECO	Commentaires d'analystes (04016006)	[reco]	[reco]
ECO	Défaillances d'entreprise (04016007)	[fai] + [faillite]	[fai] + [faillite]
ECO	Carnet des entreprises (04016008)	[dir]	[dir]
ECO	Dirigeants (04016011)	[direction] + [dirigeant] + [dirigeants]	[direction] + [dirigeant] + [dirigeants]
ECO	Contrats (04016013)	[com] + [commande] + [commandes] + [contrat] + [contrats]	[com] + [commande] + [commandes] + [contrat] + [contrats]
ECO	Contrats militaires (04016014)	[def com] + [défense contrats]	[def-com] + [défense-contrats]
ECO	Dividendes (04016015)	[Dividende] + [div]	[Dividende] + [div]
ECO	Résultats (04016018)	[res]	[res]
ECO	Introduction en bourse (04016019)	[cap] + [ipo] + [ipo bou]	[cap] + [ipo] + [ipo-bou]
ECO	Contrats publics (04016020)	[gov com]	[gov-com]
ECO	Entreprise en participation (04016023)	[coe] + [coentreprise]	[coe] + [coentreprise]
ECO	Restructurations (04016025)	[restructuration]	[restructuration]
ECO	Contentieux (04016027)	[Contentieux] + [jur]	[Contentieux] + [jur]
ECO	Marketing (04016029)	[Marketing] + [pro]	[Marketing] + [pro]
ECO	Patentes et brevets (04016031)	[brevet] + [brevets]	[brevet] + [brevets]
ECO	Fermetures d'usines (04016032)	[rst soc] + [soc rst]	[rst-soc] + [soc-rst]
ECO	Privatisations (04016034)	[privatisation]	[privatisation]
ECO	Indice et classement (04016036)	[not] + [notation]	[not] + [notation]
ECO	Recapitalisation (04016039)	[cap rst] + [rst cap]	[cap-rst] + [rst-cap]
ECO	Activité du titre (04016041)	[stockwatch] + [stw]	[stockwatch] + [stw]
ECO	économie (général) (04017000)		[brf]
EDU	Education (05000000)	[Education]	[Education]
EDU	Formation permanente (05001000)	[Formation]	[Formation]
EDU	Parents d'élèves (05003000)	[Parents]	[Parents]
EDU	écoles, collèges, lycées (05005000)	[Ecole] + [Ecoles] + [école] + [écoles]	[Collèges] + [Ecole] + [Ecoles] + [Lycées] + [baccalauréat]
EDU	Syndicats d'enseignants (05006000)	[Enseignants]	[Enseignants]
EDU	Université (05007000)	[Université] + [Universités] + [étudiants]	[Etudiants] + [Université] + [Universités]
EDU	Pédagogie (05010000)	[Enseignement]	[Enseignement]
ENV	Environnement (06000000)	[Environnement] + [env]	[env] + [environnement] + [ozone]
ENV	Espèce menacée (06002001)		[environnement-baleine] + [environnement-baleines]
ENV	Pollution de l'air (06005001)	[pollution air]	[air-pollution]
ENV	Forêts (06006003)	[forêt]	[forêt] + [forêts]
ENV	Océans (06006007)	[océans]	[mer-pollution]
ENV	Nature (06007000)	[Nature]	[Nature]
ENV	Déchets (06009000)	[Déchets]	[Déchets]
ENV	Distribution de l'eau (06010000)	[eau]	[eau]
ENV	Réchauffement climatique (06011000)	[réchauffement]	[réchauffement]
HTH	Santé (07000000)	[Santé]	[san] + [santé]
HTH	Maladies (07001000)	[Maladie] + [Maladies]	[Maladie] + [Maladies]
HTH	Maladies contagieuses (07001001)	[grippe]	[grippe]
HTH	SIDA (07001003)	[Sida]	[Sida]
HTH	Cancer (07001004)	[Cancer]	[Cancer]
HTH	Maladie cardiaque (07001005)	[cardiologie]	[cardiologie]
HTH	Alzheimer (07001006)	[Alzheimer]	[Alzheimer]
HTH	Maladie animale (07001007)	[maladie animale]	[santé-élevage] + [tremblante] + [vache-folle]
HTH	épidémie (07002000)	[Epidémie] + [Epidémies] + [épidémie] + [épidémies]	[Epidémie] + [Epidémies] + [choléra]
HTH	Traitements (07003000)	[Médecine]	[Médecine]
HTH	Organisations de santé (07004000)	[Organisations de santé]	[OMS]
HTH	Recherche médicale (07005000)	[recherche médicale]	[santé-recherche]
HTH	Personnel médical (07006000)	[Personnel médical]	[infirmier] + [infirmiers] + [infirmière] + [infirmières] + [médecin] + [médecins] + [sages-femmes]
HTH	Médicaments (07007000)	[Médicament] + [Médicaments] + [Traitements]	[Médicament] + [Médicaments] + [Traitements]
HTH	Vaccins (07008001)	[Vaccin] + [Vaccins]	[Vaccin] + [Vaccins]
HTH	Hôpitaux et cliniques (07010000)	[Clinique] + [Cliniques] + [Hopital] + [Hopitaux] + [Hôpital] + [Hôpitaux]	[Clinique] + [Cliniques] + [Hôpital] + [Hôpitaux]
HTH	Assurances médicales (07012000)	[mutuelles]	[mutuelles]
HTH	Obésité (07017003)	[obésité]	[obésité]

Figure A.2 : Table des slugs : catégories IPTC 04 (suite) à 07

HUM	Gens animaux insolite (08000000)	[Gens, animaux et insolite]	
HUM	Animaux (08001000)	[Animaux]	[Animaux] + [animal] + [chien] + [chiens]
HUM	Insolite (08002000)	[Insolite]	[Insolite]
HUM	Gens (08003000)	[Gens] + [people]	[Gens] + [people]
HUM	Célébrités (08003002)	[Célébrités]	[célébrités]
HUM	ésotérisme (08004000)	[Esotérisme] + [Esotérismes]	[Esotérisme] + [Esotérismes]
HUM	Prix et récompenses (08006000)	[Récompenses]	[Nobel] + [Oscars] + [Récompenses]
HUM	Royauté (08007000)	[Royauté]	[Royauté]
NTS	Note (08900000)	[note]	[note]
GEN	General (08999000)		[actualité] + [matin] + [revue-presse]
SOC	Social (09000000)	[Social]	[dos] + [social] + [travail]
SOC	Apprentissage (09001000)	[Apprentissage]	[apprentissage]
SOC	Convention collective (09002000)	[convention collective] + [conventions collectives]	[conventions]
SOC	Emploi (09003000)	[Emploi]	[Emploi]
SOC	Travail des enfants (09003003)		[travail-enfants]
SOC	Conditions de travail (09004000)	[Conflit social]	[Conflit-social]
SOC	Législation du travail (09005000)	[Droit travail]	[Droit-travail]
SOC	Retraite (09006000)	[Retraite] + [Retraites]	[Retraite] + [Retraites]
SOC	Reconversion (09007000)	[Reconversion]	[Reconversion]
SOC	Grèves (09008000)	[Grèves] + [grève]	[Grèves] + [grève]
SOC	Chômage (09009000)	[Chômage]	[Chômage]
SOC	Syndicats (09010000)	[Syndicats] + [syndicat]	[Syndicats] + [cfdt] + [cftc] + [cgt] + [fo] + [syndicat]
SOC	Salaires et pensions (09011000)	[Salaires] + [pensions]	[Salaires] + [pensions]
SOC	Relations du travail (09012000)	[relations travail]	[relations-travail] + [soc]
SOC	Sécurité, maladies professionnelles (09013000)	[Hygiène travail] + [sécurité travail]	[Hygiène-travail] + [sécurité-travail]
SOC	Employeurs (09015000)	[Patronat] + [Patrons]	[Medef] + [Patronat] + [Patrons]
SOC	Personnel (09016000)	[salariés]	[salariés]
LIF	Vie quotidienne et loisirs (10000000)	[Loisirs]	
LIF	Jeux (10001000)	[Jeux]	[Jeux]
LIF	échecs (10001002)	[Echecs]	[Echecs]
LIF	Jeux de hasard et loteries (10002000)	[Loterie] + [Loteries]	[Loterie] + [Loteries] + [loto]
LIF	Gastronomie (10003000)	[Gastronomie]	[Gastronomie] + [Vin]
LIF	Vacances (10005000)	[Vacances]	[Vacances]
LIF	Point circulation (10007001)	[Circulation]	[Circulation] + [Radar] + [radars]
LIF	Loisirs (général) (10010000)	[Loisirs]	[Loisirs]
LIF	Chasse (10012000)	[Chasse]	[Chasse]
POL	Politique (11000000)	[Politique]	[Politique] + [pol]
POL	Défense (11001000)	[Défense]	[Défense]
POL	Anciens combattants (11001001)	[anciens combattants]	[anciens-combattants]
POL	Sécurité nationale (11001002)	[Sécurité nationale]	[Sécurité-nationale]
POL	Forces armées (11001004)		[armée]
POL	Systèmes de missiles (11001008)	[missile] + [missiles]	[missile] + [missiles]
POL	Arme nucléaire (11001009)		[armement-nucléaire] + [nucléaire-armement]
POL	Diplomatie (11002000)	[Diplomatie]	[diplomatie]
POL	Sommet (11002001)	[Sommet]	[G20] + [G7] + [G8] + [G90] + [Sommet]
POL	élections (11003000)	[Elections] + [élections]	[Election] + [Elections]
POL	Candidats (11003001)		[législatives-bio]
POL	Elections nationales (11003004)		[Législatives] + [Présidentielle]
POL	élections locales (11003006)		[cantonales] + [municipales]
POL	Sondage (11003008)		[sondage] + [sondages]
POL	Elections européennes (11003009)		[UE-Elections]
POL	Espionnage (11004000)	[Espionnage]	[Espionnage]
POL	Aide internationale (11005000)		[aide]
POL	Sanction économique (11005001)		[sanctions]
POL	Gouvernement (11006000)	[Gouvernement]	[conseil-ministres] + [gouvernement] + [gov]
POL	Service public (11006001)	[Administration] + [fonction publique]	[Administration] + [Fonctionnaires] + [fonction-publique] + [services-publics]
POL	Sécurité publique (11006002)		[sécurité-route]
POL	Ministres (11006009)		[ministre] + [ministres]
POL	Nationalisation (11006012)		[nationalisation]
POL	Droits de l'homme (11007000)	[DroitsHomme]	[DroitsHomme]
POL	Institutions locales (11008000)	[Collectivités]	[Collectivités] + [Décentralisation] + [municipalité]
POL	Parlement (11009000)	[Parlement]	[Loi] + [Parlement] + [assemblée] + [douma] + [knesset] + [sénat] + [usa-congrès]
POL	Partis politiques (11010000)	[Partis]	[Partis] + [opposition]
POL	ONG (11010001)	[ONG]	[ONG]
POL	Réfugiés (11011000)	[Réfugiés]	[HCR] + [Réfugiés]
POL	Budget de l'Etat et impôts (11013000)	[Bud] + [Budget]	[Bud] + [Budget] + [Fiscalité]
POL	Traités et organisations internationales (11014000)	[Traités]	[AIEA] + [ASEAN] + [Alena] + [OEA] + [ONU] + [Traités] + [UA] + [UE] + [UE-25] + [UE-candidats]
POL	Relations internationales (11014001)	[Relations internationales]	[Désarmement]
POL	Négociations de paix (11014002)	[Paix]	[Paix]
POL	Alliances (11014003)	[Alliances]	[OTAN]

REL	Religion et croyance (12000000)	[Religions et croyances]	[religion]
REL	Cultes et sectes (12001000)	[Secte] + [Sectes]	[Secte] + [Sectes]
REL	Croyances (12002000)	[Croyance] + [Croyances]	[Croyance] + [Croyances] + [foi]
REL	Scientologie (12002002)	[Scientologie]	[Scientologie]
REL	Franc-maçonnerie (12003000)	[Franc-maçonnerie]	[Franc-maçonnerie]
REL	Valeurs (12006000)	[valeurs]	[Laïcité]
REL	Relations entre l'église et l'Etat (12007000)	[Eglises] + [religion]	[vatican]
REL	Christianisme (12009000)	[Christianisme]	[Christianisme] + [Chrétiens]
REL	Protestantisme (12009001)	[Protestantisme]	[Protestant] + [Protestantisme] + [Protestants]
REL	Anglicanisme (12009004)	[Anglicanisme]	[Anglican] + [Anglicanisme] + [Anglicans]
REL	Mormon (12009009)	[Mormon]	[Mormon]
REL	Catholicisme romain (12009010)	[Catholicisme romain]	[catholicisme] + [catholique] + [catholiques] + [évêques]
REL	Orthodoxie (12009012)	[Orthodoxie]	[Orthodoxe] + [Orthodoxes] + [Orthodoxie]
REL	Islam (12010000)	[Islam] + [Islam-voile] + [Musulman]	[Islam] + [Islam-voile] + [Musulman]
REL	Judaïsme (12011000)	[Judaïsme]	[Judaïsme] + [juif] + [juifs]
REL	Bouddhisme (12012000)	[Bouddhisme]	[Bouddhisme] + [Bouddhiste] + [Bouddhistes]
REL	Hindouisme (12013000)	[Hindouisme]	[Hindouisme] + [Hindouiste] + [Hindouistes]
REL	Fête religieuse (12014000)	[Fête religieuse]	[religion-fête]
REL	Ramadan (12014004)	[Ramadan]	[Ramadan]
REL	Yom Kippour (12014005)	[Yom Kippour]	[kippour]
REL	Pape (12015001)	[pape]	[pape]
REL	Bible (12023001)	[bible]	[bible]
REL	Coran (12023002)	[coran]	[coran]
REL	Torah (12023003)	[torah]	[torah]
REL	Oecuménisme (12027000)	[oecuménisme]	[oecuménisme]

FIGURE A.3 : Table des slugs : catégories IPTC 08 à 12

SCI	Science et technologie (13000000)	[Sciences et technologies]	
SCI	Physique (13001001)	[physique]	[physique]
SCI	Chimie (13001002)		[Nobel-chimie]
SCI	Histoire (13003002)	[Histoire]	[Histoire]
SCI	Psychologie (13003003)	[Psychologie]	[Psychologie]
SCI	Sociologie (13003004)	[Sociologie]	[Sociologie]
SCI	Anthropologie (13003005)	[Anthropologie]	[Anthropologie]
SCI	Sciences naturelles (13004000)	[Sciences naturelles]	
SCI	Géologie (13004001)	[Géologie]	[Géologie]
SCI	Paléontologie (13004002)	[Paléontologie]	[Paléontologie]
SCI	Botanique (13004004)	[Botanique]	[Botanique]
SCI	Zoologie (13004005)	[zoologie]	[zoologie]
SCI	Astronomie (13004007)	[astronomie]	[astronomie]
SCI	Biologie (13004008)	[biologie]	[biologie]
SCI	Recherche (13006000)	[Recherche]	[Recherche]
SCI	Exploration scientifique (13007000)	[Exploration]	[Exploration]
SCI	Programmes spatiaux (13008000)	[Espace]	[Espace]
SCI	Science (général) (13009000)	[Sciences]	[Sciences] + [science]
SCI	Technologie (général) (13010000)	[Technologie]	[Technologie]
SCI	Océanographie (13014000)	[Océanographie]	[Océanographie]
SCI	Climatologie (13015000)	[Climat]	[Climat]
SCI	Technologie de l'identification (13017000)	[biométrie]	[biométrie]
SCI	Mathématique (13018000)	[maths] + [mathématiques]	[maths] + [mathématiques]
SCI	Biotechnologie (13019000)	[Biotechnologie]	[OGM] + [clonage] + [génétique]
SCI	Nanotechnologie (13021000)	[Nanotechnologies]	[Nanotechnologie] + [Nanotechnologies]
SOI	Société (14000000)	[Société]	[société]
SOI	Toxicomanie (14001000)	[Toxicomanie]	[Toxicomanie]
SOI	Démographie (14003000)	[Démographie]	[Démographie]
SOI	Recensement (14003001)	[Population] + [Recensement]	[Population] + [Recensement]
SOI	Immigration (14003002)	[Immigration]	[Immigration]
SOI	Handicapés (14004000)	[Handicapés]	[Handicap] + [Handicapés]
SOI	Euthanasie (14005000)	[Euthanasie]	[euthanasie]
SOI	Suicide (14005001)	[suicide]	[suicide]
SOI	Famille (14006000)	[Famille]	[Famille]
SOI	Adoption (14006002)	[adoption]	[adoption]
SOI	Sexualité (14006005)	[Sexe] + [Sexualité]	[Sexe] + [Sexualité]
SOI	Contrôle des naissances (14007000)	[Contrôle des naissances]	[contraception]
SOI	Assurance maladie (14008000)	[Assurance maladie] + [Sécu]	[Sécu]
SOI	Sans-abri (14009000)	[Sans abri]	[SDF]
SOI	Minorités (14010000)	[Minorités]	[Minorités]
SOI	Homosexualité (14010001)	[gay] + [homosexualité]	[gays] + [homosexualité] + [homosexuels] + [lesbiennes]
SOI	Pornographie (14011000)	[Pornographie]	[Pornographie]
SOI	Pauvreté (14012000)	[Pauvreté]	[Pauvreté]
SOI	Prostitution (14013000)	[Prostitution]	[Prostitution]
SOI	Racisme (14014000)	[Racisme]	[Racisme] + [antisémitisme] + [xénophobie]
SOI	Avortement (14016000)	[Avortement]	[Avortement] + [ivg]
SOI	Personnes disparues (14017000)	[Disparition] + [Disparitions]	[Disparition] + [Disparitions]
SOI	Personnes disparues en temps de guerre (14017001)	[Disparus]	[Disparus]
SOI	Délinquance juvénile (14019000)	[délinquance]	[délinquance]
SOI	Esclavage (14021000)	[esclavage]	[esclavage]
SOI	Enfants (14024001)	[Enfance] + [Enfant] + [Enfants]	[Enfance] + [Enfant] + [Enfants] + [UNICEF]
SOI	Adolescent (14024003)	[Adolescence]	[Adolescent] + [Adolescents]
SOI	Personnes âgées (14024005)	[sénior] + [séniors]	[sénior] + [séniors] + [vieillesse]
SOI	Questions de société (14025000)	[Femmes]	[Femmes]
SOI	Discrimination (14025003)	[Discrimination]	[Discrimination]
SPO	Sport (15000000)	[Sport]	[spo] + [sport]
SPO	Dopage (15000027)	[Dopage] + [dopage]	[dopage]
SPO	Handisport (15000030)	[handisport]	[handisport]
SPO	Parachutisme (15001001)	[Parachutisme]	[Parachutisme]
SPO	Ski alpin (15002000)	[Ski alpin]	[Ski-alpin] + [ski]
SPO	Football américain (15003000)	[Football américain]	[Football-américain] + [foot-am]
SPO	Tir à l'arc (15004000)	[Tir arc]	[tir-arc]
SPO	Athlétisme (15005000)	[Athlétisme]	[Athlétisme]
SPO	Badminton (15006000)	[Badminton]	[Badminton]
SPO	Base-ball (15007000)	[Baseball]	[Base-ball]
SPO	Basket-ball (15008000)	[Basket]	[Basket]
SPO	Biathlon (15009000)	[Biathlon]	[Biathlon]
SPO	Billards, snooker et pool (15010000)	[Billard] + [Billard, snooker et pool] + [pool] + [snooker]	[Billard] + [Billard, snooker et pool] + [pool] + [snooker]
SPO	Bobsleigh (15011000)	[Bobsleigh]	[Bobsleigh]
SPO	Bowling (15012000)	[Bowling]	[Bowling]
SPO	Boule anglaise et pétanque (15013000)	[pétanque]	[boules] + [pétanque]
SPO	Boxe (15014000)	[Boxe]	[Boxe]
SPO	Canoë-kayak (15015000)	[Canoë kayak]	[Canoë-kayak]
SPO	Alpinisme et escalade (15016000)	[Alpinisme]	[Alpinisme]
SPO	Cricket (15017000)	[Cricket]	[Cricket]
SPO	Curling (15018000)	[Curling]	[Curling]
SPO	Cyclisme (15019000)	[Cyclisme]	[Cyclisme]
SPO	Piste (15019001)	[Piste]	[cyclisme-piste]
SPO	Course à étapes (15019013)		[cyclisme-Giro] + [cyclisme-TDF] + [cyclisme-vuelta]
SPO	VTT (15019015)	[VTT]	[Cyclisme-VTT]
SPO	Plongeon (15021000)	[Plongeon]	[Plongeon]
SPO	Equitation (15022000)	[Equitation]	[Equitation]
SPO	Escrime (15023000)	[Escrime]	[Escrime]
SPO	Hockey sur gazon (15024000)	[Hockey]	[Hockey-gazon]
SPO	Patinage artistique (15025000)	[Glace Patinage]	[Patinage-artistique]
SPO	Ski acrobatique (15026000)	[Ski acrobatique]	[Ski-acrobatique] + [Ski-freestyle]
SPO	Golf (15027000)	[Golf]	[Golf]
SPO	Gymnastique (15028000)	[Gymnastique]	[Gymnastique]
SPO	Gymnastique rythmique (15028009)	[rythmique]	[Gymnastique-rythmique]

