Modélisation d'une ontologie de domaine et des outils d'extraction de l'information associés pour l'anglais et le français

Laurie Serrano

Modélisation d'une ontologie de domaine et des outils d'extraction de l'information associés pour l'anglais et le français

Nom : SERRANO

Prénom : Laurie

UFR des Sciences du Langage

Mémoire de master 2 professionnel - 20 ECTS

Mention : Sciences du Langage

Spécialité : Modélisation et traitements en industries de la langue : parole écrit apprentissage

Parcours : Traitement automatique de la parole et de la langue écrite (TALEP)

Sous la direction de M. Olivier Kraif

Année universitaire 2009-2010

Modélisation d'une ontologie de domaine et des outils d'extraction de l'information associés pour l'anglais et le français

Nom : SERRANO

Prénom : Laurie

UFR des Sciences du Langage

Mémoire de master 2 professionnel - 20 ECTS

Mention : Sciences du Langage

Spécialité : Modélisation et traitements en industries de la langue : parole écrit apprentissage

Parcours : Traitement automatique de la parole et de la langue écrite (TALEP)

Sous la direction de M. Olivier Kraif

Année universitaire 2009-2010

Remerciements

Je remercie avant tout Stephan Brunessaux pour m'avoir accueillie au sein du département IPCC et me permettre de continuer l'aventure en thèse. Merci à mes encadrants Bruno Grilheres et Olivier Kraif pour leur implication et leurs conseils tout au long de ce stage.

Je suis particulièrement reconnaissante à Yann Mombrun et Khaled Khelif pour l'intérêt qu'ils ont porté à mes travaux et le temps qu'ils m'ont généreusement accordé. Je tiens également à remercier Loïc et Clément, les stagiaires de mon bureau, pour leur aide précieuse et leur sympathie.

Je remercie Véronique Armand pour sa disponibilité et sa bonne humeur, ainsi que tous les membres du département pour leur accueil et l'ambiance chaleureuse qu'ils entretiennent.

Enfin, je souhaite exprimer ma reconnaissance à mes amis et collègues d'Aix et de Grenoble pour leur présence et leurs encouragements.

Sommaire

Introduction 8

1 - Présentation de l'entreprise 10

1.1 - EADS - European Aeronautic Defence and Space....10

1.2 - La division DS : Defence & Security....11

1.3 - Le département IPCC : Information Processing, Control & Cognition....12

1.3.1 - Présentation générale....12

1.3.2 - La plateforme WebLab....12

2 - Présentation du stage 14

2.1 - Contexte et besoins de l'entreprise....14

2.2 - L'extraction d'information : état de l'art....14

2.2.1 - Un sous-domaine du TAL....14

2.2.2 - Les campagnes d'évaluation....16

2.2.3 - Les outils existants....18

2.2.4 - GATE : General Architecture for Text Engineering....19

2.2.4.1 - Fonctionnement général....19

2.2.4.2 - Le formalisme JAPE....23

2.2.4.3 - Quelques plugins intéressants....24

3 - Création d'une ontologie de domaine 26

3.1 - Qu'est-ce qu'une ontologie ?....26

3.2 - Modélisation d'une ontologie....26

3.2.1 - Formats....27

3.2.2 - Outils existants....27

3.3 - Méthodologie adoptée....28

3.3.1 - Tour d'horizon des ontologies disponibles....29

3.3.2 - Construction d'une taxonomie simple....30

3.3.3 - Affinage selon les besoins du domaine....31

4 - Extraction d'entités nommées 34

4.1 - Constitution de corpus....34

4.2 - ANNIE : observation et amélioration des résultats....35

4.3 - Traitement du français....41

5 - Extraction d'évènements 44

5.1 - Définition de la méthode....44

5.1.1 - Observation des travaux existants....44

5.1.2 - Élaboration d'une approche....45

5.2 - Implémentation dans GATE....47

5.2.1 - Traitement de l'anglais....47

5.2.1.1 - Installation des plugins GATE nécessaires....47

5.2.1.2 - Constitution des gazetteers....48

5.2.1.3 - Développement des règles linguistiques....49

5.2.2 - Traitement du français....53

5.3 - Analyse qualitative et améliorations possibles....54

6 - Extraction de relations 56

6.1 - Méthode d'extraction....56

6.2 - Implémentation dans GATE....56

6.3 - Analyse qualitative et améliorations possibles....58

7 - Évaluation des résultats 60

7.1 - Protocole d'évaluation....60

7.2 - Analyse des résultats....61

7.3 - Observations et améliorations envisagées....63

Conclusion 68

Introduction

Dans le cadre du master professionnel « Modélisation et Traitements Automatiques en Industries de la Langue », les étudiants doivent effectuer un stage de fin d'études pour valider l'obtention de leur diplôme et préparer leur entrée dans le monde du travail. Ce stage de 4 mois minimum s'effectue au second semestre de la dernière année, dans une entreprise dont les activités sont liées au Traitement Automatique de la Langue (TAL).

Pour ma part, j'ai choisi d'effectuer ce stage au sein du département IPCC de l'entreprise EADS située à Val de Reuil (Haute-Normandie). Mon stage s'est déroulé sur 6 mois du 8 mars au 15 septembre 2010, période durant laquelle j'ai été encadrée par Bruno Grilheres, docteur-ingénieur et responsable de projets au sein de l'équipe.

L'objectif de mes travaux est de concevoir une ontologie de domaine et les outils d'extraction de l'information associés pour l'anglais et le français. Les activités d'EADS dans le domaine du renseignement militaire ont fait émerger un certain nombre de besoins (cf. section 2.1) en traitement de l'information et notamment en extraction d'information. Pour répondre à une partie de la demande des opérationnels du métier, mon stage s'est organisé en plusieurs phases que nous détaillerons dans ce mémoire :

- Définition d'une ontologie de domaine ;
- Extraction d'entités nommées ;
- Extraction d'évènements ;
- Extraction de relations ;
- Evaluation des résultats.