Figure A.4 : Table des slugs : catégories IPTC 13 à 15

SPO	Handball (15029000)	[Hand]	[Hand]
SPO	Hippisme (15030000)	[Hippisme]	[Hippisme]
SPO	Hockey sur glace (15031000)	[Hockey Glace]	[Glace-Hockey] + [Hockey-Glace]
SPO	Pelote basque (15032000)	[Pelote basque]	[Pelote-basque]
SPO	Judo (15033000)	[Judo]	[Judo]
SPO	Karaté (15034000)	[Karaté]	[Karaté]
SPO	Lacrosse (15035000)	[Crosse]	[Crosse]
SPO	Luge (15036000)	[Luge]	[Luge]
SPO	Marathon (15037000)	[Marathon]	[Marathon]
SPO	Pentathlon moderne (15038000)	[Penthatlon]	[Pentathlon]
SPO	Automobile (15039000)	[Auto]	[Auto]
SPO	F1 (15039001)	[auto F1]	[auto-F1]
SPO	A1 (15039002)	[A1]	[Auto-A1]
SPO	Course d'endurance (15039003)	[endurance]	[Auto-endurance]
SPO	Formule Indy (15039004)	[Indy]	[Auto-Indy]
SPO	CART (15039005)	[CART]	[Auto-CART]
SPO	Rallyes (15040000)	[auto rallye] + [moto rallye]	[auto-rallye] + [moto-rallye]
SPO	Rallyes-raid (15040002)	[Auto Moto]	[Auto-Moto]
SPO	Moto (15041000)	[Moto]	[Moto]
SPO	Grand-Prix de vitesse (15041001)	[Moto vitesse]	[Moto-vitesse]
SPO	enduro (15041002)	[moto enduro]	[moto-enduro]
SPO	Endurance (15041008)	[Moto endurance]	[Moto-endurance]
SPO	Netball (15042000)	[Netball]	[Netball]
SPO	Ski nordique (15043000)	[Ski nordique]	[Ski-nordique]
SPO	Course de fond (15043001)	[fond]	[ski-fond]
SPO	Combiné nordique (15043013)	[combiné]	[combiné] + [ski-combiné]
SPO	Course d'orientation (15044000)	[orientation]	[orientation]
SPO	Polo (15045000)	[Polo]	[Polo]
SPO	Motonautisme (15046000)	[Motonautisme]	[Motonautisme]
SPO	Aviron (15047000)	[Aviron]	[Aviron]
SPO	Jeu à XIII (15048000)	[Jeu XIII]	[Rugby-XIII]
SPO	Rugby à XV (15049000)	[Rugby]	[Rugby]
SPO	Rugby à VII (15049001)		[Rugby-VII]
SPO	Voile (15050000)	[Voile]	[Voile]
SPO	Tir (15051000)	[Tir]	[Tir]
SPO	Saut à skis (15052000)	[Ski Saut]	[ski-saut]
SPO	Surf des neiges (15053000)	[Snowboard]	[ski-snowboard]
SPO	Football (15054000)	[Foot]	[Foot]
SPO	Softball (15055000)	[Softball]	[Softball]
SPO	Patinage de vitesse (15056000)	[Patinage vitesse]	[Patinage-vitesse]
SPO	Short-track (15056007)		[short-track]
SPO	CIO (15058001)	[CIO]	[CIO]
SPO	AGFIS (15058005)	[AGFIS]	[AGFIS]
SPO	Squash (15059000)	[Squash]	[Squash]
SPO	Sumo (15060000)	[Sumo]	[Sumo]
SPO	Surf (15061000)	[Surf]	[Surf]
SPO	Natation (15062000)	[Natation]	[Natation]
SPO	Natation synchronisée (15062025)	[synchronisée]	[natation-synchronisée]
SPO	Tennis de table (15063000)	[Tennis table]	[Tennis-table]
SPO	Taekwondo (15064000)	[Taekwondo]	[Taekwondo]
SPO	Tennis (15065000)	[Tennis]	[Tennis]
SPO	Triathlon (15066000)	[Triathlon]	[Triathlon]
SPO	Volley-ball (15067000)	[Volley] + [Volley ball]	[Volley] + [Volley ball]
SPO	Beach volley (15067001)	[beach volley]	[beach-volley]
SPO	Water polo (15068000)	[Water polo]	[Water-polo]
SPO	Ski nautique (15069000)	[Ski nautique]	[Ski-nautique]
SPO	Haltérophilie (15070000)	[Haltérophilie]	[Haltérophilie]
SPO	Lutte (15072000)	[Lutte]	[Lutte]
SPO	Evénements sportifs (15073000)	[JO] + [Universiade]	[JO] + [Universiade]
SPO	JO d'été (15073001)	[JO été]	[JO-2008] + [JO-2012]
SPO	JO d'hiver (15073002)		[JO-2006] + [JO-2010] + [JO-2014]
SPO	Jeux paralympiques (15073047)	[Paralympiques]	[Paralympiques]
SPO	Rodéo (15074000)	[Rodeo]	[Rodeo]
SPO	Mini golf sport (15075000)	[minigolf]	[minigolf]
SPO	Skeleton (15083000)		[Skeleton]
UNR	Guerres et conflits (16000000)	[Guerres et conflits]	
UNR	Actes de terrorisme (16001000)	[Attentat] + [Terrorisme]	[Attentat] + [Attentats] + [Terrorisme] + [attaque] + [attaques]
UNR	Conflits armés (16002000)	[combats]	[combats]
UNR	Désordres civils (16003000)	[Désordres civils]	[Violences] + [troubles]
UNR	Rébellion (16003002)	[Rébellions]	[Rébellion] + [soulèvement]
UNR	Dissidence (16003003)	[dissidents]	[dissidents]
UNR	Coup d'état (16004000)	[Coup]	[Coup]
UNR	Guérilla (16005000)	[Guérilla]	[Guérilla]
UNR	Explosion de bombes (16005002)	[bombardements]	[bombardements]
UNR	Massacre (16006000)	[Massacre]	[Massacre]
UNR	Génocide (16006001)	[Génocide]	[Génocide]
UNR	émeutes, affrontements (16007000)	[Affrontements] + [Emeutes]	[Affrontements] + [Emeutes]
UNR	Manifestations (16008000)	[Manifestation] + [Manifestations]	[Manifestation] + [Manifestations]
UNR	Guerre (16009000)	[Guerre]	[Guerre]
UNR	Prisonniers et détenus (16009003)	[prisonniers]	[prisonniers]
UNR	Conflits (général) (16010000)	[Conflit] + [Conflits]	[Conflit] + [Conflits]
UNR	Crise (16011000)	[Crise]	
UNR	Armement (16012000)	[armes] + [déminage]	[armes] + [déminage]
WEA	Météo (17000000)	[Météo]	
WEA	Bulletins (17003000)		[météo]

FIGURE A.5 : Table des slugs : catégories IPTC 15 (suite) à 17

Table A.1 : Correspondances entre slugs AFP et catégories Wikipedia normalisées

Slugs AFP	**Catégories Wikipedia normalisées**	**Slug AFP**	**Catégories Wikipedia normalisées**
accident	accident	immigration	immigration
accusation	accusation	immobilier	immobilier
acquisition	acquisition	incendie	incendie
actualité	actualité	indicateur	indicateur
administration	administration	industrie	industrie
adolescence	adolescence	infirmière	infirmière
adoption	adoption	inflation	inflation
agriculture	agriculture	informatique	informatique
aide	aide	inondation	inondation
alimentation	alimentation	internet	internet
alpinisme	alpinisme	islam	islam, islamologue, islamologie
anglican	anglican	jeux	jeux
anglicanisme	anglicanisme	judaïsme	judaïsme
animal	animal	judo	judo
animaux	animaux	juif	juif
anthropologie	anthropologie	juifs	juifs
antisémitisme	antisémitisme	justice	justice
apprentissage	apprentissage	karaté	karaté
architecture	architecture	langage	langage
archéologie	archéologie	laïcité	laïcité
armes	armes	littérature	littérature
armée	armée	livre	livre-jeu, livre
art	art	livres	livres
arts	arts	logiciel	logiciel
assemblée	assemblée	logistique	logistique
assurance	assurance	loi	loi
astronomie	astronomie	luge	luge
athlétisme	athlétisme	lutte	lutteur, lutte, lutteuse
attentat	attentat	luxe	luxe
attentats	attentats	mafia	mafia
audiovisuel	audiovisuel	maladie	maladie
audit	audit	maladies	maladies
automobile	automobile	manifestation	manifestation
avalanche	avalanche	marathon	marathon
aviron	aviron	marketing	marketing
avortement	avortement	massacre	massacre
aéronautique	aéronautique	mathématiques	mathématiques
baccalauréat	baccalauréat	matin	matinale
badminton	badminton	meurtre	meurtre
ban	ban	migration	migration
banque	banque	mine	mine
banque-centrale	banque-centrale	mines	mines
baseball	baseball	ministre	ministre
beach-volley	beach-volley	ministres	ministres
biathlon	biathlon	missile	missile
bible	bible	mode	mode
billard	billard	motonautisme	motonautisme
biologie	biologie	municipalité	municipalité
biométrie	biométrie	musique	musique
biotechnologie	biotechnologie	musulman	musulman
bobsleigh	bobsleigh	musée	musée-ou-galerie-photographique, musée
bombardements	bombardements	mécanique	mécanique
botanique	botanique	médecin	médecin
bouddhisme	bouddhisme	médecine	médecine
bouddhiste	bouddhiste	médecins	médecins

TABLE A.1 – *(suite)*

Slugs AFP	**Catégories Wikipedia normalisées**	**Slug AFP**	**Catégories Wikipedia normalisées**
bourse	bourse	média	média-participations, média-sans-publicité, média
bowling	bowling	médias	médias
boxe	boxe	médicament	médicament
brevet	brevet	métallurgie	métallurgie
btp	btp	météo	météorologie, météo-france, météoroloque, météorologue
cancer	cancer	nanotechnologie	nanotechnologie
canicule	canicule	natation	natation
canoë-kayak	canoë-kayak	nature	nature
cap	cap	naufrage	naufrage
cardiologie	cardiologie	netball	netball
carnaval	carnaval	notation	notation
casino	casino	note	note
catastrophe	catastrophe	obésité	obésité
catastrophes	catastrophes	océanographie	océanographie
catholicisme	catholicisme	omc	omc
catholique	catholique	ong	ong
cgt	cgt	opéra	opéra-ballet, opéra, opéra-comique
chantier-naval	chantier-naval	orientation	orientation
charbon	charbon	orthodoxe	orthodoxe
chasse	chasse	orthodoxie	orthodoxie
chien	chien	oscars	oscars
chimie	chimie	otage	otage
choléra	choléra	ouragan	ouragan
christianisme	christianisme	paix	paix
chrétiens	chrétiens	paléontologie	paléontologie
chômage	chômage	pape	pape
ciment	ciment	parachutisme	parachutisme
cinéma	cinéma	parlement	parlement
circulation	circulation	partis	partis
climat	climat	patrimoine	patrimoine
clinique	clinique	pauvreté	pauvreté
clonage	clonage	peinture	peinture
coentreprise	coentreprise	pentathlon	pentathlon
combiné	combiné	pharmacie	pharmacie
commande	commande	photo	photographe, photo-guide, photokinésiste, photo-journaliste, photométrie, photographe-plasticien, photojournalisme
commerce	commerce	photographie	photographie
commémoration	commémoration	physique	physique
comptabilité	comptabilité	piraterie	piraterie
conflit	conflit	piste	piste
conflits	conflits	plongeon	plongeon
consommation	consommation	police	police
constitution	constitution	politique	politique
construction	construction	pollution	pollution
contentieux	contentieux	polo	polo
contraception	contraception	pornographie	pornographie
contrat	contrat	port	port
coran	coran	poste	poste
corruption	corruption	prison	prison
coup	coup	procès	procès

Table A.1 – *(suite)*

Slugs AFP	**Catégories Wikipedia normalisées**	**Slug AFP**	**Catégories Wikipedia normalisées**
courtage	courtage	prostitution	prostitution
cricket	cricket	protestant	protestant
criminalité	criminalité	protestantisme	protestantisme
crise	crise	psychologie	psychologie
croissance	croissance	publicité	publicité
crosse	crosse	pédophilie	pédophilie
croyance	croyance	pétrole	pétrole
croyances	croyances	pêche	pêche
culture	culture	racisme	racisme
curling	curling	radar	radar
cyclisme	cyclisme	radio	radiotélescope, radiogalaxie, radiophonie, radiodiffusion, radiofréquence, radiotéléphonie, radiomessagerie, radionavigation, radio-française, radioamateurisme, radio, radiocommunications
cyclone	cyclone	recherche	recherche
danse	danse, danseuse, danseur	religion	religion
diamant	diamant	restauration	restauration
diplomatie	diplomatie	retraite	retraite
direction	direction	royauté	royauté
dirigeant	dirigeant	rugby	rugby, rugbyman
discrimination	discrimination	rébellion	rébellion
distribution	distribution	santé	santé
dopage	dopage	satellite	satellite
dos	dos	science	science
drogue	drogue	science-fiction	science-fiction
droits	droits	sciences	sciences
défense	défense	scientologie	scientologie
délinquance	délinquance	sculpture	sculpture
démographie	démographie	secte	secte
désarmement	désarmement	services	services
eau	eau	sexualité	sexualité
education	education	short-track	short-track
emballage	emballage	sida	sida
endurance	endurance	sidérurgie	sidérurgie
enfance	enfance	skeleton	skeleton
enfant	enfant	ski	ski
enlèvement	enlèvement	snooker	snooker
enseignement	enseignement	snowboard	snowboard
entreprise	entreprise	sociologie	sociologie
environnement	environnement	société	société
esclavage	esclavage	softball	softball
escrime	escrime	sommet	sommet
espace	espace	spectacle	spectacle
espionnage	espionnage	sport	sportif, sports, sportive, sport
euthanasie	euthanasie	suicide	suicide
exploration	exploration	sumo	sumo
explosion	explosion	surf	surf
exposition	exposition	syndicat	syndicat
fai	fai	séisme	séisme
famille	famille	sénat	sénatrice, sénateurs, sénateur, sénat
famine	famine	taekwondo	taekwondo
festival	festival	tauromachie	tauromachie

TABLE A.1 – *(suite)*

Slugs AFP	**Catégories Wikipedia normalisées**	**Slug AFP**	**Catégories Wikipedia normalisées**
finance	finance	taux	taux
fiscalité	fiscalité	technologie	technologie
fond	fond	tennis	tennis
foot	foot	terrorisme	terrorisme
formation	formation	tex	tex
forêt	forêt	textile	textile
franc-maçonnerie	franc-maçonnerie	théâtre	théâtre
fus	fus	tir	tir
fusion	fusion	tourisme	tourisme
g20	g20	tradition	tradition
gastronomie	gastronomie	transport	transport
gay	gay	transports	transports
gaz	gaz	travail	travail
gendarmerie	gendarmerie	triathlon	triathlon
gens	gens	troubles	troubles
golf	golf	tsunami	tsunami
gouvernement	gouvernement	typhon	typhon
grippe	grippe	télécommunications	télécommunications
grève	grève	télévision	télévision
guerre	guerre	universiade	universiade
guérilla	guérilla	université	université
gymnastique	gymnastique	vaccin	vaccin
génocide	génocide	vatican	vatican
génétique	génétique	vieillesse	vieillesse
géologie	géologie	ville	ville, villes
haltérophilie	haltérophilie	vin	vin
handicap	handicap	viol	viol
handisport	handisport	violences	violences
hindouisme	hindouisme	viticulture	viticulture
hindouiste	hindouiste	voile	voile
histoire	histoire	vol	vol
hockey	hockey	volcan	volcan
homicide	homicide	vtt	vtt
homosexualité	homosexualité	water-polo	water-polo
hôpital	hôpital	xénophobie	xénophobie
hôtellerie	hôtellerie	zoologie	zoologie

2 Dépêches et format NewsML

```
<?xml version="1.0" encoding="utf-8"?>
<NewsML Version="1.2">
<!--AFP NewsML text-photo profile evolution2 -->
<!--Processed by Xafp1-4ToNewsML1-2 rev18 -->
<Catalog Href="http://www.afp.com/dtd/AFPCatalog.xml"/>
<NewsEnvelope>
<TransmissionId>2229</TransmissionId>
<DateAndTime>20130102T210058Z</DateAndTime>
<NewsProduct FormalName="FRA"/>
<NewsProduct FormalName="FRS"/>
<NewsProduct FormalName="DAGI"/>
<NewsProduct FormalName="DPSE"/>
<Priority FormalName="4"/>
</NewsEnvelope>
<NewsItem xml:lang="fr">
<Identification>
<NewsIdentifier>
<ProviderId>afp.com</ProviderId>
<DateId>20130102T210051Z</DateId>
<NewsItemId>TX-PAR-IAN24</NewsItemId>
<RevisionId PreviousRevision="0" Update="N">1</RevisionId>
<PublicIdentifier>urn:newsml:afp.com:20130102T210051Z:TX-PAR-IAN24:1</Pu
blicIdentifier>
</NewsIdentifier>
<NameLabel>Syrie-conflit</NameLabel>
</Identification>
<NewsManagement>
<NewsItemType FormalName="News"/>
<FirstCreated>20130102T210050Z</FirstCreated>
<ThisRevisionCreated>20130102T210050Z</ThisRevisionCreated>
<Status FormalName="Usable"/>
<Urgency FormalName="4"/>
<DerivedFrom NewsItem="urn:newsml:afp.com:20130102T182013Z:TX-NIC-NYX49"/>
<AssociatedWith FormalName="Graphic"/>
<AssociatedWith FormalName="Video"/>
<AssociatedWith FormalName="PHOTOARCH"/>
<AssociatedWith FormalName="VIDEOARCH"/>
</NewsManagement>
<NewsComponent>
<NewsLines>
<DateLine xml:lang="fr">DAMAS, 2 jan 2013 (AFP) - </DateLine>
<HeadLine xml:lang="fr">Plus de 60.000 morts en 21 mois de conflit en
Syrie selon l'ONU </HeadLine>
<NewsLine>
<NewsLineType FormalName="AdvisoryLine"/>
<NewsLineText xml:lang="fr">Ajoute bilan</NewsLineText>
```