Nous commencerons par présenter succinctement l'entreprise EADS et l'équipe au sein de laquelle nous avons évolué tout au long de ce stage (cf. chapitre 1). Puis, nous décrirons les besoins de l'entreprise et le contexte théorique dans lequel s'insèrent nos travaux (cf. chapitre 2). Viendra ensuite la description détaillée des différentes étapes du stage citées plus haut (cf. chapitres 3, 4, 5, 6, 7) et nos conclusions sur cette expérience professionnelle.

1 - Présentation de l'entreprise

1.1 - EADS - European Aeronautic Defence and Space

EADS, groupe européen, mène ses activités dans le secteur de l'industrie aéronautique et spatiale, civile et militaire. Celui-ci développe et commercialise des avions civils et militaires mais également des systèmes de communications, missiles, lanceurs spatiaux, satellites, etc. Actuellement leader en Europe et seconde entreprise au niveau mondial (en concurrence avec Boeing) dans le domaine de l'aéronautique et de la défense, le groupe s'est formé le 10 juillet 2000 par la fusion de trois sociétés :

- l'allemande DaimlerChrysler Aerospace AG (DASA),
- Aérospatiale Matra, entreprise française,
- Construcciones Aeronáuticas SA (CASA) en Espagne.

Cette dimension multinationale est reflétée dans la direction du groupe par la présence d'acteurs de divers pays. À l'heure actuelle, EADS emploie environ 120 000 personnes et répartit son chiffre d'affaires sur tous les continents : 50 % en Europe, 14 % en Amérique du Nord, 20 % en Asie-Pacifique, 9 % au Moyen-Orient et 7 % dans le reste du monde. En 2009, son chiffre d'affaires est resté stable autour de 43 milliards d'euros. EADS est une société de droit néerlandais, cotée aux bourses de Francfort, Madrid et Paris et faisant partie du CAC 40.

Le groupe EADS comprend 4 divisions correspondant à ses différentes activités (cf. Figure 1.1) :

- Airbus (Airbus Commercial & Airbus Military), un des leaders de l'aviation mondiale, avec des activités dans les domaines commerciaux et militaires ;
- Eurocopter, premier constructeur d'hélicoptères au monde ;
- EADS Astrium, premier groupe spatial européen et troisième au niveau mondial ;
- EADS Defence & Security (DS), pôle des activités de défense et de sécurité du territoire.

Figure 1.1 : Divisions du groupe EADS
(source : http://www.eads.com/)

1.2 - La division DS : Defence & Security

La division Defence & Security (DS) est le pôle principal d'EADS pour ses activités de défense militaire et de sécurité du territoire. Celui-ci est centré sur l'expertise et le développement de systèmes intégrés comme les systèmes de missiles, les avions de combat, l'électronique de défense ainsi que les communications et services militaires. La division dispose d'une longue expérience et d'un grand savoir-faire en tant que fournisseur de systèmes complexes de défense qui intègrent des technologies développées par les autres divisions du groupe EADS et les sous-divisions de DS, présentées ci-après (cf. Figure 1.2).

EADS DS est constituée de 5 sous-divisions :

- Defence and Communications Systems, architecte et intégrateur de systèmes ;
- Defence Electronics, spécialiste des capteurs, avionique et guerre électronique ;
- Military Air Systems, fabricant d'aéronefs de combat ;
- MBDA, leader mondial en conception et production de missiles ;
- Eurofighter GMBH, constructeur d'avions de combat.

Au sein de la composante Defence and Communications Systems, le pôle System Design Center (SDC) est chargé de la conception des systèmes et de leur interopérabilité au sein d'EADS Defence and Security. Celui-ci est également divisé en plusieurs entités, dont le SCDE (Studies Concept Development & Experimentation), service en charge de la création de prototypes expérimentaux, au sein duquel se trouve le département IPCC, lieu de notre stage.

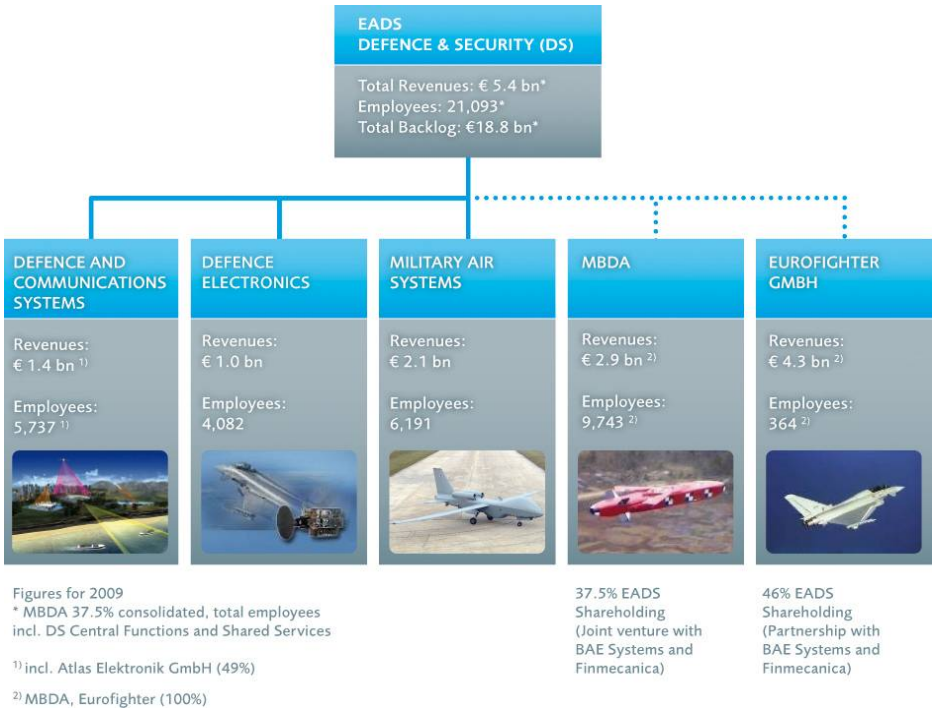


Figure 1.2 : Organisation de la division Defence & Security
(source : http://www.eads.com/)

1.3 - Le département IPCC : Information Processing, Control & Cognition

1.3.1 - Présentation générale

Situé à Val de Reuil (Haute-Normandie), le département IPCC est une entité R&D / R&T[1] placée sous la direction de Stephan Brunessaux. L'équipe comprend 21 personnes dont 6 doctorants en convention CIFRE, preuve de l'aspect innovant des thèmes abordés au sein du département. Notre maître de stage, Bruno Grilheres, est docteur-ingénieur dans cette équipe.

Les activités d'IPCC se concentrent autour des problématiques de traitement de l'information et s'articulent en trois composantes principales :

- le « media mining » (fouille de documents multimédias / fouille de données non-structurées),
- la fusion de données,
- les technologies Web et sécurité.

Ce service est chargé de l'innovation dans l'extraction, la gestion et l'exploitation des informations et des connaissances au sein des systèmes. Il participe à divers projets de recherche, français ou européens, dont la majorité est centrée sur le traitement de l'information, depuis la collecte de documents en passant par l'extraction des informations d'intérêt et jusqu'à leur exploitation par des utilisateurs. Ceux-ci sont souvent financés par l'Agence Nationale de la Recherche (ANR) et mettent en collaboration plusieurs acteurs industriels et universitaires. Parmi les entreprises avec lesquelles IPCC a pu travailler, nous pouvons citer, parmi les plus connues, Sinequa, Mondeca, Temis, Synapse, Thalès, Xerox, Exalead, Systran, etc. Le département est également en lien avec plusieurs entités universitaires telles que le GREYC[2] (laboratoire de Caen), le LITIS (laboratoire de Rouen/Le Havre), le CEA-List, le LIP6 (laboratoire de Paris VI), etc.

1.3.2 - La plateforme WebLab[3]

Les compétences mentionnées plus haut ont donné le jour au WebLab, plateforme dédiée à l'intégration de divers services de « media mining » en vue de l'exploitation de documents multimédias. Cette architecture est au centre de nombreux projets menés au sein d'IPCC. L'un des objectifs du WebLab est de proposer des solutions dédiées au traitement de l'information dans un but stratégique et ce, dans des domaines tels que la veille économique et technologique et le renseignement militaire (OSINT[4]). Celles-ci constituent le niveau « WebLab Applications » du schéma présenté ci-dessous (cf. Figure 1.3). Comme nous pouvons le voir, le WebLab est composé de deux autres couches : WebLab Services et WebLab Core. En effet, pour construire ces

1 Recherche et développement / Recherche et technologie
2 Groupe de Recherche en Informatique, Image, Automatique et Instrumentation de Caen
3 © EADS 2008 : All rights reserved.
4 Open Source Intelligence

applications, cette plateforme permet d'intégrer divers composants (COTS[5]) commerciaux, open source, développés par EADS ou ses partenaires. Ces COTS ont des fonctions diverses et permettent de construire une chaine de traitement très complète proposant entre autres :

- l'acquisition de données (Web, bases de données),
- la normalisation de texte, images, etc.
- la transcription parole-texte,
- l'extraction d'entités nommées et de relations,
- l'analyse sémantique,
- la catégorisation thématique d'un document,
- le résumé automatique de texte,
- l'indexation de données,
- la recherche plein texte et sémantique.

Pour développer de tels composants, le département IPCC adopte une approche pluridisciplinaire mêlant analyse linguistique et sémantique, techniques statistiques et méthodes d'apprentissage. Enfin, le WebLab Core, cœur open source de la plateforme, permet, par l'utilisation de standards reconnus, l'intégration et l'interopérabilité entre ces différents services.

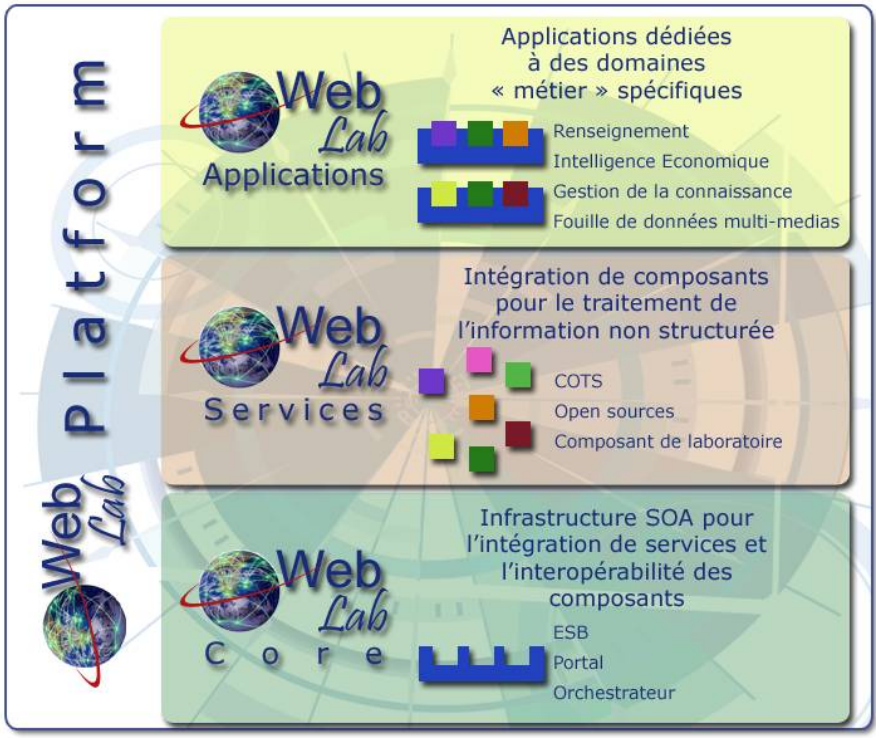


Figure 1.3 : La plateforme WebLab
(© EADS 2008 : All Rights Reserved)