Figure A.6 : Dépêche AFP 20130102T210051Z-TX-PAR-IAN24 (1/4)

```
</NewsLine>
<NewsLine>
<NewsLineType FormalName="ProductLine"/>
<NewsLineText xml:lang="fr">=(Infographie+Video+Photo Archives+Video
archives)=</NewsLineText>
</NewsLine>
</NewsLines>
<AdministrativeMetadata>
<Provider>
<Party FormalName="AFP"/>
</Provider>
</AdministrativeMetadata>
<DescriptiveMetadata>
<Language FormalName="fr"/>
<Genre FormalName="Prev"/>
<Genre FormalName="Article"/>
<Genre FormalName="Update"/>
<SubjectCode>
<SubjectMatter FormalName="16010000"/>
</SubjectCode>
<OfInterestTo FormalName="MOA-TFG-1=MOA"/>
<OfInterestTo FormalName="FRE-TFG-1=ELU"/>
<OfInterestTo FormalName="DAB-TFG-1=DAB"/>
<OfInterestTo FormalName="AMN-TFG-1=AMW"/>
<OfInterestTo FormalName="EUA-TFG-1=EUA"/>
<OfInterestTo FormalName="FRE-TFG-5=FRSGL"/>
<DateLineDate>20130102T175118+0200</DateLineDate>
<Location HowPresent="Origin">
<Property FormalName="Country" Value="SYR"/>
<Property FormalName="City" Value="DAMAS"/>
</Location>
<Property FormalName="Keyword" Value="Syrie"/>
<Property FormalName="Keyword" Value="conflit"/>
<Property FormalName="GeneratorSoftware" Value="Cafp32"/>
</DescriptiveMetadata>
<ContentItem>
<MediaType FormalName="Text"/>
<Format FormalName="NITF3.1"/>
<Characteristics>
<SizeInBytes>4252</SizeInBytes>
<Property FormalName="Words" Value="708"/>
</Characteristics>
<DataContent>
<nitf>
<body>
<body.content>
```

FIGURE A.7 : Dépêche AFP 20130102T210051Z-TX-PAR-IAN24 (2/4)

```
<p>Les Nations unies ont annoncé mercredi la mort de plus de 60.000
personnes dans le conflit qui fait rage depuis 21 mois en Syrie où les
redoutables chasseurs-bombardiers du régime intensifiaient les raids
meurtriers et les combats ne connaissaient aucun répit. </p>
<p>Le Haut-Commissariat de l'ONU aux droits de l'Homme a affirmé à
Genève avoir recensé 59.648 personnes tuées en Syrie entre les premières
manifestations lancées dans le sillage du Printemps arabe le 15 mars
2011 et la fin novembre 2012. </p>
<p>"Etant donné que le conflit s'est poursuivi sans relâche depuis fin
novembre, nous pouvons supposer que plus de 60.000 personnes ont été
tuées jusqu'au début 2013", a dit la Haut-Commissaire Navi Pillay dans
un communiqué, estimant ce nombre "bien plus élevé qu'attendu et
réellement choquant". </p>
<p>"On assiste à une prolifération de crimes graves par les deux
parties, y compris des crimes de guerre et, très probablement, des
crimes contre l'humanité", a-t-elle dénoncé. </p>
<p>La Syrie a basculé dans la guerre civile après que cette révolte
populaire violemment réprimée par le régime se soit militarisée, et les
combats opposent désormais les soldats à des déserteurs aidés par des
civils ayant pris les armes mais aussi des jihadistes venus de
l'étranger.</p>
<p>Le bilan de l'ONU est supérieur à celui de l'Observatoire syrien des
droits de l'Homme (OSDH), une organisation qui s'appuie sur un large
réseau de militants et de médecins à travers le pays et qui chiffre le
nombre des morts à plus de 46.000 morts en 21 mois. </p>
<p>Mais l'OSDH ne recense ni les milliers de personnes disparues ou en
détention ni la plupart des morts parmi les "chabbihas", les miliciens
du régime de Bachar al-Assad, ni les combattants étrangers. </p>
<p>De plus "les rebelles et l'armée ne révèlent pas le nombre de morts
dans leurs rangs pour ne pas porter un coup au moral des troupes",
explique à l'AFP le directeur de l'OSDH, Rami Abdel Rahmane, qui estime
qu'en totalisant toutes ces catégories, le nombre de morts pourrait
dépasser les 100.000. </p>
<p/>
<hl2>Journaliste américain enlevé </hl2>
<p/>
<p>Alors que les journalistes continuent d'entrer clandestinement en
Syrie en raison des restrictions imposées par les autorités, le reporter
de guerre indépendant américain James Foley, 39 ans, qui a fourni des
reportages vidéo à l'AFP, a été enlevé, a annoncé sa famille qui est
sans nouvelle de lui. </p>
<p>Il a été arrêté, selon les témoignages recueillis par l'AFP, le 22
novembre par quatre hommes armés de Kalachnikov. </p>
<p>"Nous voulons que Jim revienne à la maison sain et sauf, ou au moins,
nous avons besoin de lui parler pour savoir qu'il va bien", a dit son
père, John Foley. "Jim est un journaliste sans parti pris et nous
```

Figure A.8 : Dépêche AFP 20130102T210051Z-TX-PAR-IAN24 (3/4)

```
appelons à sa libération". </p>
<p>Devant le blocage des efforts internationaux pour un règlement
politique, la Syrie a débuté l'année 2013 dans la violence avec
l'aviation, le principal atout du régime dans la guerre, lançant de
nouveaux raids sanglants. </p>
<p>Des dizaines de personnes, dont plusieurs rebelles, ont été tuées ou
blessées dans l'explosion d'une station d'essence touchée par un raid
aérien à Milha, une région bordant la capitale où les rebelles ont
installé leurs bases-arrières, selon l'OSDH. </p>
<p>L'ONG a précisé ne pas être en mesure de fournir dans l'immédiat un
bilan plus précis. Une vidéo mise en ligne par des militants a montré
des habitants pris de panique courant au milieu des flammes à la
recherche de survivants. </p>
<p>Sur d'autres images, un homme hurle en tenant dans ses bras un corps,
dont ne restent que la tête et une partie du torse ensanglantées. Le
corps d'un autre homme en feu tient encore sur une mobylette au milieu
de l'incendie. </p>
<p>Toujours dans la ceinture sud de la capitale, douze membres d'une
même famille, en majorité des enfants, ont été fauchés par une bombe
larguée par l'aviation à Mouadamiyat al-Cham, a dit l'OSDH. </p>
<p>Selon un bilan provisoire de l'ONG, 127 personnes ont péri mercredi
dans le pays, dont 60 civils et une vingtaine de rebelles qui ont été
tués dans des assauts contre des aéroports militaires et la base de Wadi
Deif (nord), aux mains de l'armée. Selon l'OSDH, un combattant
australien a péri près de cette base. </p>
<p>bur-sbh/tp</p>
</body.content>
</body>
</nitf>
</DataContent>
</ContentItem>
</NewsComponent>
</NewsItem>
</NewsML>
```

FIGURE A.9 : Dépêche AFP 20130102T210051Z-TX-PAR-IAN24 (4/4)

```
<?xml version="1.0" encoding="utf-8"?>
<NewsML Version="1.2">
  <!--AFP NewsML text-photo profile evolution2 -->
  <!--Processed by Xafp1-4ToNewsML1-2 rev18 -->
  <Catalog Href="http://www.afp.com/dtd/AFPCatalog.xml"/>
  <NewsEnvelope>
    <TransmissionId>3082</TransmissionId>
    <DateAndTime>20130110T175236Z</DateAndTime>
    <NewsService FormalName="DGTE"/>
    <NewsProduct FormalName="FRS"/>
    <NewsProduct FormalName="FIL"/>
    <NewsProduct FormalName="FRA"/>
    <Priority FormalName="4"/>
  </NewsEnvelope>
  <NewsItem xml:lang="fr">
    <Identification>
      <NewsIdentifier>
        <ProviderId>afp.com</ProviderId>
        <DateId>20130110T175231Z</DateId>
        <NewsItemId>TX-PAR-ITP91</NewsItemId>
        <RevisionId PreviousRevision="1" Update="N">2</RevisionId>

<PublicIdentifier>urn:newsml:afp.com:20130110T175231Z:TX-PAR-ITP91:2</Pu
blicIdentifier>
      </NewsIdentifier>
      <NameLabel>Social-emploi-syndicat-patronat-gouvernement</NameLabel>
    </Identification>
    <NewsManagement>
      <NewsItemType FormalName="News"/>
      <FirstCreated>20130110T175223+0000</FirstCreated>
      <ThisRevisionCreated>20130110T175223+0000</ThisRevisionCreated>
      <Status FormalName="Usable"/>
      <Urgency FormalName="4"/>
      <AssociatedWith
NewsItem="urn:newsml:urn:newsml:afp.com:20130110:03446623:879b:4553-9244
-110a3d91101e"/>
    </NewsManagement>
    <NewsComponent>
      <NewsLines>
        <DateLine>PARIS, 10 jan 2013 (AFP) - </DateLine>
        <HeadLine xml:lang="fr">Le nouveau texte patronal exclut
toujours une taxation des contrats courts </HeadLine>
        <NewsLine>
          <NewsLineType FormalName="AdvisoryLine"/>
          <NewsLineText xml:lang="fr">merci bien lire au 5e para que le
projet prévoit bien une prise en charge moitié-salarié,
```

Figure A.10 : Dépêche AFP 20130110T175231Z-TX-PAR-ITP91 (1/4)

```
moitié-entreprise des mutuelles d'entreprise </NewsLineText>
        </NewsLine>
      </NewsLines>
      <AdministrativeMetadata>
        <Provider>
          <Party FormalName="AFP"/>
        </Provider>
      </AdministrativeMetadata>
      <DescriptiveMetadata>
        <Language FormalName="fr"/>
        <Genre FormalName="Lead"/>
        <SubjectCode>
          <Subject FormalName="09000000"/>
        </SubjectCode>
        <SubjectCode>
          <SubjectMatter FormalName="09003000"/>
        </SubjectCode>
        <SubjectCode>
          <SubjectMatter FormalName="09010000"/>
        </SubjectCode>
        <SubjectCode>
          <SubjectMatter FormalName="09015000"/>
        </SubjectCode>
        <SubjectCode>
          <SubjectMatter FormalName="11006000"/>
        </SubjectCode>
        <SubjectCode>
          <Subject FormalName="11000000"/>
        </SubjectCode>
        <OfInterestTo FormalName="FRF-TFG-1=FAF"/>
        <DateLineDate>20130110T175223+0000</DateLineDate>
        <Location HowPresent="Origin">
          <Property FormalName="Country" Value="FRA"/>
          <Property FormalName="City" Value="PARIS"/>
        </Location>
        <Property FormalName="Version" Value="CORRECTION"/>
        <Property FormalName="Keyword" Value="Social"/>
        <Property FormalName="Keyword" Value="emploi"/>
        <Property FormalName="Keyword" Value="syndicat"/>
        <Property FormalName="Keyword" Value="patronat"/>
        <Property FormalName="Keyword" Value="gouvernement"/>
        <Property FormalName="GeneratorSoftware" Value="Cafp32"/>
      </DescriptiveMetadata>
      <ContentItem>
        <MediaType FormalName="Text"/>
        <Format FormalName="NITF3.1"/>
```

FIGURE A.11 : Dépêche AFP 20130110T175231Z-TX-PAR-ITP91 (2/4)

```
        <Characteristics>
          <SizeInBytes>2328</SizeInBytes>
          <Property FormalName="Words" Value="388"/>
        </Characteristics>
        <DataContent>
          <nitf>
            <body>
              <body.content>
                <p>Le patronat, dont les divisions ont de nouveau éclaté
jeudi, a transmis dans l'après-midi aux syndicats un texte remanié, mais
qui exclut toujours une taxation des contrats courts, condition posée
par les syndicats à un accord sur la "sécurisation de l'emploi".  </p>
                <p>Les discussions, interrompues vers 13H30, ont repris
peu après 17H00 avec la distribution du nouveau texte, avant d'être de
nouveau suspendues à 17H30, le temps de laisser les organisations
syndicales en prendre connaissance. </p>
                <p>La modulation des cotisations sociales, destinée à
décourager le recours aux contrats précaires abusifs et réclamée par les
syndicats, ne figure pas dans la dernière mouture du patronat,
contrairement à ce qu'espéraient les syndicats à la mi-journée.  </p>
                <p>Le patronat a en revanche retiré son projet de créer
des CDI très flexibles, liés à la durée d'un "projet" de 9 mois minimum.
Il propose d'aussi d'élargir les CDI intermittents, "à titre
expérimental" seulement. </p>
                <p>S'agissant de la généralisation des complémentaires
santé collectives, le projet réduit de 4 à 3 ans le délai maximal de
mise en oeuvre et prévoit qu'il soit "partagé par moitié entre salariés
et employeurs". </p>
                <p>Peu de temps avant la transmission du nouveau texte,
ce dernier round, qui doit s'achever au plus tard vendredi, a été marqué
par un "clash de l'UPA", selon les mots du Medef. Des dissensions entre
la CGPME et le Medef étaient déjà apparues au grand jour les semaines
précédentes.</p>
                <p>Dans un communiqué, l'UPA (artisans) a déclaré "ne
pas accepter un texte qui lèse la grande majorité des entreprises",
estimant que "le projet d'accord en cours de finalisation organise la
flexibilité de l'emploi au seul profit de quelques grandes entreprises
françaises", alors que la majorité "supportera l'essentiel des surcoûts
générés".</p>
                <p>La question du choix des prestataires chargés
d'offrir les mutuelles d'entreprise généralisées froisse
particulièrement l'UPA. "On a fait sauter le contrat de projet, le CDI
intermittent devient expérimental: l'UPA n'a plus rien", commentait
Marie-Françoise Leflon (CFE-CGC). </p>
                <p>Pour Joseph Thouvenel (CFTC), "le patronat aurait dû
faire avant ses arbitrages". "On a perdu temps", regrettait-il.  </p>
```

Figure A.12 : Dépêche AFP 20130110T175231Z-TX-PAR-ITP91 (3/4)

```
                <p>Depuis le coup d'envoi, le 4 octobre 2012, de cette
négociation cruciale, syndicats (CDFT, CGT, FO, CFE-CGC, CFTC) et
patronat (Medef, CGPME, UPA) cherchent les moyens de fluidifier le
marché du travail en donnant plus de souplesse aux entreprises et de
protection aux salariés. </p>
                <p>L'objectif des négociations est de conclure un accord
d'ici à vendredi. </p>
                <p>shu/db/bma</p>
              </body.content>
            </body>
          </nitf>
        </DataContent>
      </ContentItem>
    </NewsComponent>
  </NewsItem>
</NewsML>
```

FIGURE A.13 : Dépêche AFP 20130110T175231Z-TX-PAR-ITP91 (4/4)

```
<?xml version="1.0" encoding="utf-8"?>
<NewsML Version="1.2">
  <!--AFP NewsML text-photo profile evolution2 -->
  <!--Processed by Xafp1-4ToNewsML1-2 rev18 -->
  <Catalog Href="http://www.afp.com/dtd/AFPCatalog.xml"/>
  <NewsEnvelope>
    <TransmissionId>0574</TransmissionId>
    <DateAndTime>20130113T094806Z</DateAndTime>
    <NewsService FormalName="DGTE"/>
    <NewsProduct FormalName="JER"/>
    <NewsProduct FormalName="DAB"/>
    <NewsProduct FormalName="AMW"/>
    <NewsProduct FormalName="ELU"/>
    <NewsProduct FormalName="MOA"/>
    <NewsProduct FormalName="BRU"/>
    <NewsProduct FormalName="GVA"/>
    <NewsProduct FormalName="EUA"/>
    <NewsProduct FormalName="FRA"/>
    <NewsProduct FormalName="FRS"/>
    <NewsProduct FormalName="DILI"/>
    <NewsProduct FormalName="DGTE"/>
    <NewsProduct FormalName="DVBP"/>
    <NewsProduct FormalName="DAGI"/>
    <NewsProduct FormalName="DPSE"/>
    <NewsProduct FormalName="DVBA"/>
    <NewsProduct FormalName="DVBF"/>
    <Priority FormalName="3"/>
  </NewsEnvelope>
  <NewsItem xml:lang="fr">
    <Identification>
      <NewsIdentifier>
        <ProviderId>afp.com</ProviderId>
        <DateId>20130113T094803Z</DateId>
        <NewsItemId>TX-PAR-JAF12</NewsItemId>
        <RevisionId PreviousRevision="0" Update="N">1</RevisionId>

<PublicIdentifier>urn:newsml:afp.com:20130113T094803Z:TX-PAR-JAF12:1</Pu
blicIdentifier>
      </NewsIdentifier>
      <NameLabel>Somalie-France-otage-combats</NameLabel>
    </Identification>
    <NewsManagement>
      <NewsItemType FormalName="News"/>
      <FirstCreated>20130113T094758+0000</FirstCreated>
      <ThisRevisionCreated>20130113T094758+0000</ThisRevisionCreated>
      <Status FormalName="Usable"/>
```

Figure A.14 : Dépêche AFP 20130113T094803Z-TX-PAR-JAF12 (1/3)

```
      <Urgency FormalName="3"/>
      <AssociatedWith
NewsItem="urn:newsml:urn:newsml:afp.com:20130113:8ef985cc:1228:44d9-8ad9
-6e2dc894cf2c"/>
    </NewsManagement>
    <NewsComponent>
      <NewsLines>
        <DateLine>MOGADISCIO, 13 jan 2013 (AFP) - </DateLine>
        <HeadLine xml:lang="fr">Huit civils tués dans le raid français
en Somalie (témoins) </HeadLine>
      </NewsLines>
      <AdministrativeMetadata>
        <Provider>
          <Party FormalName="AFP"/>
        </Provider>
      </AdministrativeMetadata>
      <DescriptiveMetadata>
        <Language FormalName="fr"/>
        <SubjectCode>
          <SubjectDetail FormalName="02001007"/>
        </SubjectCode>
        <SubjectCode>
          <SubjectMatter FormalName="16002000"/>
        </SubjectCode>
        <SubjectCode>
          <Subject FormalName="16000000"/>
        </SubjectCode>
        <SubjectCode>
          <Subject FormalName="02000000"/>
        </SubjectCode>
        <SubjectCode>
          <SubjectMatter FormalName="02001000"/>
        </SubjectCode>
        <OfInterestTo FormalName="MOA-TFG-1=ORI"/>
        <OfInterestTo FormalName="DAB-TFG-1=DAB"/>
        <OfInterestTo FormalName="AMN-TFG-1=AMW"/>
        <OfInterestTo FormalName="FRE-TFG-1=ELU"/>
        <OfInterestTo FormalName="MOA-TFG-1=MOA"/>
        <OfInterestTo FormalName="EUA-TFG-1=EUA"/>
        <OfInterestTo FormalName="EUA-TFG-1=EUA"/>
        <OfInterestTo FormalName="EUA-TFG-1=EUA"/>
        <OfInterestTo FormalName="FRE-TFG-5=FRSGL"/>
        <DateLineDate>20130113T094758+0000</DateLineDate>
        <Location HowPresent="Origin">
          <Property FormalName="Country" Value="SOM"/>
          <Property FormalName="City" Value="MOGADISCIO"/>
```

FIGURE A.15 : Dépêche AFP 20130113T094803Z-TX-PAR-JAF12 (12/3)

```
        </Location>
        <Property FormalName="Keyword" Value="Somalie"/>
        <Property FormalName="Keyword" Value="France"/>
        <Property FormalName="Keyword" Value="otage"/>
        <Property FormalName="Keyword" Value="combats"/>
        <Property FormalName="GeneratorSoftware" Value="Cafp32"/>
      </DescriptiveMetadata>
      <ContentItem>
        <MediaType FormalName="Text"/>
        <Format FormalName="NITF3.1"/>
        <Characteristics>
          <SizeInBytes>552</SizeInBytes>
          <Property FormalName="Words" Value="92"/>
        </Characteristics>
        <DataContent>
          <nitf>
            <body>
              <body.content>
                <p>Au moins huit civils ont été tués samedi au cours du
raid français infructueux pour libérer un otage en Somalie, qui a fait
aussi plusieurs morts islamistes ou français, ont affirmé dimanche à
l'AFP des habitants sur place. </p>
                <p>Quatre de ces civils ont été tués lors de la
progression au sol des commandos français vers la localité de Bulomarer,
où l'otage était réputé être détenu. Quatre autres civils sont morts
dans les combats entre ces commandos et les insurgés islamistes à
Bulomarer, ont rapporté ces témoins, interrogés par téléphone depuis
Mogadiscio.</p>
                <p>nur-bb/jms</p>
              </body.content>
            </body>
          </nitf>
        </DataContent>
      </ContentItem>
    </NewsComponent>
  </NewsItem>
</NewsML>
```

Figure A.16 : Dépêche AFP 20130113T094803Z-TX-PAR-JAF12 (3/3)

3 Classification automatique de documents sur la taxonomie IPTC

La taxonomie de l'IPTC qui régit la classification thématique employée par l'AFP peut être appliquée de façon utile à d'autres données. Il s'agit notamment des connaissances rassemblées au sujet d'entités dans les bases de connaissances utilisées par les systèmes de Liage. Des corpus documentaires issus d'autres sources que l'AFP peuvent également faire l'objet d'une telle classification, qu'ils soient l'objet de la tâche de Liage ou intégrés dans le processus d'acquisition du modèle comme données d'apprentissage. Dans ces deux cas, l'emploi de catégories communes permet une meilleure adéquation entre modèle et données à traiter.

Une classification automatique de documents non AFP selon la taxonomie IPTC peut être envisagée grâce à la disponibilité d'archives de dépêches de l'agence en grande quantités, formant un large corpus étiqueté : chaque dépêche est en effet associée à des catégories thématiques de l'IPTC lors de sa production (cf. chapitre 4, section 2) : les combinaisons entre dépêches et catégories IPTC constituent ainsi des données de référence, pour lesquelles aucune tâche d'annotation supplémentaire n'est requise par ailleurs. L'apprentissage d'un classifieur peut être envisagé partir des données de référence ainsi mises à disposition.

Les données d'entraînement et d'évaluation rassemblées pour cette tâche comprennent 133 406 dépêches du fil généraliste en français de l'AFP, datées de l'année 2012. Chaque dépêche est associée à un ou plusieurs codes de sujets IPTC, ce qui porte le nombre d'association entre code IPTC et dépêche à 235 309. La distribution de ces sujets est indiquée à la table A.2. Chaque asso-

POL	**ECO**	**CLJ**	**CLT**	**UNR**	**HTH**	**SOI**	**ENV**
58 356	31 442	26 019	10 378	20 700	5 113	7 632	4 752
25%	**13%**	**11%**	**4%**	**8%**	**2%**	**3%**	**2%**
HUM	**SOC**	**SCI**	**REL**	**EDU**	**LIF**	**DIS**	**SPO**
10 065	12 707	4 061	7 332	3 635	1 620	5 548	25 123
4%	**5%**	**1,5%**	**3%**	**1,5%**	**0,6%**	**2%**	**11%**

TABLE A.2 : Distribution des catégories IPTC sur les 133 406 dépêches-exemples pour l'apprentissage de la classification.

ciation d'un code IPTC à une dépêche donne lieu à un exemple d'entraînement ou d'évaluation. L'ensemble de traits associés à ce code correspond au sac de mots dérivé du contenu textuel de la dépêche, chaque mot étant pondéré par le nombre de ses occurrences dans la dépêche considérée. Le sac de mots est obtenu après une série d'opérations de prétraitement usuelles : segmentation en tokens, filtrage des tokens non pertinents (mots grammaticaux, mots les plus fréquents du français, tokens non lexicaux) et des hapax (seuil minimal d'occurrences fixé à 10), stemming[1] par troncation des terminaisons courantes du français. La figure A.17 illustre le résultat de ce processus par quelques exemples ainsi obtenus. On peut observer que les deuxième et troisième exemples d'entraînement sur cette figure correspondent à la même dépêche, à laquelle les catégories SOI (Société) et CLT (Culture) ont été associées.

1. Comme cela a été indiqué au chapitre 2.1 (section 5) ette méthode est choisie pour sa simplicité, notamment comparée à la lemmatisation, et permet en outre de capturer une partie de la dérivation des formes traitées. Les terminaisons considérées pour la troncation sont *-e, -s, -t, -x, -es, -et, -nt, -ent*.

UNR troi|1 bomb|1 violenc|3 teresa|1 fermetur|1 meurtrier|1 kamikaz|1 veu|1 group|1 auparava|1 avai|4 2011|1 pay|3 sourc|1 assau|1 dan|5 discour|1 cultivateur|1 habita|2 fusillad|1 tuer|1 haram|4 martin|1 ebonyi|1 grossi|1 urgenc|1 parol|2 affrontem|2 entr|3 sec|1 sorti|1 islami|1 sud|1 ezza|1 propo|2 quelqu|1 rendu|1 proi|1 capital|1 boko|4 ordr|1 plu|1 million|1 comm|1 confli|1 qu|1 eta|2 contr|1 onyekachi|1 cancer|1 por|2 gouvernem|1 anti|1 attenta|3 foncier|3 onu|1 plusieur|1 suicid|1 lor|1 nigeria|4 jonathan|2 abuja|2 terrestr|1 nord|1 moin|1 nombreu|1 servic|1 deu|2 litig|1 faubourg|1 policier|1 sai|1 centr|1 ajouta|1 catholiqu|1 voisin|1 samedi|2 jusqu|1 renseignem|1 polic|2 eni|1 zon|1 goodluck|1 afriqu|1 centain|1 afp|1 villageoi|1 dernier|1 interreligieu|1 attaqu|2 personn|5 plac|1 parti|1 2008|1 gouverneur|1

SOI violenc|2 vill|1 jaun|1 incid|4 dimanch|1 trop|1 assoir|1 bastion|1 avai|1 centain|1 halakha|1 publiqu|1 dan|7 parfoi|1 fai|1 autr|1 montra|1 librairi|1 habita|1 hostil|1 porta|1 autobu|2 benjamin|1 verbal|1 ailleur|1 lign|1 hain|1 nombreu|1 dirigea|1 homm|1 entr|2 sexuell|1 majeur|1 sui|1 ministr|1 public|2 campagn|1 frang|2 expansion|1 compri|1 quotidienn|1 emprisonnem|1 qu|1 eta|1 discrimination|2 comm|2 orthodo|11 lesquel|1 initial|1 stric|1 contrai|1 femm|8 viv|1 shearim|1 fair|1 1980|1 plusieur|1 rigori|1 loi|1 semain|1 exhortation|1 depui|1 religieu|5 pri|1 imag|1 opinion|1 politiqu|1 rassemblem|1 manifestation|3 adop|1 lectur|1 population|1 insulter|1 cesser|1 bei|3 shemesh|3 ultra|11 majoritairem|1 objectif|1 san|1 premier|1 refu|1 quartier|1 soir|1 importa|2 samedi|1 sharim|1 polic|1 pay|1 juiv|1 physiqu|1 protester|1 craigna|1 juif|6 souligna|1 pratiqu|1 dernier|1 tension|2 attaqu|2 contr|9 netanyahu|1 cracher|1 mai|2 notamm|3 occidental|1 membr|1 parti|1 position|1 clima|1

CLT violenc|2 vill|1 jaun|1 incid|4 dimanch|1 trop|1 assoir|1 bastion|1 avai|1 centain|1 halakha|1 publiqu|1 dan|7 parfoi|1 fai|1 autr|1 montra|1 librairi|1 habita|1 hostil|1 porta|1 autobu|2 benjamin|1 verbal|1 ailleur|1 lign|1 hain|1 nombreu|1 dirigea|1 homm|1 entr|2 sexuell|1 majeur|1 sui|1 ministr|1 public|2 campagn|1 frang|2 expansion|1 compri|1 quotidienn|1 emprisonnem|1 qu|1 eta|1 discrimination|2 comm|2 orthodo|11 lesquel|1 initial|1 stric|1 contrai|1 femm|8 viv|1 shearim|1 fair|1 1980|1 plusieur|1 rigori|1 loi|1 semain|1 exhortation|1 depui|1 religieu|5 pri|1 imag|1 opinion|1 politiqu|1 rassemblem|1 manifestation|3 adop|1 lectur|1 population|1 insulter|1 cesser|1 bei|3 shemesh|3 ultra|11 majoritairem|1 objectif|1 san|1 premier|1 refu|1 quartier|1 soir|1 importa|2 samedi|1 sharim|1 polic|1 pay|1 juiv|1 physiqu|1 protester|1 craigna|1 juif|6 souligna|1 pratiqu|1 dernier|1 tension|2 attaqu|2 contr|9 netanyahu|1 cracher|1 mai|2 notamm|3 occidental|1 membr|1 parti|1 position|1 clima|1

Figure A.17 : Exemples de classification dérivés des associations de sujets IPTC aux dépêches du corpus d'apprentissage.

On dispose donc de 235 309 exemples de classification selon la taxonomie IPTC avec pour traits discriminants un vocabulaire pondéré dérivé de la dépêche considérée. Plusieurs catégories IPTC pouvant être associées à un même vocabulaire, on envisage la tâche comme un cas de classification multiple dans laquelle chaque instance à classifier peut être associée à plusieurs catégories parmi les 16 possibles. Chaque catégorie c donne lieu au développement d'un classifieur binaire spécialisé, devant assigner pour chaque instance x constituée d'un vocabulaire pondéré un label y tel que

$$y = \left\{ \begin{array}{ll} 1 & \text{si } x \text{ est associé à } c \\ 0 & \text{sinon} \end{array} \right\}$$

Toutes les catégories pour lesquelles $y = 1$ pour un exemple x sont ainsi associées à x et donc au document représenté par x. Les classifieurs binaires sont des classifieurs à maximum d'entropie développés à l'aide du logiciel *MEGA Model Optimization Package* de Daumé III [DI04] [2].

Classification automatique de documents non-AFP

Les articles de l'encyclopédie Wikipedia, utilisés dans l'acquisition de connaissances sur les entités (cf. chapitre 5, section 2.2.2), font l'objet d'une classification automatique selon la méthode présentée ici afin de munir ces entités de caractéristiques comparables aux document de l'AFP lors du processus d'alignement. La table A.3 présente la distribution des catégories IPTC sur l'ensemble des articles Wikipedia dont ont été dérivées les entités (personnes et organisations) de la base Aleda, pour un usage dans la base de connaissances associée au système Nomos (Nomos-KB, cf. chapitre 5, section 2.2.2).

CLT	**ECO**	**SPO**	**HUM**
115 624	57 194	49 593	44 858
24,99%	**12,36%**	**10,72%**	**9,70%**
SCI	**SOI**	**POL**	**UNR**
37 080	31 703	27 502	18 439
8,01%	**6,85%**	**5,94%**	**3,99%**
LIF	**CLJ**	**REL**	**SOC**
17 302	17 003	12 353	10 533
3,74%	**3,67%**	**2,67%**	**2,28%**
EDU	**ENV**	**HTH**	**DIS**
9 540	5 402	5 227	3 334
2,06%	**1,17%**	**1,13%**	**0,72%**

TABLE A.3 : Distribution des catégories IPTC sur l'ensemble des articles de l'encyclopédie Wikipedia donnant lieu à une entité d'Aleda.

2. Disponible à l'adresse `http://www.umiacs.umd.edu/~hal/megam`

4 Enrichissement de dépêches à l'aide de métadonnées

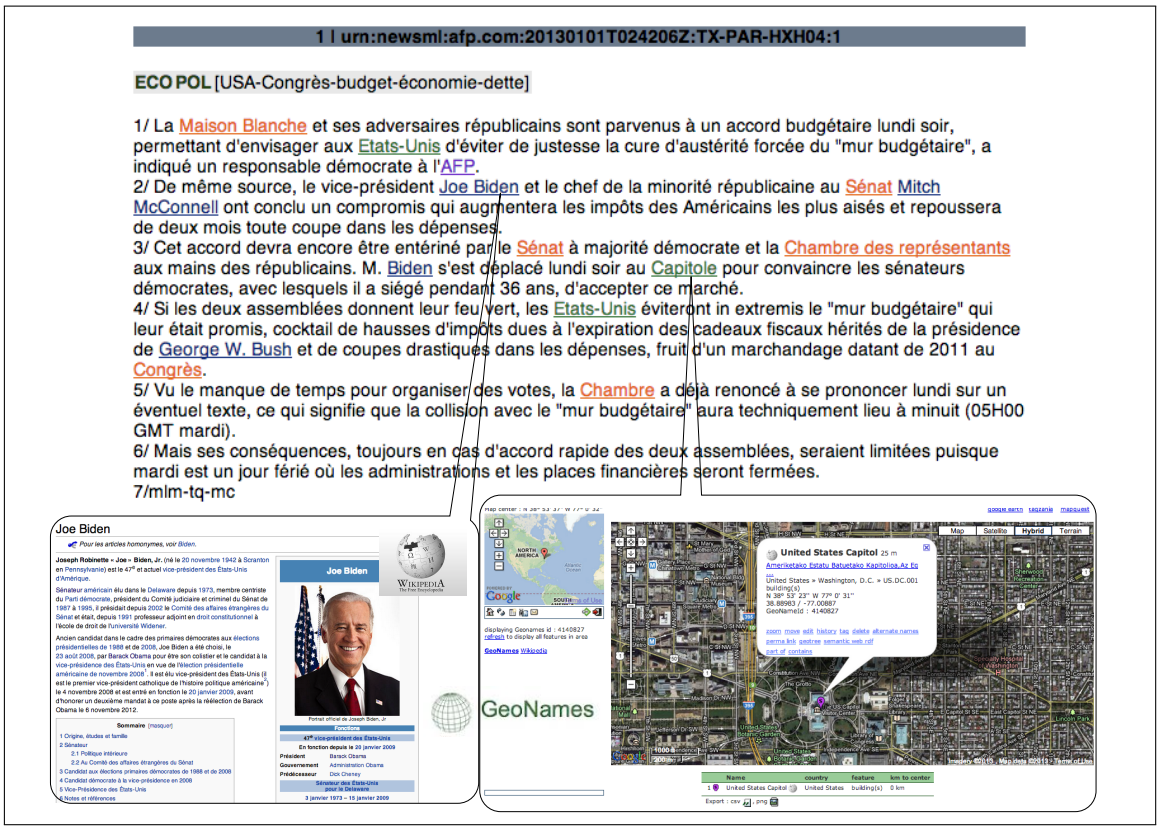


Figure A.18 : Dépêche enrichie : visualisation en HTML (métadonnées au format RDFa)

```
<html>
    <head>
        <meta http-equiv="Content-Type" content="text/html; charset=utf-8"
            vocab="http://afp.com/AMO/metadata/" />
        <title> AFP News - Metadata </title>
    </head>
    <body>

        <h4>1 | urn:newsml:afp.com:20130101T024206Z:TX-PAR-HXH04:1 </h4>
        <table>
            <tr>
                <td>ECO</td>
                <td>POL</td>
                <td>[USA-Congrès-budget-économie-dette]</td>
            </tr>
        </table>
        <div>
            <span>1/</span> La<a title="Maison Blanche [1000000002872401]"
                href="http://fr.wikipedia.org/wiki/Maison_blanche"
resource="#1000000002872401"
                typeof="Organization">Maison Blanche</a> et ses
adversaires républicains sont
            parvenus à un accord budgétaire lundi soir, permettant
d'envisager aux <a
                title="United States [2000000006252001]"
href="http://www.geonames.org/6252001"
                resource="#2000000006252001"
typeof="Country">Etats-Unis</a> d'éviter de justesse la
            cure d'austérité forcée du "mur budgétaire", a indiqué un
responsable démocrate à l' <a
                title="Agence France-Presse [1000000000055197]"
                href="http://fr.wikipedia.org/wiki/Agence_France-Presse"
                resource="#1000000000055197"
typeof="Organization">AFP</a>. </div>
        <div>
            <span>2/</span> De même source, le vice-président <a
                title="Joe Biden (m) [1000000000255711]"
                href="http://fr.wikipedia.org/wiki/Joe_Biden"
resource="#1000000000255711"
                typeof="Person">Joe Biden</a> et le chef de la minorité
républicaine au <a
                title="Sénat [1000000000089073]"
                href="http://fr.wikipedia.org/wiki/Sénat des États-Unis"
                resource="#1000000000089073"
typeof="Organization">Sénat</a>
            <a title="Mitch McConnell (m) [1000000000263731]"
```