5 Commercial Off-The-Shelf (« composant pris sur étagère »)

2 - Présentation du stage

2.1 - Contexte et besoins de l'entreprise

Comme nous avons pu le constater lors de la présentation du département IPCC, ses projets sont centrés sur le traitement de l'information dans des documents multimédias. Leurs activités dans le renseignement militaire impliquent une étude des informations en sources ouvertes (ISO), à savoir tout document accessible publiquement et légalement (presse écrite, blogs, sites internet, radio, télévision, etc.). En effet, dans un but stratégique, il est crucial de « fouiller » cette masse d'informations, dont la majorité s'avère non-structurée, afin d'en extraire des connaissances pertinentes et utiles dans un but donné. Cela implique divers traitements textuels et donc une analyse approfondie des phénomènes linguistiques. Les collaborations entre IPCC et certaines entreprises de TAL permettent déjà à ceux-ci de proposer au sein du WebLab des composants d'analyse linguistique et en particulier d'extraction d'information. Toutefois, l'équipe souhaite proposer ses propres COTS d'extraction d'information pour pouvoir les adapter à sa convenance (algorithmes spécifiques à un domaine donné) mais aussi les réutiliser de manière totalement libre au sein du département et notamment dans le cadre des thèses en cours. En effet, beaucoup des travaux actuels de l'équipe peuvent tirer bénéfice d'un tel système, comme les recherches sur la cotation de l'information[6] en sources ouvertes, la construction automatique de fiches biographiques à partir de données Web ou encore la modélisation de l'information géographique en sources ouvertes. La construction d'un tel outil nécessitant des compétences en linguistique mais également des bases avancées en informatique, le choix d'un stagiaire avec un profil de linguiste informaticien a été privilégié.

2.2 - L'extraction d'information : état de l'art

2.2.1 - Un sous-domaine du TAL

Depuis les débuts du Traitement Automatique du Langage dans les années 60-70, la compréhension automatique de textes est l'objet de nombreuses recherches et vise à saisir le sens global d'un document. Les échecs récurrents des systèmes alors développés mettent rapidement en cause une vision trop générique de la compréhension automatique. En effet, de tels outils s'avèrent inutilisables dans un contexte opérationnel en raison du coût élevé des adaptations nécessaires (bases de connaissances et ressources lexicales spécifiques). Conscients d'être trop ambitieux au regard des possibilités technologiques, les chercheurs s'orientent alors vers des techniques plus réalistes d'extraction d'information. S'il n'est pas directement possible de comprendre automatiquement un texte, le repérage et l'extraction des principaux éléments de sens apparaît comme un objectif plus raisonnable. Cette réorientation théorique est reprise de façon détaillée par [Poibeau, 2003].

6 Évaluation de la crédibilité/plausibilité de l'information

L'extraction d'information est donc une discipline assez récente qui consiste en une analyse partielle d'un texte afin d'en extraire des informations spécifiques. Celles-ci permettent de construire une représentation structurée (bases de données, fiches, tableaux) d'un document à l'origine non-structuré. Il s'agit donc d'une tâche plus limitée où l'on détermine à l'avance le type d'entité à extraire automatiquement. Cela en fait une approche guidée par le but de l'application dans laquelle elle s'intègre, dépendance qui reste, à l'heure actuelle, une limite majeure des systèmes d'extraction. Les tâches les plus communes en extraction d'information restent l'extraction d'entités nommées [Nadeau et al., 2007], de relations entre entités et d'évènements. Comme nous l'avons déjà précisé, l'extraction d'information est un domaine dépendant du but, le nombre et le type des entités-cibles est donc variable selon l'application.

L'extraction d'entités nommées a été et est encore beaucoup étudiée car celles-ci constituent des éléments indispensables pour saisir le sens d'un texte. Leur définition est encore aujourd'hui le sujet de nombreuses discussions, nous retiendrons ici celle admise par la majorité : entités fortement référentielles, elles désignent directement un objet du monde (*« désignateurs rigides »* de Kripke) et correspondent de façon générale aux noms propres de personne, organisation, lieu, mais aussi aux dates, unités monétaires, pourcentages, unités de mesure, etc. Par ailleurs, les différents outils d'extraction d'information s'attachent à extraire des éléments textuels plus complexes tels que les relations et les évènements. Les relations correspondent aux liens existants entre différentes entités nommées repérées dans un texte : il peut s'agir par exemple de détecter les relations entre une personne et une organisation (appartenance, direction, etc.) ou encore d'extraire les attributs d'une personne (date de naissance, e-mail, adresse, etc.). Enfin, une dernière tâche est l'extraction d'évènements, particulièrement utile dans les activités de veille économique et stratégique. Celle-ci peut-être conçue comme une forme particulière d'extraction de relations : une action est reliée avec une date, un lieu et des participants. Cette définition peut varier légèrement mais au moins un de ces éléments est nécessaire pour la présence d'un événement. Le modèle TimeML, par exemple, considère qu'un événement se définit au minimum par la présence d'un indicateur temporel.

Les dix dernières années ont vu apparaître un intérêt grandissant pour cette discipline avec notamment l'apparition de campagnes d'évaluation que nous aborderons plus en détail dans la partie suivante. Deux approches principales émergent alors : l'extraction basée sur des techniques linguistiques d'un côté et les systèmes statistiques à base d'apprentissage de l'autre. Celles-ci se basent, de façon commune, sur des pré-traitements linguistiques « classiques » comme la « tokenization » (découpage en mots), la lemmatisation (attribution de la forme non-fléchie associée), l'analyse morphologique (structure et propriétés d'un mot) ou syntaxique (structure d'une phrase et relations entre éléments d'une phrase).

La première approche exploite les avancées en TAL et repose principalement sur l'utilisation de grammaires formelles construites par la main d'un expert-linguiste. Les pré-traitements cités plus haut servent de base à la construction de règles et patrons linguistiques qui définissent les contextes d'apparition de telle entité ou relation. Notons ici l'importance particulière accordée à l'analyse syntaxique (en constituants ou dépendance) dans le repérage et le typage des relations et des évènements.

La seconde approche utilise des techniques statistiques pour « apprendre » des régularités sur de larges corpus de textes où les entités-cibles ont été préalablement annotées. Ces méthodes

d'apprentissage ou « machine learning » sont plus ou moins supervisées[7] et exploitent des caractéristiques textuelles plus ou moins linguistiques issues des pré-traitements précédemment évoqués. Parmi celles-ci nous pouvons citer les « modèles de Markov Caché » (HMM), les « Conditionals Random Fields » (CRF), les « Support Vector Machine » (SVM) ou encore les techniques de « bootstrapping » et de « clustering » [Ireson et al., 2005]. Par ailleurs, actuellement, de plus de plus de recherches portent sur l'apprentissage symbolique de ressources linguistiques dans divers domaines.

Ces dernières années, un nouveau type d'approche tend à se généraliser : ce sont les méthodes hybrides. Les limites de chacune des approches que nous venons de mentionner ont amené les acteurs du domaine à mêler les techniques existantes pour augmenter les performances de leurs outils. En effet, un certain nombre de problèmes en extraction d'information constitue un réel frein à la commercialisation des systèmes existants. Tout d'abord, la plupart des solutions sont développées pour un domaine ou un genre de texte particulier et voient leurs performances décroître rapidement face à un texte différent de ce point de vue. Le même problème survient lorsque les outils sont développés à partir de corpus très homogènes (sur la forme ou le contenu) et que ceux-ci sont réutilisés sur d'autres corpus de natures plus variées. Ces limites concernent à la fois les méthodes symboliques et statistiques et nécessitent une ré-adaptation constante des techniques. Les approches à base de règles linguistiques, elles, souffrent également du coût de leur développement manuel et de la nécessité d'une expertise en linguistique pour pouvoir les modifier et les adapter. Pour tenter de résoudre cela, les experts se penchent actuellement vers des méthodes d'apprentissage automatique de patrons linguistiques. Pour finir, les approches statistiques nécessitent, lors de la phase d'apprentissage, une grande quantité de textes pré-annotés et, ces données n'étant pas toujours disponibles, cela constitue une réelle contrainte. Des recherches sont menées dans ce sens avec notamment l'utilisation d'un apprentissage dit « semi-supervisé », qui vise à améliorer les performances en combinant les données étiquetées et non-étiquetées.

2.2.2 - Les campagnes d'évaluation

Afin de stimuler le développement de ces techniques et de dégager les pistes de recherche les plus prometteuses, des campagnes d'évaluation sont menées au niveau européen et international. Celles-ci ont pour but de mettre en place un protocole d'évaluation permettant aux experts de mesurer les performances des outils qu'ils ont développés. Les campagnes définissent généralement plusieurs tâches à accomplir telles que l'extraction d'entités nommées, de relations ou encore d'évènements, la résolution de coréférence, etc. Le protocole le plus courant est de fournir un corpus d'entrainement et un corpus de test où les éléments à extraire ont été pré-annotés ainsi qu'un ou plusieurs scripts de « scoring » (i.e. d'évaluation). Le corpus d'entraînement permet de préparer l'outil à la tâche d'extraction pour pouvoir ensuite s'auto-évaluer sur le corpus de test et estimer son score grâce aux scripts fournis. Les métriques d'évaluation les plus communes sont la précision, le rappel et la F-mesure. En extraction d'information, ces métriques peuvent être définies ainsi[8] :

7 Leurs données d'apprentissage peuvent être des exemples préalablement annotés ou des données brutes.

8 La variable E désigne le type d'étiquette

$$précision = \frac{nombre\ d'entités\ correctement\ étiquetées\ E}{nombre\ d'entités\ étiquetées\ E}$$

$$rappel = \frac{nombre\ d'entités\ correctement\ étiquetées\ E}{nombre\ d'entités\ E}$$

$$F_{mesure} = \frac{2 \cdot précision \cdot rappel}{précision + rappel}$$

Figure 2.1 : Métriques d'évaluation

Une fois leurs systèmes préparés à la tâche d'évaluation, ceux-ci sont évalués et classés par les organisateurs de la campagne. Ces évaluations s'accompagnent le plus souvent de publications d'articles dans lesquels les participants décrivent leur(s) outil(s) et les techniques mises en œuvre. Cela permet de mettre en avant les nouvelles approches et de faire le point sur les performances de celles déjà connues.

Dans le domaine de l'extraction d'information, les campagnes MUC (Message Understanding Conference) restent les pionnières et les plus connues au niveau international. Créées au début des années 1990 par la DARPA[9], elles constituent les premières initiatives pour encourager l'évaluation des systèmes d'extraction et ont fortement contribué à l'essor de ce domaine. À l'origine destinées au domaine militaire, les sept séries d'évaluation menées ont permis de diversifier les applications. Celles-ci se caractérisent par la tâche d'extraction consistant à remplir un formulaire à partir d'un texte (participants d'un évènement, par exemple). Certains jeux de données de ces campagnes sont actuellement mis à disposition gratuitement. Notons pour finir, que le terme « entité nommée » est né lors d'une conférence MUC.