FIGURE A.19 : Dépêche enrichie : source HTML (métadonnées au format RDFa) (1/3)

```
                href="http://fr.wikipedia.org/wiki/Mitch_McConnell"
resource="#1000000000263731"
                typeof="Person">Mitch McConnell</a> ont conclu un
compromis qui augmentera les
            impôts des Américains les plus aisés et repoussera de deux
mois toute coupe dans les
             dépenses. </div>
        <div>
            <span>3/</span> Cet accord devra encore être entériné par le <a
                title="Sénat [1000000000089073]"
                href="http://fr.wikipedia.org/wiki/Sénat des États-Unis"
                resource="#1000000000089073"
typeof="Organization">Sénat</a> à majorité démocrate et
             la <a title="Chambre des représentants des États-Unis
[1000000000089459]"
                href="http://fr.wikipedia.org/wiki/Chambre des
représentants des États-Unis"
                resource="#1000000000089459"
typeof="Organization">Chambre des représentants</a> aux
            mains des républicains. M. <a title="Joe Biden (m)
[1000000000255711]"
                href="http://fr.wikipedia.org/wiki/Joe_Biden"
resource="#1000000000255711"
                typeof="Person">Biden</a> s'est déplacé lundi soir au <a
                title="United States Capitol [2000000004140827]"
                href="http://www.geonames.org/4140827"
resource="#2000000004140827" typeof="POI"
                >Capitole</a> pour convaincre les sénateurs démocrates,
avec lesquels il a siégé
            pendant 36 ans, d'accepter ce marché. </div>
        <div>
            <span>4/</span> Si les deux assemblées donnent leur feu
vert, les <a
                title="United States [2000000006252001]"
href="http://www.geonames.org/6252001"
                resource="#2000000006252001"
typeof="Country">Etats-Unis</a> éviteront in extremis
            le "mur budgétaire" qui leur était promis, cocktail de
hausses d'impôts dues à
            l'expiration des cadeaux fiscaux hérités de la présidence de <a
                title="George W. Bush (m) [1000000000680078]"
                href="http://fr.wikipedia.org/wiki/George_W._Bush"
resource="#1000000000680078"
                typeof="Person">George W. Bush</a> et de coupes
drastiques dans les dépenses, fruit
            d'un marchandage datant de 2011 au <a title="Congrès des
```

Figure A.20 : Dépêche enrichie : source HTML (métadonnées au format RDFa) (2/3)

```
États-Unis[1000000000088935]"
                href ="http://fr.wikipedia.org/wiki/Congrès des États-Unis"
                resource ="#1000000000088935"
typeof="Organization">Congrès</a>. </div>
        <div>
            <span>5/</span> Vu le manque de temps pour organiser des
votes,  la <a
                title ="Chambre des représentants des États-Unis
[1000000000089459]"
                href ="http://fr.wikipedia.org/wiki/Chambre des
représentants des États-Unis"
                resource ="#1000000000089459"
typeof="Organization">Chambre</a> a déjà renoncé à se
            prononcer lundi sur un éventuel texte, ce qui signifie que
la collision avec le "mur
            budgétaire" aura techniquement lieu à minuit (05H00 GMT
mardi). </div>
        <div>
            <span>6/</span> Mais ses conséquences, toujours en cas
d'accord rapide des deux
            assemblées, seraient limitées puisque mardi est un jour
férié où les administrations et
            les places financières seront fermées. </div>
        <div>
            <span>7/</span>mlm-tq-mc</div>
    </body>
</html>
```

FIGURE A.21 : Dépêche enrichie : source HTML (métadonnées au format RDFa) (3/3)

```
<?xml version="1.0" encoding="utf-8"?>
<NewsML Version="1.2">
    <Catalog Href="http://www.afp.com/dtd/AFPCatalog.xml"/>
    <NewsEnvelope>
        <TransmissionId>0149</TransmissionId>
        <DateAndTime>20130101T024211Z</DateAndTime>
        <NewsService FormalName="DGTE"/>
        <NewsProduct FormalName="FRS"/>
        <NewsProduct FormalName="ECF"/>
        <Priority FormalName="4"/>
    </NewsEnvelope>
    <NewsItem xml:lang="fr">
        <Identification>
            <NewsIdentifier>
                <ProviderId>afp.com</ProviderId>
                <DateId>20130101T024206Z</DateId>
                <NewsItemId>TX-PAR-HXH04</NewsItemId>
                <RevisionId PreviousRevision="0" Update="N">1</RevisionId>

<PublicIdentifier>urn:newsml:afp.com:20130101T024206Z:TX-PAR-HXH04:1</Pu
blicIdentifier>
            </NewsIdentifier>
            <NameLabel>USA-Congrès-budget-économie-dette</NameLabel>
        </Identification>
        <NewsManagement>
            <NewsItemType FormalName="News"/>
            <FirstCreated>20130101T024149+0000</FirstCreated>
            <ThisRevisionCreated>20130101T024149+0000</ThisRevisionCreated>
            <Status FormalName="Usable"/>
            <Urgency FormalName="4"/>
            <AssociatedWith
NewsItem="urn:newsml:urn:newsml:afp.com:20130101:0f829142:6492:456b-93a3
-5fbfda9b21a2"/>
        </NewsManagement>
        <NewsComponent>
            <NewsLines>
                <DateLine>WASHINGTON, 01 jan 2013 (AFP) - </DateLine>
                <HeadLine xml:lang="fr">USA: accord budgétaire entre
Maison Blanche et républicains (source démocrate) </HeadLine>
            </NewsLines>
            <AdministrativeMetadata>
                <Provider><Party FormalName="AFP"/></Provider>
            </AdministrativeMetadata>
            <DescriptiveMetadata>
                <Language FormalName="fr"/>
                <SubjectCode><SubjectMatter
```

Figure A.22 : Dépêche enrichie : format NewsML (métadonnées au format XML) (1/3)

```
FormalName="04008000"/></SubjectCode>
                <SubjectCode><SubjectDetail
FormalName="04008008"/></SubjectCode>
                <SubjectCode><SubjectMatter
FormalName="11009000"/></SubjectCode>
                <SubjectCode><SubjectMatter
FormalName="11013000"/></SubjectCode>
                <SubjectCode><SubjectMatter
FormalName="11006000"/></SubjectCode>
                <SubjectCode><Subject FormalName="04000000"/></SubjectCode>
                <SubjectCode><Subject FormalName="11000000"/></SubjectCode>
                <SubjectCode><SubjectMatter
FormalName="04017000"/></SubjectCode>
                <OfInterestTo FormalName="FRE-TFG-5=FRSGL"/>
                <OfInterestTo FormalName="SEF-TFE-1=ECF"/>
                <DateLineDate>20130101T024149+0000</DateLineDate>
                <Location HowPresent="Origin">
                    <Property FormalName="Country" Value="USA"/>
                    <Property FormalName="City" Value="WASHINGTON"/>
                </Location>
                <Property FormalName="Keyword" Value="USA"/>
                <Property FormalName="Keyword" Value="Congrès"/>
                <Property FormalName="Keyword" Value="budget"/>
                <Property FormalName="Keyword" Value="économie"/>
                <Property FormalName="Keyword" Value="dette"/>
                <Property FormalName="GeneratorSoftware" Value="Cafp32"/>
            </DescriptiveMetadata>
            <ContentItem>
                <MediaType FormalName="Text"/>
                <Format FormalName="NITF3.1"/>
                <Characteristics>
                    <SizeInBytes>1470</SizeInBytes>
                    <Property FormalName="Words" Value="245"/>
                </Characteristics>
                <DataContent>
                <nitf>
                <body>
                <body.content>
                <p>La <ENAMEX aleda_id="1000000002872401">Maison
Blanche</ENAMEX> et ses adversaires républicains sont parvenus à
              un accord budgétaire lundi soir, permettant d'envisager aux
                <ENAMEX aleda_id="2000000006252001">Etats-Unis</ENAMEX>
d'éviter de justesse la cure d'austérité forcée du
               "mur budgétaire", a indiqué un responsable démocrate à
l'<ENAMEX aleda_id="1000000000055197">AFP</ENAMEX>.</p>
                <p>De même source, le vice-président <ENAMEX
```

FIGURE A.23 : Dépêche enrichie : format NewsML (métadonnées au format XML) (2/3)

```
aleda_id="1000000000255711">Joe Biden</ENAMEX> et le chef de la
                minorité républicaine au <ENAMEX
aleda_id="1000000000089073">Sénat</ENAMEX>
                <ENAMEX aleda_id="1000000000263731">Mitch
McConnell</ENAMEX> ont conclu un compromis qui augmentera les
                impôts des Américains les plus aisés et repoussera de
deux mois toute coupe dans les dépenses. </p>
                <p>Cet accord devra encore être entériné par le <ENAMEX
aleda_id="1000000000089073">Sénat</ENAMEX> à majorité
                démocrate et la <ENAMEX
aleda_id="1000000000089459">Chambre des représentants </ENAMEX> aux mains
des
                républicains. M. <ENAMEX
aleda_id="1000000000255711">Biden</ENAMEX> s'est déplacé lundi soir au
                <ENAMEX aleda_id="2000000004140827">Capitole</ENAMEX>
pour convaincre les sénateurs démocrates, avec
                lesquels il a siégé pendant 36 ans, d'accepter ce
marché.</p>
                <p>Si les deux assemblées donnent leur feu vert, les
<ENAMEX aleda_id="2000000006252001">Etats-Unis</ENAMEX>
                éviteront in extremis le "mur budgétaire" qui leur était
promis, cocktail de hausses d'impôts dues à
                l'expiration des cadeaux fiscaux hérités de la
présidence de
                <ENAMEX aleda_id="1000000000680078">George W.
Bush</ENAMEX> et de coupes drastiques dans les dépenses,
                fruit d'un marchandage datant de 2011 au <ENAMEX
aleda_id="1000000000088935">Congrès</ENAMEX>.</p>
                <p>Vu le manque de temps pour organiser des votes, la
<ENAMEX aleda_id="1000000000089459">Chambre</ENAMEX> a déjà
                renoncé à se prononcer lundi sur un éventuel texte, ce
qui signifie que la collision avec le "mur budgétaire"
                aura techniquement lieu à minuit (05H00 GMT mardi). </p>
                <p>Mais ses conséquences, toujours en cas d'accord
rapide des deux assemblées, seraient limitées puisque mardi
                est un jour férié où les administrations et les places
financières seront fermées. </p>
                <p>mlm-tq/mc</p>
                </body.content>
                </body>
                </nitf>
                </DataContent>
            </ContentItem>
        </NewsComponent>
    </NewsItem>
</NewsML>
```

Figure A.24 : Dépêche enrichie : format NewsML (métadonnées au format XML) (3/3)

Annexe B

Nomos

1 Traits pour l'apprentissage supervisé

[1] c_wpw Poids Wikipedia indiqué par la base Aleda : taille du texte de l'article Wikipedia correspondant à l'entité, en nombre de lignes, reflétant sa notoriété. Celle-ci détermine un sens par défaut d'une mention pour laquelle cette entité est candidate. La valeur de ce trait est plus précisément le logarithme de la taille.

[2] c_geonw Poids GeoNames indiqué par la base Aleda : nombre d'habitants du lieu, reflétant son importance relativement à d'autres lieux du même nom. La valeur de ce trait est plus précisément le logarithme du nombre d'habitants.

[3] c_weightnull Les articles Wikipedia et les entrées de GeoNames ne donnent pas systématiquement lieu à un poids : certains articles Wikipedia ne comprennent pas de contenu textuel (page en attente de rédaction ou contenu graphique seulement, par exemple), et le nombre d'habitants n'est pas indiqué pour certains lieux dans GeoNames, notamment de petites communes ou lieux-dits. Lorsqu'Aleda n'indique donc pas d'attribut correspondant au poids Wikipedia ou GeoNames, ce trait prend la valeur 1.

[4] c_wp_occs Nombre d'occurrences du candidat dans le corpus Wikipedia, soit le nombre d'occurrences d'un wikilink dont ce candidat est la référence. Ce trait peut être vu comme un second type de modélisation de la notoriété, parallèlement à c_wpw.

[5] c_wp_occs_rel Valeur de c_wp_occs normalisée par la somme des valeurs du même trait pour tous les candidats de la mention courante.

[6] c_wp_occs_m Nombre d'occurrences du candidat dans le corpus Wikipedia avec la mention courante, soit le nombre d'occurrences d'un wikilink dont ce candidat est la référence et dont la mention est l'ancre textuelle. Ce trait modélise la tendance d'une entité à être mentionnée en contexte avec chacune de ses mentions possibles et peut ainsi favoriser la paire (m, c) en fonction de sa valeur.

[7] c_titlepar_in_m_suw On retient les tokens autour de la mention courante dans le document traité, avec une fenêtre de 3 tokens à gauche et 3 tokens à droite. Si le candidat présente un e-`titlepar` (mot entre parenthèses figurant à la suite du titre de l'article Wikipedia), renseigné par Nomos-KB, ce trait prend la valeur 1 si le e-`titlepar` figure parmi les tokens environnants, 0 sinon. Si le e-`titlepar` a par exemple la valeur *ministre* (entité `Auguste Paris (ministre)`), en association avec une mention entourée du même mot (*M. Paris, ministre des Travaux Publics*), le trait prend la valeur 1.

[8] c_name_in_doc Un nom canonique est indiqué pour les entités d'Aleda. Ce trait prend la valeur 1 si le nom canonique du candidat figure parmi l'ensemble des mentions du document traité, 0 sinon. Par exemple, la mention *Obama* présente parmi ses candidats une entité dont le nom canonique est `Barack Obama` ; si une autre mention du document est identique à ce nom, ce trait a la valeur 1. Il permet de tenir compte des cas d'homonymie provoqués notamment par les mentions par noms de famille seuls : ici, si le document contient les mentions *Obama* et *Barack Obama*, le candidat possible pour la mention *Obama* dont le nom canonique est `Michelle Obama` ne donnera pas de valeur positive à ce trait. On peut observer que ce trait introduit une approximation du regroupement de mentions par résolution de coréférence ; celle-ci est en effet utilisée par plusieurs systèmes de Liage évoqués au chapitre 3 et peut être utile afin de traiter plusieurs mentions, considérées comme coréférentes et donc à aligner sur une même entité, comme une requête unique, et ainsi de contraindre davantage la sélection des candidats. Le trait c_name_in_doc vise

à représenter le contenu de telles contraintes sous la forme d'une préférence pour un candidat, sans que les mentions éventuellement concernées par la même entité ne soient explicitement regroupées.

[9] c_noname Lorsqu'aucun des candidats issus de la BC ne présente le trait c_name_indoc avec la valeur 1, ce trait prend la valeur 1, 0 sinon.

[10] c_name_m_dist Distance d'édition entre le nom canonique du candidat et la mention.

[11] c_aleda_type_in_m_types Les systèmes de REN retournent des mentions typées (PERSON, ORGANIZATION ou LOCATION). Lorsque la REN est non-déterministe et que plusieurs mentions de même forme, c'est-à-dire de mêmes frontières, sont retournées avec des types différents, Nomos considère une mention unique munie d'un ensemble de types. Dans l'analyse finalement retournée, les mentions identifiées portent le type de l'entité d'alignement lorsque celle-ci est issue de la BC. Les types assignés à une mention par la REN ne contraignent donc pas l'alignement, qui peut être fait sur une entité de type différent ; ils sont en revanche intégrés sous la forme de ce trait, selon l'idée que l'indication de type de la REN, même ambigüe, peut guider la recherche de l'alignement correct. Ce trait prend donc la valeur 1 lorsque le candidat partage son type avec la mention.

[12] c_gender_equals_m_gender Les entités de type PERSON présentent dans Aleda un attribut indiquant leur genre (masculin ou féminin), lorsqu'il peut être déterminé par le prénom inclus dans le nom canonique de l'entité. Le module de REN SxPipe/NP retourne par ailleurs, pour un certain nombre de mentions de personnes, un genre. Cette information est déterminée soit par le prénom s'il est inclus dans la mention et qu'il indique de façon très sure le genre, soit d'après les informations associées à la mention dans le lexique de variantes utilisé par SxPipe/NP et dérivé de la base Aleda. Le trait c_gender_equals_m_gender prend la valeur 1 lorsque le genre du candidat est le même que celui de la mention, dans les configurations du système intégrant les analyses de SxPipe/NP.

[13] c_name_m_longer_dist Pour une mention donnée de type PERSON ou ORGANIZATION, si au moins une mention du document traité présente une forme lexicale plus longue (par exemple *Barack Obama* pour la mention *Obama*), on calcule une distance d'édition entre ces mentions plus longues et le nom canonique du candidat s'il est de type PERSON ou ORGANIZATION. Ce trait modélise une forme de coréférence légère entre les mentions similaires (prénom et nom de famille vs. noms de famille seuls, acronymes et formes étendues).

[14] c_multi_m Ce trait prend la valeur 1 si le candidat est également candidat pour d'autres mentions de la dépêche. On peut en effet considérer que son association avec plusieurs mentions, y compris différentes de la mention courante, rend plus vraisemblable sa dénotation comme entité dans le document traité.

[15] c_iptc_d_iptc_sim Similarité cosinus entre les catégories IPTC du document (D_{iptc}) et celles associées au candidat dans Nomos-KB (E_{iptc1} et E_{iptc2}). Ce trait cherche à capturer une proximité de contexte thématique entre l'entité et la mention à partir du document et ainsi à discriminer sur ce plan les différents candidats.

[16] c_slugs_d_slugs_sim Similarité cosinus entre les slugs de document (D_{slugs}) et ceux associés au candidat lorsque les catégories de son article Wikipedia donnent lieu à des correspondances avec les slugs de l'AFP (E_{slugs1} et E_{slugs2}).

[17] c_sw_d_sw_sim d et c sont associés à des ensembles de mots saillants, respectivement D_{sbow} et E_{sbow}. L'ensemble E_{sbow} est déterminé de façon statique dans Nomos-KB : le

score de saillance de chaque mot est relatif à l'article de l'entité et au corpus entier. Il est ici de nouveau calculé pour chaque paire (m, c) présentée au système. La saillance des mots associés à un candidat est alors calculée non pas relativement à l'ensemble du corpus Wikipedia mais au sous-corpus formé par les candidats de la mention courante, plus précisément par leurs articles Wikipedia, afin d'être discriminante entre ces candidats et non pas entre un candidat et toutes les autres entités PERSON et ORGANIZATION d'Aleda ; on obtient ainsi un autre ensemble E'_{SBOW}, dont la similarité cosinus avec D_{sbow} donne la valeur du trait c_sw_d_sw_sim.

[18] c_sw_d_locvars_sim Similarité cosinus entre les mots saillants associés au candidat, selon le même calcul que précédemment (E'_{SBOW}), et $D_{locvars}$, c'est-à-dire l'ensemble des variantes des entités de type LOCATION possiblement mentionnées dans les métadonnées de la dépêche (de façon certaine pour les pays mais non univoque pour les villes et autres sous-types). Ce trait permet de capturer les co-occurrences de c avec les lieux caractéristiques de la dépêche, par le biais de leurs variantes seulement. Ainsi, la variante *Michael Jackson* présente parmi ses candidats un chanteur américain célèbre et un écrivain britannique spécialiste de la bière et du whisky, tous deux ayant pour nom canonique `Michael Jackson`. Si le document traité est muni d'une métadonnée indiquant un pays avec la valeur *GBR*, la variante *Grande-Bretagne* figurera dans $D_{locvars}$. Si le nom des pays de nationalité des deux candidats figure parmi leurs E'_{SBOW} respectifs, on aura alors une similarité non nulle entre E'_{SBOW} et $D_{locvars}$ pour le second.

[19] c_sw_m_suw_sim Similarité cosinus entre les mots saillants associés au candidat, selon le même calcul que précédemment (E'_{SBOW}), et les mots autour de la mention dans le texte de la dépêche. On cherche notamment à capturer avec ce trait les descriptions nominales en apposition, comme dans *M. Paris, ministre des Travaux Publics*, qui peuvent également figurer parmi les mots de E'_{SBOW} et ainsi discriminer plusieurs candidats pour une même mention.

[20] c_peers_vars_d_mentions_sim Les entités de types PERSON et ORGANIZATION sont associées dans Nomos-KB à des entités co-occurrentes ou reliées dans Wikipedia (cf. ensembles E_{ewl1}, E_{ewl2} et E_{rel}, chapitre 5, section 2.2.2, soit les entités *parentes*). Le document traité présente quant à lui un ensemble de mentions. La valeur de ce trait correspond à la similarité cosinus entre l'ensemble des variantes des entités parentes du candidat et l'ensemble des mentions du document.

[21] c_peers_vars_d_sw_sim Similarité cosinus entre les variantes d'entités co-occurrentes du candidat, comme dans le trait précédent, et les mots saillants associés au document.

[22] c_d_sim Ce trait tient compte des différents traits indiquant le degré de similarité du candidat avec le document : [] c_sw_d_sw_sim, [] c_sw_m_suw_sim, [] c_slugs_d_slugs_sim, [] c_iptc_d_iptc_sim, [] c_peers_vars_d_mentions_sim, [] c_peers_vars_d_sw_sim et [] c_sw_d_locvars_sim. La valeur du trait [] c_d_sim est la valeur maximale de la liste des valeurs correspondant à chacun de ces traits.

[23] c_d_sim_level Pour les candidats NIL et NAE, on indique par ce trait le niveau de similarité de l'ensemble des candidats issus de la BC de type PERSON ou ORGANIZATION pour la même mention, en supposant qu'une similarité faible à ce niveau, combinée à d'autres critères, peut favoriser le cas NIL ou l'interprétation NAE. La valeur du trait [] c_d_sim_level est la valeur maximale de la liste des valeurs du trait [] c_d_sim pour chaque candidat concerné.

[24] c_countrydiv Si le candidat est de type LOCATION et présente un sous-type dans Aleda, et que ce sous-type correspond à une division administrative (département, conté, canton..., le trait prend la valeur 1, 0 sinon. Ce trait discrimine les candidats de type LOCATION de même nom, tels que `Paris`, capitale de la France et `Paris`, département français.

[25] c_countrycode_equals_d_countrycode Les entités de type LOCATION sont associées dans Aleda à un pays identifié par un code ISO-3166-2, lui-même correspondant à l'une des entrées de sous-type COUNTRY d'Aleda. Les dépêches AFP sont munies d'une métadonnée indiquant le pays concerné par la dépêche — plus précisément le pays de rédaction de la dépêche, non nécessairement identique au pays concerné par l'événement rapporté — par un code ISO-3166-3 lui aussi mis en correspondance avec l'un des pays recensés par Aleda, par le bias de Nomos-KB. Ce trait prend la valeur 1 lorsque le pays de l'entité et celui du document sont identiques.

[26] c_country_equals_d_country Ce trait prend la valeur 1 lorsque le candidat est une entité de type COUNTRY et qu'il s'agit du pays indiqué dans les métadonnées de la dépêche.

[27] c_size Ce trait prend pour valeur le nombre de candidats issus de la BC associés à la mention courante. On suppose qu'un nombre réduit de candidats issus de la BC peut favoriser une interprétation non dénotationnelle de la mention.

[28] c_nil_sole Ce trait prend la valeur 1 lorsqu'aucun candidat issu de la BC n'est trouvé pour la mention. On suppose que l'absence de candidats issus de la BC peut favoriser une interprétation non dénotationnelle de la mention.

[29] m_conf Score de confiance retourné par LIANE pour la détection de chaque mention.

[30] m_spos1 Ce trait prend la valeur 1 lorsque la mention se trouve en tête de phrase, 0 sinon ; il peut indiquer une ambiguïté de la mention avec une forme non dénotationnelle.

[31] m_wpoccs Nombre d'occurrences de la mention dans Wikipedia, en tant qu'ancre textuelle de wikilinks d'articles correspondant à des entrées d'Aleda. La valeur de ce trait peut indiquer une propension de la mention à être employée de façon dénotationnelle ou non.

[32] m_doccs Nombre d'occurrences de la mention dans la dépêche. En supposant que des mentions identiques dans un même document sont coréférentes et que de multiples occurrences d'une même forme lexicale analysée comme dénotationnelle est peu probablement un faux positif, ce trait peut favoriser ou non un alignement sur NAE.

[33] m_inlefff Si la mention est présente dans un lexique de formes du français[1] en lettres minuscules, ce trait prend la valeur 1, 0 sinon. Il modélise ainsi la possible ambiguïté de la mention avec une forme non dénotationnelle.

[34] m_inlefff_nocap Identique au trait précédent, si la mention est présente dans le lexique sans capitale à l'initiale.

[35] m_dagu Dans les configurations de REN non déterministes, ce trait prend la valeur 1 si la mention constitue l'unique transition dénotationnelle

1. Nous utilisons pour cette tâche le lexique morpho-syntaxique Le*fff* [Sag10], développé dans le cadre de la plateforme de modélisation et d'acquisition d'informations lexicales Alexina (`http://gforge.inria.fr/projects/alexina/`)

[36] m_short Dans les configurations de REN non déterministes, au sein des zones ambigues, ce trait prend la valeur 1 si la mention se trouve dans le chemin le plus court de la zone. On oppose ainsi, dans *Le maire d'Orange Alain Labé enseignait...* la lecture avec la mention *Orange Alain Labé* à la lecture avec les deux mentions *Orange* et *Alain Labé*.

[37] m_long Symétriquement au trait précédent, ce trait prend la valeur 1 si la mention se trouve dans le chemin le plus long de la zone.

[38] m_nolongerexists Pour les mentions de type PERSON, ce trait prend la valeur 1 si le document ne présente aucune variante plus longue que la mention. Il s'agit notamment de cas où une mention isolée, qui pourrait correspondre à un nom de famille défini dans le lexique, n'est pas précédé dans la dépêche de la mention avec le nom complet correspondant ; l'interprétation non dénotationnelle de la mention peut alors être favorisée.

[39] m_multiner Dans la configuration de REN par union entre LIANE et SxPipe/NP, ce trait prend la valeur 1 lorsque la mention est retournée par les deux systèmes.

Le système Nomos réalise l'alignement de chaque mention de façon distincte : ses résultats ne sont pas soumis à des contraintes globales comme c'est le cas dans certains systèmes, où l'alignement de toutes les mentions doit respecter une cohérence globale de co-occurrences d'entités. Les traits [] c_sw_d_locvars_sim, [] c_peers_vars_d_mentions_sim, [] c_peers_vars_d_sw_sim, [] c_countrycode_equals_d_countrycode et [] c_country_equals_d_country sont introduits ici afin d'intégrer de façon non contrainte les co-occurrences courantes d'entités, ou d'entités avec un certain contexte lexical ou géographique.

2 Modèles : configurations de REN et traits

REN	Traits
Gold	c_wpw, c_geonw, c_admloc, c_wp_occs, c_iptc_d_iptc_sim, c_multi_m, c_peers_vars_d_mentions_sim, c_name_in_doc, c_sw_d_sw_sim, c_wp_occs_rel, c_sw_d_locvars_sim
L1b & L1b′	c_wpw, c_geonw, c_admloc, c_wp_occs, c_multi_m, c_sw_d_sw_sim, c_name_in_doc, c_iptc_d_iptc_sim, c_country_equals_d_country, c_wp_occs_rel, c_slugs_d_slugs_sim
L1	c_wpw, c_geonw, m_inlefff_nocap, c_wp_occs, c_admloc, c_iptc_d_iptc_sim, c_peers_vars_d_mentions_sim, m_inlefff, c_sw_d_sw_sim, c_multi_m, c_name_in_doc, c_country_equals_d_country
LN	c_wpw, c_geonw, c_wp_occs, m_inlefff_nocap, c_admloc, m_conf, c_wp_occs_rel, m_long, c_peers_vars_d_mentions_sim, c_country_equals_d_country, m_spos1, c_titlepar_in_m_suwt
SA	c_wpw, c_geonw, c_wp_occs_m, m_inlefff_nocap, c_admloc, c_name_in_doc, c_name_m_longer_sim, c_countrycode_equals_d_countrycode
LNSAU	c_wpw, c_geonw, c_wp_occs_m, m_inlefff_nocap, c_admloc, c_weightnull, m_long, c_name_in_doc, c_countrycode_equals_d_countrycode, m_multiner, m_inlefff, c_peers_vars_d_mentions_sim, m_short
LNSAI	c_wpw, c_geonw, c_admloc, c_name_in_doc, c_wp_occs_rel, c_iptc_d_iptc_sim

Table B.1 : Traits sélectionnés dans chaque configuration de Nomos.

Annexe C

Corpus Arboré de Paris 7

1 Corpus Arboré de Paris 7

En supplément au corpus GAFP, présenté au chapitre 5 (section 2.1), un second corpus de référence pour la tâche d'identification d'entités a été constitué à partir du contenu brut du Corpus Arboré de Paris 7 (ou French Treebank, [ACT03]), principalement destiné aux travaux en analyse syntaxique, tels que ceux de Candito et al. [Can+09]. Ce corpus, nommé GFTB dans le présent contexte, se distingue formellement du corpus GAFP dans la mesure où les articles extraits du journal *Le Monde* dont il se compose ne sont pas associés à la structuration et aux métadonnées caractérisant les dépêches de l'AFP. Ces articles ne présentent notamment pas de titre, mots-clés ou information de catégorisation thématique. Les articles composant ce corpus datent par ailleurs des années 1990 et présentent donc une distribution des entités différente, les mentions d'entités et leur fréquence étant largement dépendante de l'actualité. Le contenu de ces articles ainsi que le style rédactionnel employé peuvent être considérés comme relevant du même domaine que la production de l'AFP, même si les modalités de rédaction et de publication de l'information issue d'une agence de presse diffèrent du travail journalistique de la presse traditionnelle. Les informations relatives à la taille du corpus GFTB (695 documents) sont renseignées à la table C.1.

Phrases		**Tokens**	
#	**12 351**	**#**	**390 383**
# moy. / doc.	18	**# moy. / phr.**	32

TABLE C.1 : Description du corpus GFTB.

Annotation pour l'identification d'entités Le corpus GFTB, destiné de façon similaire au corpus GAFP à l'entraînement et l'évaluation du système d'identification automatique d'entités, a été constitué en tant que corpus de référence selon la même procédure générale : pre-annotation automatique puis revue manuelle par un annotateur humain. La production de ce corpus annoté est décrite dans Sagot et al. [SRS12]. Les balises d'annotation sont également nommées ENAMEX et présentent les mêmes attributs, dont des identifiants d'entités renvoyant à la ressource adoptée dans le présent travail. Les informations de distribution et de taux d'ambiguïté des mentions et entités, établies relativement à cette ressource, sont présentées dans les tables C.2 à C.5 avec un schéma identique aux données du corpus GAFP.

	moyenne	**min**	**max**
Mentions	17	1	132
connues	12		
inconnues	4		
Entités	4	1	61
connues	2		
inconnues	2		

TABLE C.2 : Distribution des mentions et entités par article dans le corpus GFTB.

Catégorisation thématique Un rapprochement des documents du corpus GFTB vers une forme comparable à celle des dépêches de l'AFP est réalisé afin de le rendre utilisable pour l'élaboration d'un système spécifique au traitement de documents tels qu'ils sont produits par l'Agence. Ce rapprochement est notamment possible pour ce qui concerne les informations ou métadonnées de catégorisation thématique : chaque article du corpus GFTB peut en effet être classifié selon

	Type	Connues	Inconnues	Total
Mentions	PERSON	1 245	780	2 025
	LOCATION	3 491	274	3 765
	ORGANIZATION	3 730	2 001	5 731
	Total	**8 466 (73%)**	**3 055 (27%)**	**11 521**
Entités	PERSON	418	476	894
	LOCATION	584	87	671
	ORGANIZATION	748	832	1 580
	Total	**1 750 (56%)**	**1 395 (44%)**	**3 145**

Table C.3 : Distribution des mentions et entités par type dans le corpus GFTB.

# Entités candidates	min	min \ 0	max	moy.	moy. \ 0
par mention distincte (toutes les entités)	0	1	118	1	2
par mention distincte (entités inconnues seulement)	0	1	113	5	0,82

Table C.4 : Taux minimaux (à partir de 0 et à partir de 1), maximaux et moyens d'ambiguïté entre candidats (entités de la ressource) par mention distincte dans le corpus GFTB.

# Entités candidates	0	1	> 1	≥ 1	Total
Mentions distinctes (toutes les entités)	1 934	1 345	657	2 002	
Mentions distinctes (entités inconnues seulement)	1 383	146	124	270	1 923

Table C.5 : Distribution des mentions distinctes selon le taux d'ambiguïté (nul, égal à 1, supérieur à 1) entre candidats (entités de la ressource) dans le corpus GFTB.

la taxonomie de l'IPTC (cf. chapitre 4, section 2.2 et annexe A), dont la couverture en termes de domaines est *a priori* aussi adaptée au journal *Le Monde* qu'à la production de l'AFP. Cette catégorisation est effectuée selon la méthode automatique décrite à l'annexe A (section 3), également appliquée aux articles de l'encyclopédie Wikipedia pour la constitution de la base de connaissances du système Nomos (cf. chapitre 5, section 2.2).

On obtient ainsi un corpus de référence non seulement similaire au corpus GAFP en termes de domaines couverts mais comportant également des informations de catégorisation selon la même taxonomie. La distribution des catégories IPTC au travers du corpus GFTB est présentée à la table C.6.

POL	**ECO**	**CLJ**	**CLT**	**UNR**	**HTH**	**SOI**	**ENV**
214	506	37	121	28	41	59	64
30,5%	**72,5%**	**5%**	**17%**	**4%**	**6%**	**8,5%**	**9%**
HUM	**SOC**	**SCI**	**REL**	**EDU**	**LIF**	**DIS**	**SPO**
26	215	60	7	28	55	17	3
3,5%	**31%**	**8,5%**	**1%**	**4%**	**8%**	**2%**	**0,5%**

Table C.6 : Distribution des catégories IPTC sur les 695 articles du corpus de référence GFTB.

Entités dans le corpus GFTB Comme cela a été évoqué au cours du présent travail, le problème de l'Annotation Sémantique et de l'identification des entités mentionnées dans les contenus textuels, de nature notamment journalistique, est étroitement associé à la constitution de ressources rendues disponibles sous la forme d'inventaires d'entités. Même dans le cas de ressources constituées à l'aide de larges dépôts encyclopédiques tels que Wikipedia, l'adéquation entre les corpus à traiter et les entités à identifier se pose en termes de couverture et de synchronisation. Les entités mentionnées dans un corpus données sont en effet liées à la date de rédaction, autrement dit à une forme d'actualité, de laquelle les ressources peuvent se distinguer en fonction de leur époque de constitution et de l'étendue à la fois temporelle et thématique des entités qu'elle recense. Dans le cas particulier du corpus GFTB, dont les contenus datent des années 1990 et dont la thématique dominante est de nature économique, on peut observer une distribution d'entités dont l'alignement avec la base d'entités Aleda, en partie dérivée de Wikipedia, est moindre en comparaison avec le corpus GAFP. En effet, 44% des entités qui y sont mentionnées sont absentes d'Aleda (27% des mentions), contre 27% pour le corpus GAFP (19% des mentions). La thématique économique du corpus GFTB semble participer de cette tendance dans la mesure où elle fait intervenir nombre de personnalités du milieu des affaires, sujettes à mention dans les articles du *Monde* concernant les entreprises et organisations au sein desquelles elles évoluent ; ces deux groupes d'entités — cadres d'entreprises ou organisations diverses, notamment non institutionnelles — donnent lieu à des biographies et articles descriptifs de façon moins systématique que les entités en lien avec la thématique culturelle, notamment, dans une ressource telle que Wikipedia (cf. distribution des catégories IPTC sur les articles Wikipedia donnant lieu à une entité Aleda, annexe A, table A.3).

Le taux d'entités mentionnées dans le corpus GFTB et absentes de la base Aleda, qui s'élève à 44%, permet dès avant la conduite d'expériences de constater que la tâche d'identification est, pour ce corpus, davantage déterminée par la capacité à reconnaître les mentions d'entités hors inventaire que par la discrimination à effectuer entre plusieurs entités candidates de la base pouvant être dénotées par une mention donnée.

2 Nomos : expériences avec le corpus GFTB

En complément des expériences menées sur le corpus GAFP pour le développement du système Nomos, nous présentons ici les résultats de travaux menés sur le corpus GFTB. Il s'agit de déterminer dans quelle mesure le système Nomos présente une capacité d'adaptation à un corpus distinct du corpus original de développement et d'identifier les problèmes rencontrés dans la tâche d'identification lors de cette opération de transposition.

Comme cela a été évoqué précédemment, un enjeu notable quant au traitement du corpus GFTB réside dans la forte proportion d'entités absentes de l'inventaire cible (Aleda). Il s'agit également d'évaluer l'efficacité de la méthode de sélection des entités candidates étant donné une mention à identifier, ce que les systèmes Nomos et NPNORMALIZER accomplissent par l'accès à un index pré-calculé recensant un ensemble figé de variantes possibles pour chaque entité d'Aleda. Nous avons pu mesurer ce taux de rappel sur le corpus GFTB : il atteint 74%, contre 88% sur le corpus GAFP (cf. chapitre 6, section 2.2.1), ce qui induit une correction maximale du Liage ne pouvant dépasser ces 74%. Ce constat permet de remettre en cause l'efficacité de cette méthode de sélection, qu'il apparaît utile de modifier dans les développements futurs du système Nomos.

Les expériences présentées ici ont été réalisées avec le même ensemble de traits (chapitre 6, table 6.3) que dans le cas du corpus GAFP. Deux aspects de développement du système Nomos ont été explorés :

- Les meilleurs modèles obtenus à la suite des développements (entraînements, tests et sélection de traits) sur le corpus GAFP ont été appliqués au corpus GFTB dans son ensemble, qui constitue alors un unique corpus de test.

- Une phase d'entraînement et de test a été menée sur le corpus GFTB lui-même, à l'aide d'une validation croisée où la proportion du test équivaut à 10% du corpus total, avec les deux ensembles de traits *baseline* : traits 1 à 39 de la table 6.3 au chapitre 6 (B^{all}), et uniquement les traits 1 et 2 (poids indiqué dans la base Aleda, B^2).
- Une seconde phase d'entraînement et de test a été menée sur le corpus GFTB avec les traits issus de la sélection par méthode gloutonne lors du développement de Nomos sur le corpus GAFP, chaque configuration employant l'ensemble de traits correspondant.

Les configurations de reconnaissance d'entités nommées (REN) retenues pour ces expériences sont les suivantes :

- REN Gold,
- REN automatique non déterministe de LIANE,
- REN automatique non déterministe avec intersection de LIANE et SxPipe/NP.

Les résultas des expériences menées sont présentés à la table C.7. Les résultats obtenus sur le corpus GAFP avec les meilleurs modèles développés sont rappelés à titre de comparaison.