Par ailleurs, nous pouvons citer le programme ACE (Automatic Content Extraction) qui, sous la direction du NIST[10], mène également des campagnes d'évaluation. Spécialisées dans l'analyse d'articles de presse, celles-ci évaluent les systèmes sur l'extraction d'entités nommées et sur la résolution de coréférences (mentions d'entités nommées). Aujourd'hui, la campagne TAC (Text Analysis Conference) a pris la suite des actions menées dans le cadre du programme ACE.

Toujours à l'échelle mondiale, les campagnes CoNLL (Conference on Natural Language Learning) évaluent et font la promotion des méthodes d'extraction par apprentissage. Celles-ci sont classées parmi les meilleures conférences internationales dans le domaine de l'intelligence artificielle. Ce succès est en partie du au fait que ces conférences sont dirigées par l'ACL (Association of Computational Linguistics), la plus réputée des associations de linguistique et informatique. Celle-ci est aussi à l'origine des conférences Senseval/Semeval spécialisées dans l'évaluation des outils de désambiguïsation sémantique, point crucial en extraction d'information.

En Europe, l'association ELRA (European Language ressources Association) a mis en place les conférences LREC (Language Ressources and Evaluation Conference). Lors de celles-ci les différents acteurs en ingénierie linguistique présentent de nouvelles méthodes d'évaluation ainsi que divers outils liés aux ressources linguistiques. De plus, cette association participe à l'évaluation de systèmes divers en fournissant les corpus et données nécessaires. Enfin, il nous faut citer la campagne française ESTER (Évaluation des Systèmes de Transcription Enrichie d'Émissions

9 Defense Advanced Research Projects Agency
10 National Institute of Standards and Technology

Radiophoniques) qui, entre autres activités, évalue le repérage d'entités nommées par des systèmes de transcription enrichie.

La DARPA a également initié le « Machine Reading Program » (MRT) : projet visant à construire un système universel de lecture de texte capable d'extraire automatiquement la connaissance du langage naturel pour la transformer en représentation formelle. Celui-ci est destiné à faire le lien entre le savoir humain et les systèmes de raisonnement nécessitant ce savoir. Il s'agit pour cela de combiner les avancées en Traitement Automatique du Langage Naturel (TALN) et en Intelligence Artificielle (IA).

Précisons, pour finir, que ces quelques projets et campagnes d'évaluation ne sont évidemment pas les seules existantes et que nous avons fait le choix de les mentionner car ce sont celles que nous avons pu rencontrer durant nos recherches. Cette remarque s'applique également à la description de quelques outils d'extraction d'information dans la partie suivante.

2.2.3 - Les outils existants

Pour poursuivre ce bref état de l'art, citons quelques outils d'extraction d'information dont nous avons pu avoir connaissance durant notre stage. Au vu du nombre important de solutions commerciales proposées par des entreprises de TAL, nous faisons le choix de n'aborder que des outils open source et/ou gratuits.

Tout d'abord, la société Thomson Reuters (qui a racheté ClearForest) propose plusieurs services autour de l'extraction d'information regroupés sous le nom OpenCalais [Thomson Reuters, 2008]. Celle-ci a mis en place OpenCalais Web Service, un outil en ligne d'extraction d'entités nommées, faits et évènements. Celui-ci ainsi que les divers plugins qui l'accompagnent (Marmoset, Tagaroo, Gnosis, etc.) sont utilisables gratuitement pour usage commercial ou non-commercial. Le service d'annotation en ligne permet de traiter des textes en anglais, français et espagnol grâce à une détection automatique de la langue du texte fourni. Il extrait pour toutes ces langues un nombre conséquent d'entités nommées (villes, organisations, monnaies, personnes, e-mails, etc.) et attribue également un indice de pertinence/intérêt à chacune d'elles. L'analyse des textes anglais est plus complète : extraction d'évènements et de relations, désambiguïsation d'entités, détection de thème, association automatique de mots-clés (« semantic tags »), etc. Toutes ces annotations peuvent être récupérées au format RDF[11] dont nous reparlerons (cf. section 3.2.1). Enfin, précisons qu'OpenCalais est le fruit de l'utilisation de techniques linguistiques couplées avec des méthodes statistiques et d'apprentissage.

Par ailleurs, LingPipe développé par Alias-i constitue une véritable « boîte à outils » pour l'analyse automatique de textes [Alias-i, 2003]. Divers outils y sont disponibles gratuitement pour la recherche et, parmi ceux-ci, une majorité relèvent plus ou moins directement du domaine de l'extraction d'information. D'une part, des modules de pré-traitement permettent de « préparer » le texte pour la phase d'extraction : l'analyse morpho-syntaxique, le découpage en phrases, la désambiguïsation sémantique de mots. D'autre part, LingPipe met à disposition des modules de détection d'entités nommées et de phrases d'intérêt, d'analyse d'opinion et de classification thématique. Ces traitements sont tous réalisés par approche statistique et notamment par l'utilisation de CRF et de modèles d'apprentissage (spécifiques à une langue, un genre de texte ou un type de

11 Ressource Description Framework

corpus).

Également renommé, le groupe OpenNLP[12] rassemble un nombre important de projets open source autour du Traitement Automatique du Langage [OpenNLP, 2008]. Son objectif principal est de promouvoir ces initiatives et de favoriser la communication entre acteurs du domaine pour une meilleure interopérabilité entre systèmes. En extraction d'information, nous pouvons retenir les projets NLTK[13] [Bird, 2008] et MALLET[14] [McCallum, 2009]. Le premier correspond à plusieurs modules en Python pouvant servir de base au développement de son propre outil d'extraction. Le second projet, MALLET, est un logiciel développé par les étudiants de l'université du Massachussetts Amherst sous la direction d'Andrew McCallum, expert du domaine [McCallum, 2005]. Ce logiciel inclut différents outils pour l'annotation de segments (entités nommées et autres), tous basés sur des techniques statistiques de type CRF, HMM et MEMM (« Maximum Entropy Markov Models »).

Pour finir, mentionnons également les groupes de recherche Stanford NLP[15] Group [Stanford, 2004] et Ontotext [Ontotext, 2000] dont les travaux sont intégrés dans GATE (cf. section suivante). L'équipe de l'université de Stanford en Californie, a créé différents outils de TAL très utiles pour l'extraction d'information : un analyseur syntaxique probabiliste pour l'anglais, un étiqueteur morpho-syntaxique ainsi qu'un système d'extraction d'entités nommées qui reconnaît les noms de personne, d'organisation et de lieu. Ontotext développe ses activités autour des technologies sémantiques et diffuse gratuitement la plateforme KIM[16] pour un usage non-commercial. Celle-ci propose de créer des liens sémantiques entre documents mais aussi d'extraire les entités nommées, relations et évènements d'un texte et de les stocker automatiquement dans une base de données.

2.2.4 - GATE : General Architecture for Text Engineering

GATE [Cunningham et al., 2002] ayant déjà été utilisé lors de projets antérieurs au sein d'IPCC et nos diverses recherches n'ayant pas révélé de système plus complet, nous avons choisi cet environnement pour réaliser les objectifs de notre stage. Plateforme open source en Java dédiée à l'ingénierie textuelle, celle-ci nous est apparue bien adaptée au développement de notre système. Créée il y a 15 ans par les chercheurs de l'université de Sheffield (Royaume-Uni), GATE est largement utilisé par les experts en TAL et dispose d'une grande communauté d'utilisateurs. Cela lui permet de disposer d'un ensemble de solutions d'aide et de support (forum, liste de diffusion, foire aux questions, wiki, tutoriels, etc.), point particulièrement important lorsque l'on débute avec un tel outil. Par ailleurs, les créateurs de GATE propose des formations pour améliorer son niveau ainsi que des certifications permettant de faire valoir ses compétences à un niveau professionnel.

2.2.4.1 - Fonctionnement général

L'environnement GATE peut être utilisé en tant que librairie ou grâce à une interface graphique, solution que nous avons privilégiée ici. L'outil repose sur le principe d'une chaine de

12 Open Natural Language Processing
13 Natural Language ToolKit
14 Machine Learning for LanguagE Toolkit
15 Natural Language Processing
16 Knowledge and Information Management

traitement (« pipeline ») composée de plusieurs modules (dits « Processing Resources » PR) appliqués successivement sur un ou plusieurs textes (dits « Language Resources » LR). Les documents donnés en entrée peuvent être un simple texte ou un corpus (cf. Figure 2.3), fournis soit par un « copier-coller » soit par une URL. Les différents composants annotent chacun à leur tour le texte en prenant en compte les annotations précédentes puis le document est retourné à l'utilisateur au format XML[17] (cf. Erreur : source de la référence non trouvée). Les annotations correspondent généralement à l'association d'attributs (« features ») à une zone de texte. Ceux-ci se présentent sous la forme *propriété = « valeur »*, où la valeur peut être une chaîne de caractères ou un nombre. Voici par exemple l'annotation obtenue grâce à un composant de segmentation en mots :

*Token {category=*NNP*, kind=word, length=5, orth=upperInitial, string=Obama}*

Figure 2.2 : Exemple d'annotation dans GATE

Il s'agit ici d'une annotation de type « Token » (c'est-à-dire une chaine de caractères entre deux espaces) à laquelle sont associés la catégorie grammaticale du « token », son type, sa longueur en nombre de caractères, sa casse et la chaîne de caractères associée.