Meilleurs modèles GAFP	P_{REN}	R_{REN}	F_{REN}	Acc	Acc_{BC}	Acc_{NIL}	P_{ALL}	R_{ALL}	F_{ALL}	P_{NAE}	R_{NAE}	F_{NAE}
Gold	100	100	100	75,97	73,48	83,94	75,97	75,97	75,97	-	-	-
Gold - GAFP	*100*	*100*	*100*	*84,72*	*83,64*	*89,35*	*84,72*	*84,72*	*84,72*	-	-	-
LN	77,48	61,63	68,65	83,37	85,13	72,21	64,59	51,38	57,23	68,35	66,70	67,51
LN - GAFP	*87,64*	*83,88*	*85,71*	*87,60*	*86,88*	*91,35*	*76,77*	*73,47*	*75,08*	*72,08*	*43,83*	*54,51*
LNSAI	84,33	68,25	75,44	82,56	85,81	58,91	69,62	56,35	62,28	-	0	-
LNSAI - GAFP	*92,95*	*72,82*	*81,66*	*91,96*	*93,17*	*80,91*	*85,48*	*66,97*	*75,10*	-	*0*	-

B^2	P_{REN}	R_{REN}	F_{REN}	Acc	Acc_{BC}	Acc_{NIL}	P_{ALL}	R_{ALL}	F_{ALL}	P_{NAE}	R_{NAE}	F_{NAE}
Gold	100	100	100	74,52	71,94	82,74	74,52	74,52	74,52	-	-	-
LN	60,93	80,49	69,36	75,20	74,57	77,72	45,82	60,53	52,16	-	0	-
LNSAI	84,22	68,17	75,35	78,72	81,44	58,97	66,30	53,67	59,32	-	0	-

B^{all}	P_{REN}	R_{REN}	F_{REN}	Acc	Acc_{BC}	Acc_{NIL}	P_{ALL}	R_{ALL}	F_{ALL}	P_{NAE}	R_{NAE}	F_{NAE}
Gold	100	100	100	77,28	74,91	84,82	77,28	77,28	77,28	-	-	-
LN	61,28	80,96	69,76	78,40	78,02	79,98	48,05	63,47	54,69	-	0	-
LNSAI	84,54	68,43	75,64	82,60	85,01	65,07	69,83	56,52	62,48	-	0	-
LN+NAE	89,25	52,62	66,20	83,47	87,04	08,11	74,50	43,92	55,26	70,73	89,03	78,83
LNSAI+NAE	90,07	59,31	71,52	84,79	88,17	18,27	76,37	50,28	60,64	41,63	49	45,01

Meilleurs traits GAFP	P_{REN}	R_{REN}	F_{REN}	Acc	Acc_{BC}	Acc_{NIL}	P_{ALL}	R_{ALL}	F_{ALL}	P_{NAE}	R_{NAE}	F_{NAE}
Gold	100	100	100	75,23	72,42	84,20	75,23	75,23	75,23	-	-	-
LN	83,53	63,34	72,05	75,73	74,51	80,73	63,26	47,97	54,56	80,03	78,29	79,15
LNSAI	84,64	68,51	75,73	81,43	84,48	59,13	68,93	55,79	61,66	-	0	-

TABLE C.7 : Résultats de Nomos sur le corpus GFTB.

Discussion On constate à l'aide de la table C.7 que les performances de Nomos sont bien moins satisfaisantes sur le corpus GFTB que sur le corpus GAFP. En ce qui concerne la correction du Liage avec la REN Gold, on note que le score relatif au Liage des entités présentes dans la base Aleda (Acc_{BC}) ne peut excéder les 74% du rappel des candidats, évoqués précédemment. À ce titre, ce score qui atteint de près de 76%, obtenu avec le meilleur modèle issu du développement sur le corpus GAFP, est relativement satisfaisant. La correction Acc_{NIL} qui concerne la capacité de Nomos à identifier une entité hors inventaire peut atteindre des score supérieurs à 80%, mais également tomber en-deçà des 60%, voire en-deçà des 10% sans sélection de traits. Ces scores reflètent les difficultés d'adaptation du système à un corpus où les cas de NIL sont particulièrement fréquents.

Dans toutes les configurations et avec tous les modèles, les performances générales de Nomos (F_{ALL}) sur le corpus GFTB sont inférieures à leurs équivalents sur le corpus GAFP, avec des différences allant de la dizaine à la vingtaine de points. On remarque que le score de rappel de REN (R_{REN}) est relativement bas dans la plupart des cas (entre 52% et 68%) et que, lorsqu'il atteint 80%, la précision (P_{REN}) est considérablement réduite. Ce taux de rappel bas est manifeste également lorsque l'élimination de mentions présente une précision satisfaisante (P_{NAE}) et avec des taux de correction (Acc_{BC} et Acc_{NIL}) plutôt bons, allant de 74 à 84%.

Le corpus GFTB semble ainsi plus difficile à traiter notamment en termes de rappel des mentions d'entités. Afin de vérifier que le faible rappel n'est pas uniquement lié à une élimination trop abrupte des mentions par le système Nomos lui-même, mais bien à une faiblesse de reconnaissance du module de REN, nous donnons à la table C.8 les résultats du système LIANE dans sa version déterministe au niveau de la REN seulement.

P_{REN}	R_{REN}	F_{REN}
83,09	64,60	72,69

TABLE C.8 : Résultats de LIANE en REN sur le corpus GFTB.

Ces résultats, et notamment la mesure de rappel de 64,60%, montrent que les mentions d'entités présentes dans le corpus GFTB sont, au niveau de la reconnaissance, plus difficiles à collecter que celles du corpus GFTB. Dans la configuration jointe du système Nomos, les résultats finaux, c'est-à-dire après l'accomplissement du Liage, ne peuvent donc pas dépasser cet ordre de performance.

On remarque en revanche que la fonctionnalité de repérage des faux positifs est relativement efficace sur le corpus GFTB avec la REN de LIANE : elle avoisine les 80% (F_{NAE}), quand le même score atteignait 57% dans le meilleur cas avec le corpus GAFP.

Le système NPNORMALIZER paraît plus robuste devant le changement de corpus : même si ses performances restent en-deçà de celles réalisées sur le corpus GAFP, elles dépassent les différentes configurations de Nomos expérimentées ici, comme le montre la table C.9.

P_{REN}	R_{REN}	F_{REN}	Acc	Acc_{BC}	Acc_{NIL}	P_{ALL}	R_{ALL}	F_{ALL}
83,83	69,91	76,24	83,98	86,25	67,07	70,40	58,71	64,02

TABLE C.9 : Résultats du système NPNORMALIZER sur le corpus GFTB.

La liste présentée en C.10 donne un aperçu des mentions d'entités non repérées par le module de REN dans les différentes configurations de Nomos.

Banque de France	Antenne 2	TF1
Sécurité sociale	Amérique latine	Bourse de Paris
Communauté européenne	Commission de Bruxelles	Nouvel économiste
A2	Sema group	Générale des eaux
Assemblée nationale	Amérique centrale	Société générale
Réserve fédérale	M6	Club de Paris
Banque mondiale	Banque fédérale d'Allemagne	Vieux-Colombier
Rhône-Poulenc	Elf Aquitaine	Crédit local de France
Conseil des Bourses de valeurs	BTF Gmb.H	

TABLE C.10 : Exemples de mentions d'entités non reconnues par le module de REN de Nomos.

Références

[AC75] A. V. Aho et M. J. Corasick. “Efficient string matching : an aid to bibliographic search”. In *Communications of the ACM* 18.6 (1975) [103].

[ACT03] A. Abeillé, L. Clément et F. Toussenel. “Building a treebank for French”. In *Treebanks*. Sous la dir. d’A. Abeillé. Kluwer, Dordrecht, 2003 [169, 243, 300].

[Age10] Agence France-Presse. *Manuel de l’agencier*. 2010 [135].

[AH09] G. Antoniou et F. van Harmelen. “Web Ontology Language : OWL”. In *Handbook on ontologies*. Springer, 2009 [32].

[BB98] A. Bagga et B. Baldwin. “Entity-based cross-document coreferencing using the vector space model”. In *Proceedings of the 17th international conference on Computational linguistics*. Vol. 1. Association for Computational Linguistics. 1998 [116].

[BC08] P. Buitelaar et P. Cimiano. *Ontology learning and population : bridging the gap between text and knowledge*. Ios Press, 2008 [33].

[BC10] F. Béchet et E. Charton. “Unsupervised knowledge acquisition for extracting named entities from speech”. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE. 2010 [65, 182].

[BD88] P. Boullier et P. Deschamp. *Le système Syntax™ – Manuel d’utilisation et de mise en œuvre sous Unix™*. En ligne sur `http://syntax.gforge.inria.fr/syntax3.8-manual.pdf`. 1988 [180].

[BFL98] C. F. Baker, C. J. Fillmore et J. B. Lowe. “The berkeley framenet project”. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*. Vol. 1. Association for Computational Linguistics. 1998 [55].

[Biz+09] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak et S. Hellmann. “DBpedia-A crystallization point for the Web of Data”. In *Web Semantics : Science, Services and Agents on the World Wide Web* 7.3 (2009) [86, 91].

[BLHL01] T. Berners-Lee, J. Hendler et O. Lassila. “The semantic web”. In *Scientific american* 284.5 (2001) [12, 25, 36].

[BO07] C. Brewster et K. O’Hara. “Knowledge representation with ontologies : Present challenges—Future possibilities”. In *International Journal of Human-Computer Studies* 65.7 (2007) [31].

[Bor+98] A. Borthwick, J. Sterling, E. Agichtein et R. Grishman. “Exploiting diverse knowledge sources via maximum entropy in named entity recognition”. In *Proc. of the Sixth Workshop on Very Large Corpora*. 1998 [182].

[BP06] R. Bunescu et M. Pasca. “Using encyclopedic knowledge for named entity disambiguation”. In *Proceedings of EACL*. Vol. 6. 2006 [108, 118, 120].

[BR04] D. Bean et E. Riloff. “Unsupervised learning of contextual role knowledge for coreference resolution”. In *Proc. of HLT/NAACL*. 2004 [71].

[BSS11] F. Béchet, B. Sagot et R. Stern. “Coopération de méthodes statistiques et symboliques pour l’adaptation non-supervisée d’un système d’étiquetage en entités nommées”. In *TALN’2011 - Traitement Automatique des Langues Naturelles*. 2011 [178].

[BSW99] D. M. Bikel, R. Schwartz et R. M. Weischedel. “An algorithm that learns what’s in a name”. In *Machine learning* 34.1 (1999) [182].

[BT06] D. G. Brizan et A. U. Tansel. “A survey of entity resolution and record linkage methodologies”. In *Communications of the IIMA* 6.3 (2006) [74].

[Bui+08] P. Buitelaar, P. Cimiano, A. Frank, M. Hartung et S. Racioppa. “Ontology-based information extraction and integration from heterogeneous data sources”. In *International Journal of Human-Computer Studies* 66.11 (2008) [61].

[Bus45] V. Bush. “As we may think”. In 176 (1945) [42].

[Can+09] M. Candito, B. Crabbé, P. Denis, F. Guérin et al. “Analyse syntaxique du français : des constituants aux dépendances”. In *Actes de la 16e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2009)*. 2009 [300].

[Cao+07] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai et H. Li. “Learning to rank : from pairwise approach to listwise approach”. In *Proceedings of the 24th international conference on Machine learning*. ACM. 2007 [122].

[CG12] E. Charton et M. Gagnon. “A disambiguation resource extracted from Wikipedia for semantic annotation”. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*. European Language Resources Association (ELRA), 2012 [91, 93].

[CGO11] E. Charton, M. Gagnon et B. Ozell. “Automatic semantic web annotation of named entities”. In *Advances in Artificial Intelligence* (2011) [99, 104–106, 154].

[Cha+99] J.-C. Chappelier, M. Rajman, R. Aragüés et A. Rozenknop. “Lattice parsing for speech recognition”. In *Proc. of 6ème conférence sur le Traitement Automatique du Langage Naturel (TALN 99)*. 1999 [163].

[CHS04] P. Cimiano, S. Handschuh et S. Staab. “Towards the self-annotating web”. In *Proceedings of the 13th international conference on World Wide Web*. ACM. 2004 [98, 99].

[Cir+04] F. Ciravegna, S.Chapman, A. Dingli et Y. Wilks. “Learning to harvest information for the semantic web”. In *The Semantic Web : Research and Applications* (2004) [97, 98].

[CJ11] Z. Chen et H. Ji. “Collaborative ranking : A case study on entity linking”. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. 2011 [124].

[CTM10] E. Charton et J. Torres-Moreno. “NLGbAse : a free linguistic resource for Natural Language Processing systems”. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*. European Language Resources Association (ELRA), 2010 [91].

[Cuc07] S. Cucerzan. “Large-scale named entity disambiguation based on Wikipedia data”. In *Proceedings of EMNLP-CoNLL*. Vol. 6. 2007 [108, 118, 120].

[Cuc11] S. Cucerzan. “TAC entity linking by performing full-document entity extraction and disambiguation”. In *Proc. of TAC* (2011) [124].

[Cun+11a] H. Cunningham, V. Tablan, I. Roberts, M. Greenwood et N. Aswani. “Information Extraction and Semantic Annotation for Multi-Paradigm Information Management”. In *Current Challenges in Patent Information Retrieval* (2011) [98].

[Cun+11b] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Damljanovic, T. Heitz, M. A. Greenwood, H. Saggion, J. Petrak, Y. Li et W. Peters. *Text Processing with GATE (Version 6)*. 2011 [59, 65, 98, 100, 129].

[CW00] J. Cowie et Y. Wilks. “Information extraction”. In *Handbook of Natural Language Processing* (2000) [56, 57].

[DB09] P. Denis et J. Baldridge. “Global joint models for coreference resolution and named entity classification”. In *Procesamiento del Lenguaje Natural* 42 (2009) [158].

[DeJ77] G. DeJong. “Skimming newspaper stories by computer”. In *Proceedings of the 5th international joint conference on Artificial intelligence*. Vol. 1. Morgan Kaufmann Publishers Inc. 1977 [55].

[DeJ82] G. DeJong. “An overview of the {FRUMP} system”. In (1982) [55].

[Den07] P. Denis. “New learning models for robust reference resolution”. Thèse de doctorat. University of Texas at Austin, 2007 [71].

[DFH11] J. Domingue, D. Fensel et J. A. Hendler. *Handbook of semantic web technologies*. Vol. 1. Springer, 2011 [79].

[DI04] H. Daumé III. “Notes on CG and LM-BFGS optimization of logistic regression”. In *Paper available at http ://pub. hal3. name# daume04cg-bfgs, implementation available at http ://hal3. name/megam* (2004) [198, 282].

[Dil+03] S. Dill, N. Eiron, D. Gibson, D. Gruhl, R. Guha, A. Jhingran, T. Kanungo, S. Rajagopalan, A. Tomkins et J. Tomlin. “SemTag and Seeker : Bootstrapping the semantic web via automated semantic annotation”. In *Proceedings of the 12th international conference on World Wide Web*. ACM. 2003 [97–99].

[Dod+04] G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel et R. Weischedel. “The automatic content extraction (ACE) program–tasks, data, and evaluation”. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC’04)*. Vol. 4. European Language Resources Association (ELRA), 2004 [63, 110].

[Dre+10] M. Dredze, P. McNamee, D. Rao, A. Gerber et T. Finin. “Entity Disambiguation for Knowledge Base Population”. In *Proceedings of the 23rd International Conference on Computational Linguistics*. 2010 [116, 122, 123].

[DSS10] L. Danlos, B. Sagot et R. Stern. “Analyse discursive des incises de citation”. In *2ème Congrès Mondial de Linguistique Française - CMLF 2010*. 2010 [241].

[DSS93] R. Davis, H. Shrobe et P. Szolovits. “What is a knowledge representation ?” In *AI magazine* 14.1 (1993) [31].

[Ehr08] M. Ehrmann. “Les entités nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation”. Thèse de doctorat. Université Paris 7 Denis Diderot, 2008 [56, 60, 63–65, 67–69, 252].

[Erd+00] M. Erdmann, A. Maedche, H. Schnurr et S. Staab. “From manual to semi-automatic semantic annotation : About ontology-based text annotation tools”. In *Proceedings of the COLING 2000 Workshop on Semantic Annotation and Intelligent Content*. Vol. 10. 2000 [97].

[FGM05] J. R. Finkel, T. Grenager et C. Manning. “Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling”. In (2005) [66, 119].

[FH02] M. Fleischman et E. Hovy. “Fine grained classification of named entities”. In *Proceedings of the 19th international conference on Computational linguistics*. Vol. 1. Association for Computational Linguistics. 2002 [69].

[FI71] G. Frege et C. Imbert. *Écrits logiques et philosophiques*. Éditions du Seuil, 1971 [68].

[Fre92] G. Frege. “Ausführungen über Sinn und Bedeutung”. In *Nachgelassene Schriften* (1892) [68].

[GCY92] W. A. Gale, K. W. Church et D. Yarowsky. “One sense per discourse”. In *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics. 1992 [119].

[GGC09] S. Galliano, G. Gravier et L. Chaubard. “The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts”. In *Interspeech 2009* (2009) [66].

[GHC98] N. Ge, J. Hale et E. Charniak. “A statistical approach to anaphora resolution”. In *Proceedings of the Sixth Workshop on Very Large Corpora*. 1998 [71].

[GJ11] S. Gottipati et J. Jiang. “Linking entities to a knowledge base with query expansion”. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. 2011 [118, 119].

[GN87] M. R. Genesereth et N. J. Nilsson. *Logical foundations of artificial intelligence*. Vol. 9. Morgan Kaufmann Los Altos, CA, 1987 [30].

[GOS09] N. Guarino, D. Oberle et S. Staab. “What is an ontology ?” In *Handbook on ontologies*. Springer, 2009 [30].

[Gra+08] B. C. Grau, I. Horrocks, B. Motik, B. Parsia, P. Patel-Schneider et U. Sattler. “OWL 2 : The next step for OWL”. In *Web Semantics : Science, Services and Agents on the World Wide Web* 6.4 (2008) [55].

[Gri12] R. Grishman. “Information Extraction : Capabilities and Challenges”. In *Notes prepared for the 2012 International Winter School in Language and Speech Technologie* (2012) [50, 56, 61].

[Gro81] M. Gross. “Les bases empiriques de la notion de prédicat sémantique”. In *Langages* 15.63 (1981) [238].

[GS96] R. Grishman et B. Sundheim. “Message understanding conference-6 : A brief history”. In *Proceedings of COLING*. Vol. 96. 1996 [16, 56, 63].

[Har54] Z Harris. *Distributional structure.(J. Katz, Ed.) Word Journal Of The International Linguistic Association, 10 (23), 146-162*. 1954 [114].

[Har58] Z. S. Harris. “Linguistic transformations for information retrieval”. In *Proceedings of the International Conference on Scientific Information*. Vol. 2. 1958 [54].

[HB11] T. Heath et C. Bizer. “Linked data : Evolving the web into a global data space”. In *Synthesis Lectures on the Semantic Web : Theory and Technology* 1.1 (2011) [34].

[Hir98] L. Hirschman. “The Evolution of Evaluation : Lessons from the Message Understanding Conferences”. In (1998) [56].

[HK10] A. Haghighi et D. Klein. “Coreference resolution in a modular, entity-centered model”. In *Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics. 2010 [71].

[Hob86] J. Hobbs. “Resolving pronoun references”. In *Readings in natural language processing*. Morgan Kaufmann Publishers Inc. 1986 [71].

[Hoe+07] R. Hoekstra, J. Breuker, M. D. Bello et A. Boer. “The LKIF Core ontology of basic legal concepts”. In (2007). Sous la dir. de P. Casanovas, M. A. Biasiotti, E. Francesconi et M. T. Sagri [33].

[Hof+11] J. Hoffart, M. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater et G. Weikum. “Robust disambiguation of named entities in text”. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. 2011 [118, 124, 202].

[Hor08] I. Horrocks. “Ontologies and the semantic web”. In *Communications of the ACM* 51.12 (2008) [24, 25, 27, 30].

[HS11] X. Han et L. Sun. “A generative entity-mention model for linking entities with knowledge base”. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*. Vol. 1. Association for Computational Linguistics. 2011 [118, 123].

[HSZ11] X. Han, L. Sun et J. Zhao. “Collective entity linking in web text : a graph-based method”. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM. 2011 [124, 202].

[HZ10] X. Han et J. Zhao. “Structural semantic relatedness : a knowledge-based method to named entity disambiguation”. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. 2010 [121].

[IV98] N. Ide et J. Véronis. “Introduction to the special issue on word sense disambiguation : the state of the art”. In *Computational linguistics* 24.1 (1998) [64, 117].

[JGD11] H. Ji, R. Grishman et H. T. Dang. “An Overview of the TAC2011 Knowledge Base Population Track”. In *Proceedings of Text Analysis Conference (TAC2011)*. 2011 [17, 111, 113, 118, 120, 124, 154, 202, 208].

[Ji+10] H. Ji, R. Grishman, H. Dang, K. Griffitt et J. Ellis. “Overview of the TAC 2010 knowledge base population track”. In *Proceedings of the Third Text Analysis Conference*. 2010 [17, 111, 113, 117, 120, 123, 154].

[Jij+08] V. Jijkoun, M. A. Khalid, M. Marx et M. De Rijke. “Named entity normalization in user generated content”. In *Proceedings of the second workshop on Analytics for noisy unstructured text data*. ACM. 2008 [73].

[Joa06] T. Joachims. “Training Linear SVMs in Linear Time”. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*. 2006 [122].

[Kir+04] A. Kiryakov, B. Popov, I. Terziev, D. Manov et D. Ognyanoff. “Semantic annotation, indexing, and retrieval”. In *Web Semantics : Science, Services and Agents on the World Wide Web* 2.1 (2004) [96–102, 106].

[Kiy+09] N. Kiyavitskaya, N. Zeni, J. Cordy, L. Mich et J. Mylopoulos. “Cerno : Light-weight tool support for semantic annotation of textual documents”. In *Data & Knowledge Engineering* 68.12 (2009) [98].

[KJDR08] M. A. Khalid, V. Jijkoun et M. De Rijke. “The impact of named entity normalization on information retrieval for question answering”. In *Advances in Information Retrieval*. Springer, 2008 [73].

[Kob+09] G. Kobilarov, T. Scott, Y. Raimond, S. Oliver, C. Sizemore, M. Smethurst, C. Bizer et R. Lee. “Media Meets Semantic Web–How the BBC Uses DBpedia and Linked Data to Make Connections”. In *The Semantic Web : Research and Applications* (2009) [12, 128].

[KR11] Z. Kozareva et S. Ravi. “Unsupervised name ambiguity resolution using a generative model”. In *Proceedings of the First Workshop on Unsupervised Learning in NLP*. Association for Computational Linguistics. 2011 [123].

[LC05] E. V. de La Clergerie. “From metagrammars to factorized TAG/TIG parsers”. In *Proceedings of the Ninth International Workshop on Parsing Technology*. Association for Computational Linguistics. 2005 [163, 240].

[LC08] B. Lamiroy et M. Charolles. “Les verbes de parole et la question de l’(in) transitivité”. In *Discours. Revue de linguistique, psycholinguistique et informatique* 2 (2008) [238].

[Leh82] W. G. Lehnert. “Plot units : A narrative summarization strategy”. In (1982). Sous la dir. de W. G. Lehnert et M. H. Ringle [55].

[Li+02] H. Li, R. K. Srihari, C. Niu et W. Li. “Location normalization for information extraction”. In *Proceedings of the 19th international conference on Computational linguistics*. Vol. 1. Association for Computational Linguistics. 2002 [73].

[Li+09] F. Li, Z. Zheng, F. Bu, Y. Tang, X. Zhu et M. Huang. “Thu quanta at tac 2009 kbp and rte track”. In *Proceedings of Test Analysis Conference 2009 (TAC 09)*. 2009 [122, 123].

[Lil1] H. Li. “Learning to rank for information retrieval and natural language processing”. In *Synthesis Lectures on Human Language Technologies* 4.1 (2011) [121].

[Lin98] D. Lin. “Using collocation statistics in information extraction”. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. 1998 [66].

[Liu+12] X. Liu, M. Zhou, F. Wei, Z. Fu et X. Zhou. “Joint inference of named entity recognition and normalization for tweets”. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics : Long Papers*. Vol. 1. Association for Computational Linguistics. 2012 [73].

[LL94] S. Lappin et H. J. Leass. “An algorithm for pronominal anaphora resolution”. In *Computational linguistics* 20.4 (1994) [71].

[LMP01] J. Lafferty, A. McCallum et F. C. Pereira. “Conditional random fields : Probabilistic models for segmenting and labeling sequence data”. In (2001) [65, 104, 243].

[Mai07] E. Maier. “Mixed quotation : between use and mention”. In *Proceedings of LENLS2007* (2007) [240].

[Man07] C. Mangold. “A survey and classification of semantic search approaches”. In *International Journal of Metadata, Semantics and Ontologies* 2.1 (2007) [44, 233].

[MBC03] D. Maynard, D. Bontcheva et D. Cunningham. “Towards a semantic extraction of named entities”. In (2003) [63].

[McD96] D. McDonald. “Internal and external evidence in the identification and semantic categorization of proper names”. In *Corpus processing for lexical acquisition* (1996) [65].

[McN+10] P. McNamee, H. Dang, H. Simpson, P. Schone et S. Strassel. “An evaluation of technologies for knowledge base population”. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*. European Language Resources Association (ELRA), 2010 [120].

[MD09] P. McNamee et H. T. Dang. “Overview of the TAC 2009 knowledge base population track”. In *Text Analysis Conference (TAC)*. 2009 [17, 111, 113, 116, 154, 167].

[Men+11a] P. Mendes, M. Jakob, A. García-Silva et C. Bizer. “Dbpedia spotlight : Shedding light on the web of documents”. In *Proceedings of the 7th International Conference on Semantic Systems*. ACM. 2011 [98, 99, 102, 118, 120].

[Men+11b] P. Mendes, J. Daiber, M. Jakob et C. Bizer. “Evaluating dbpedia spotlight for the tac-kbp entity linking task”. In *Proceedings of the TACKBP 2011 Workshop*. 2011 [116, 118, 120].

[MH02] K. Markert et U. Hahn. “Understanding metonymies in discourse”. In *Artificial Intelligence* 135.1 (2002) [252].

[Min74] M. Minsky. “A framework for representing knowledge”. In (1974). Sous la dir. de P. Winston [31].

[Mit02] R. Mitkov. *Anaphora resolution*. Vol. 134. Longman London, 2002 [71].

[MK12] W. Maass et T. Kowatsch. *Semantic Technologies in Content Management Systems : Trends, Applications and Evaluations*. Springer, 2012 [46, 128, 130, 131].

[ML03] A. McCallum et W. Li. “Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons”. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*. Vol. 4. Association for Computational Linguistics. 2003 [182].

[MN03] K. Markert et M. Nissim. “Corpus-based metonymy analysis”. In *Metaphor and symbol* 18.3 (2003) [252].

[MN06] K. Markert et M. Nissim. “Metonymic proper names : A corpus-based account”. In *TRENDS IN LINGUISTICS STUDIES AND MONOGRAPHS* 171 (2006) [252].

[MN07a] K. Markert et M. Nissim. “Metonymy resolution at semeval i : Guidelines for participants”. In *Rapport technique, SemEval* (2007) [252].

[MN07b] K. Markert et M. Nissim. “Semeval-2007 task 08 : Metonymy resolution at semeval-2007”. In *Proceedings of the 4th International Workshop on Semantic Evaluations*. Association for Computational Linguistics. 2007 [252].

[MOC96] R. Merchant, M. Okurowski et N. Chinchor. “The multilingual entity task (MET) overview”. In (1996) [63].

[Moe06] M. Moens. *Information extraction : algorithms and prospects in a retrieval context*. Springer, 2006 [51, 55, 58, 59].

[MP98] E. Marsh et D. Perzanowski. “MUC-7 evaluation of IE technology : Overview of results”. In 20 (1998) [66].

[MW04] *Models of Identity Uncertainty with Application to Noun Coreference*. 2004 [157].

[Nas+11] A. Nasr, F. Béchet, J.-F. Rey, B. Favre et J. Le Roux. “MACAON : An NLP tool suite for processing word lattices”. In *Proceedings of the ACL 2011 System Demonstration* (2011) [163].

[Nel65] T. H. Nelson. “A File Structure for the Complex, the Changing and the indeterminate”. In (1965) [42].

[Ng08] V. Ng. “Unsupervised models for coreference resolution”. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. 2008 [71].

[Ng10] V. Ng. “Supervised noun phrase coreference research : The first fifteen years”. In *Proceedings of the 48th annual meeting of the association for computational linguistics*. Association for Computational Linguistics. 2010 [71].

[NNB09] C. Nédellec, A. Nazarenko et R. Bossy. “Information Extraction”. In *Handbook on ontologies*. Springer, 2009 [61].

[Nou12] D. Nouvel. “Reconnaissance des entités nommées par exploration de règles d’annotation - Interpréter les marqueurs d’annotation comme instructions de structuration locale”. Thèse de doctorat. Université François Rabelais - Tours, 2012 [66].

[OGS09] D. Oberle, S. Grimm et S. Staab. “An Ontology for Software”. In *Handbook on ontologies*. Springer, 2009 [33].

[Otl34] P. Otlet. *Traité de documentation : le livre sur le livre, théorie et pratique*. Editions Mundaneum, Bruxelles, 1934 [41, 42].

[Pag+99] L. Page, S. Brin, R. Motwani et T. Winograd. “The PageRank citation ranking : bringing order to the web.” In *Technical Report. Stanford InfoLab* (1999) [24].

[Ped+06] T. Pedersen, A. Kulkarni, R. Angheluta, Z. Kozareva et T. Solorio. “An unsupervised language independent method of name discrimination using second order co-occurrence features”. In *Computational Linguistics and Intelligent Text Processing*. Springer, 2006 [72].

[Plo11] D. Ploch. “Exploring entity relations for named entity disambiguation”. In *ACL HLT 2011* (2011) [118, 123].

[PN99] T. Poibeau et A. Nazarenko. “L’extraction d’information, une nouvelle conception de la compréhension de texte ?” In *TAL. Traitement automatique des langues* 40.2 (1999) [55, 57].

[Pop+03] B. Popov, A. Kiryakov, A. Kirilov, D. Manov, D. Ognyanoff et M. Goranov. “KIM–semantic annotation platform”. In *The Semantic Web-ISWC 2003* (2003) [96, 100].

[PP04] A. Purandare et T. Pedersen. “Word sense discrimination by clustering contexts in vector and similarity spaces”. In *Proceedings of the Conference on Computational Natural Language Learning*. 2004 [72].

[PP09] A. Pilz et G. Paaß. “Named entity resolution using automatically extracted semantic information”. In *Proceedings of LWA*. 2009 [74, 123].

[PPK05] T. Pedersen, A. Purandare et A. Kulkarni. “Name discrimination by clustering similar contexts”. In *Computational Linguistics and Intelligent Text Processing*. Springer, 2005 [72].

[Qui80] W. V. O. Quine. *From a Logical Point of View*. Harvard Univ. Press, 1980 [31].

[Rat+11] L. Ratinov, D. Roth, D. Downey et M. Anderson. “Local and global algorithms for disambiguation to wikipedia”. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics (ACL)*. 2011 [118, 120, 124, 202].

[RFC1738] T. Berners-Lee, L. Masinter et M. McCahill. *Uniform Resource Locators (URL)*. 1994 [24].

[RFC866] B.-L. Tim et C. Dan. *Hypertext markup Language - 2.0*. 1996 [24].

[RH05] L. Reeve et H. Han. “Survey of semantic annotation platforms”. In *Proceedings of the 2005 ACM symposium on Applied computing*. ACM. 2005 [97, 98].

[RM08] C. Rosse et J. L. Mejino. “The foundational model of anatomy ontology”. In *Anatomy Ontologies for Bioinformatics* (2008) [33].

[Ros02] L. Rosier. “La presse et les modalités du discours rapporté : l’effet d’hyperréalisme du discours direct surmarqué”. In *L’information grammaticale* 94.1 (2002) [239].

[Ros08] L. Rosier. *Le discours rapporté en français*. Editions OPHRYS, 2008 [238].

[Ros+11] S. Rosset, C. Grouin, O. Galibert, P. Zweigenbaum, K. Fort et L. Quintard. “Les entités nommées dans le programme Quaero : Proposition pour une extension de la définition des EN, de la définition à l’évaluation”. In (2011) [63].

[Rum75] D. E. Rumelhart. “Notes on a schema for stories”. In (1975). Sous la dir. de D. Bobrow et A. Collins [55].

[Rum77] D. E. Rumelhart. *Introduction to human information processing*. Wiley Chichester, 1977 [55].

[SA77] R. C. Schank et R. P. Abelson. “Scripts, plans, goals and understanding : An inquiry into human knowledge structures.” In (1977) [54].

[Sab90] G. Sabah. “L’intelligence artificielle et le langage”. In (1990) [54].

[Sag10] B. Sagot. “The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French”. In (2010) [180, 240, 295].

[Sag81] N. Sager. *Natural language information processing*. Addison-Wesley Publishing Company, Advanced Book Program, 1981 [55].

[SB08] B. Sagot et P. Boullier. “SxPipe 2 : architecture pour le traitement pré-syntaxique de corpus bruts”. In *Traitement Automatique des Langues* 49.2 (2008) [163, 179].

[SBF98] R. Studer, V. R. Benjamins et D. Fensel. “Knowledge engineering : principles and methods”. In *Data & knowledge engineering* 25.1 (1998) [30].

[Sch72] R. C. Schank. “Conceptual dependence : A theory of natural language understanding.” In *Cognitive Psychology ; Cognitive Psychology* (1972) [54].

[Sch98] H. Schütze. “Automatic word sense discrimination”. In *Computational linguistics* 24.1 (1998) [72].

[SDS10] B. Sagot, L. Danlos et R. Stern. “A Lexicon of French Quotation Verbs for Automatic Quotation Extraction”. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*. European Language Resources Association (ELRA), 2010 [241].

[Sha+11] I. Shafran, R. Sproat, M. Yarmohammadi et B. Roark. “Efficient determinization of tagged word lattices using categorial and lexicographic semirings”. In *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. IEEE. 2011 [163].

[SHBL06] N. Shadbolt, W. Hall et T. Berners-Lee. “The semantic web revisited”. In *Intelligent Systems, IEEE* 21.3 (2006) [26].

[She65] J. Shera. *Libraries and the organization of knowledge*. C. Lockwood, London, 1965 [42].

[SI99] S. Sekine et H. Isahara. “IREX project overview”. In (1999) [66, 67].

[Sim+10] H. Simpson, S. Strassel, R. Parker et P. McNamee. “Wikipedia and the web of confusable entities : Experience from entity linking query creation for tac 2009 knowledge base population”. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*. Citeseer. European Language Resources Association (ELRA), 2010 [115, 116].

[SRS12] B. Sagot, M. Richard et R. Stern. “Annotation référentielle du Corpus Arboré de Paris 7 en entités nommées”. In *Traitement Automatique des Langues Naturelles (TALN)*. 2012 [300].

[SSl10] R. Stern et B. Sagot. “Resources for Named Entity Recognition and Resolution in News Wires”. In *Entity 2010 Workshop at LREC'10*. 2010 [183].

[SSl12a] B. Sagot et R. Stern. “Aleda, a free large-scale entity database for French”. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (2012) [93, 94].

[SSl12b] R. Stern et B. Sagot. “Population of a Knowledge Base for News Metadata from Unstructured Text and Web Data”. In *AKBC-WEKEX 2012 - Proceedings of the Knowledge Extraction Workshop at NAACL-HLT 2012*. 2012 [224].

[SSN02] S. Sekine, K. Sudo et C. Nobata. “Extended named entity hierarchy”. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*. Vol. 2. European Language Resources Association (ELRA), 2002 [68].

[Sto+09] V. Stoyanov, N. Gilbert, C. Cardie et E. Riloff. “Conundrums in noun phrase coreference resolution : Making sense of the state-of-the-art”. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Vol. 2. Association for Computational Linguistics. 2009 [156].

[Sun95] B. Sundheim. “Overview of results of the MUC-6 evaluation”. In (1995) [63].

[Sur+08] M. Surdeanu, R. Johansson, A. Meyers, L. Màrquez et J. Nivre. “The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies”. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*. Association for Computational Linguistics. 2008 [106].

[TC+10] Z. C. Taylor Cassidy, J. Artiles, H. Ji, H. Deng, L.-A. Ratinov, J. Zheng, J. Han et D. Roth. “CUNY-UIUC-SRI TAC-KBP2011 Entity Linking System Description”. In *Proc. Text Analysis Conference (TAC2011)*. 2010 [119, 120].

[Tét02] J.-F. Tétu. “Les stratégies de la citation dans la presse”. In *Citation et détournement* (2002) [236].

[TKS02] E. Tjong Kim Sang. “Introduction to the CoNLL- 2002 shared task : language-independent named entity recognition”. In (2002) [63].

[TKSDM03] E. F. Tjong Kim Sang et F. De Meulder. “Introduction to the CoNLL-2003 shared task : Language-independent named entity recognition”. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*. Vol. 4. Association for Computational Linguistics. 2003 [63].

[Ure+06] V. Uren, P. Cimiano, J. Iria, S. Handschuh, M. Vargas-Vera, E. Motta et F. Ciravegna. “Semantic annotation for knowledge management : Requirements and a survey of the state of the art”. In *Web Semantics : science, services and agents on the World Wide Web* 4.1 (2006) [97, 98].

[VR02] D. Van Raemdonck. “Discours rapporté et frontières de phrase : l'épreuve de l'intégration syntaxique”. In *Faits de langues* 19 (2002) [238].

[WB09] Y. Wilks et C. Brewster. “Natural language processing as a foundation of the semantic web”. In *Foundations and Trends in Web Science* 1.3-4 (2009) [15, 36, 38, 50, 51, 53, 97].

[Wei07] D. Weinberger. *Everything is miscellaneous : The power of the new digital disorder*. Times Books, 2007 [41].

[Wil08] Y. Wilks. “The semantic web : Apotheosis of annotation, but what are its semantics ?” In *Intelligent Systems, IEEE* 23.3 (2008) [15, 36, 38].

[Woo10] D. Wood. *Linking Enterprise Data*. Springer, 2010 [12, 15, 41, 47, 131–133].

[Yan+03] X. Yang, G. Zhou, J. Su et C. L. Tan. “Coreference resolution using competition learning approach”. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*. Vol. 1. Association for Computational Linguistics. 2003 [157].

[Zha+10] W. Zhang, J. Su, C. Tan et W. Wang. “Entity linking leveraging : automatically generated annotation”. In *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics. 2010 [118].

[Zha+11] W. Zhang, Y. Sim, J. Su et C. Tan. “Entity linking with effective acronym expansion, instance selection and topic modeling”. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence*. Vol. 3. AAAI Press. 2011 [118].

[Zhe+10] Z. Zheng, F. Li, M. Huang et X. Zhu. “Learning to Link Entities with Knowledge Base”. In *Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 2010 [121–123].