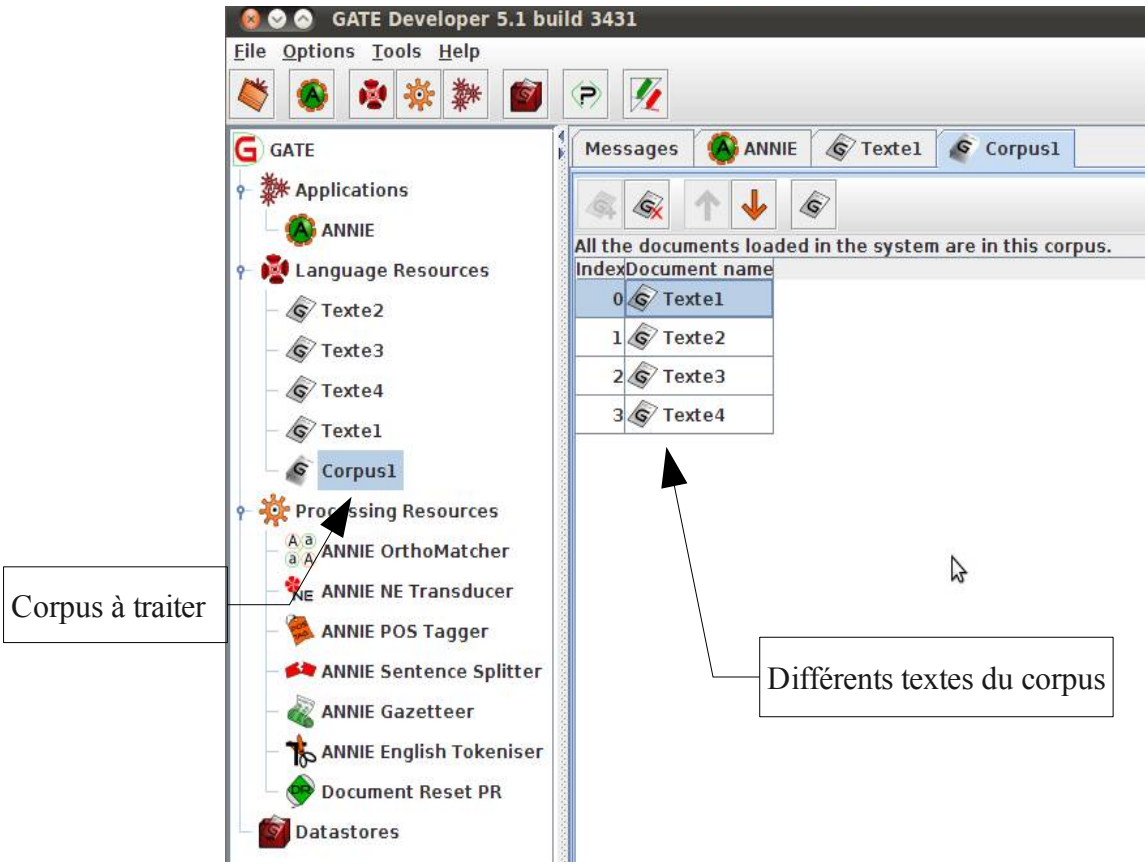


Figure 2.3 : Exemple de corpus dans GATE

17 eXtensible Markup Language

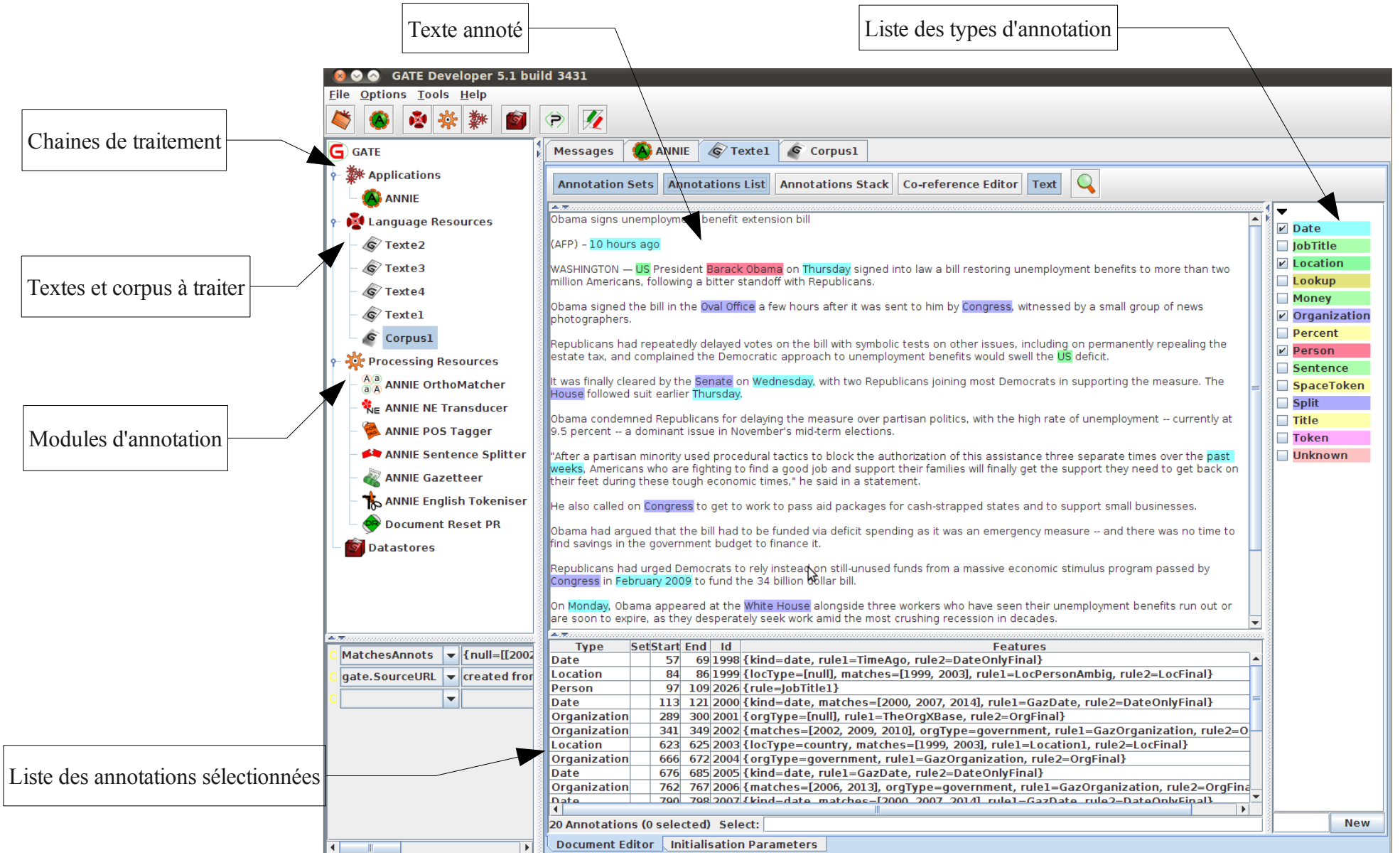


Figure 2.4 : Exemple de texte annoté dans GATE

Par un système de plugins, GATE met à disposition de ses utilisateurs un grand nombre de modules dédiés à l'analyse textuelle. Les plus courants sont les segmenteurs (*Tokenizers*), les étiqueteurs morpho-syntaxiques (*Part Of Speech Taggers*), les lexiques (*Gazetteers*) ou encore les transducteurs (*JAPE*[18] *transducers*). L'interface graphique permet de charger de nouveaux plugins et ressources, de les paramétrer et de les combiner au sein d'une même chaine de traitement (cf. Figure 2.5).

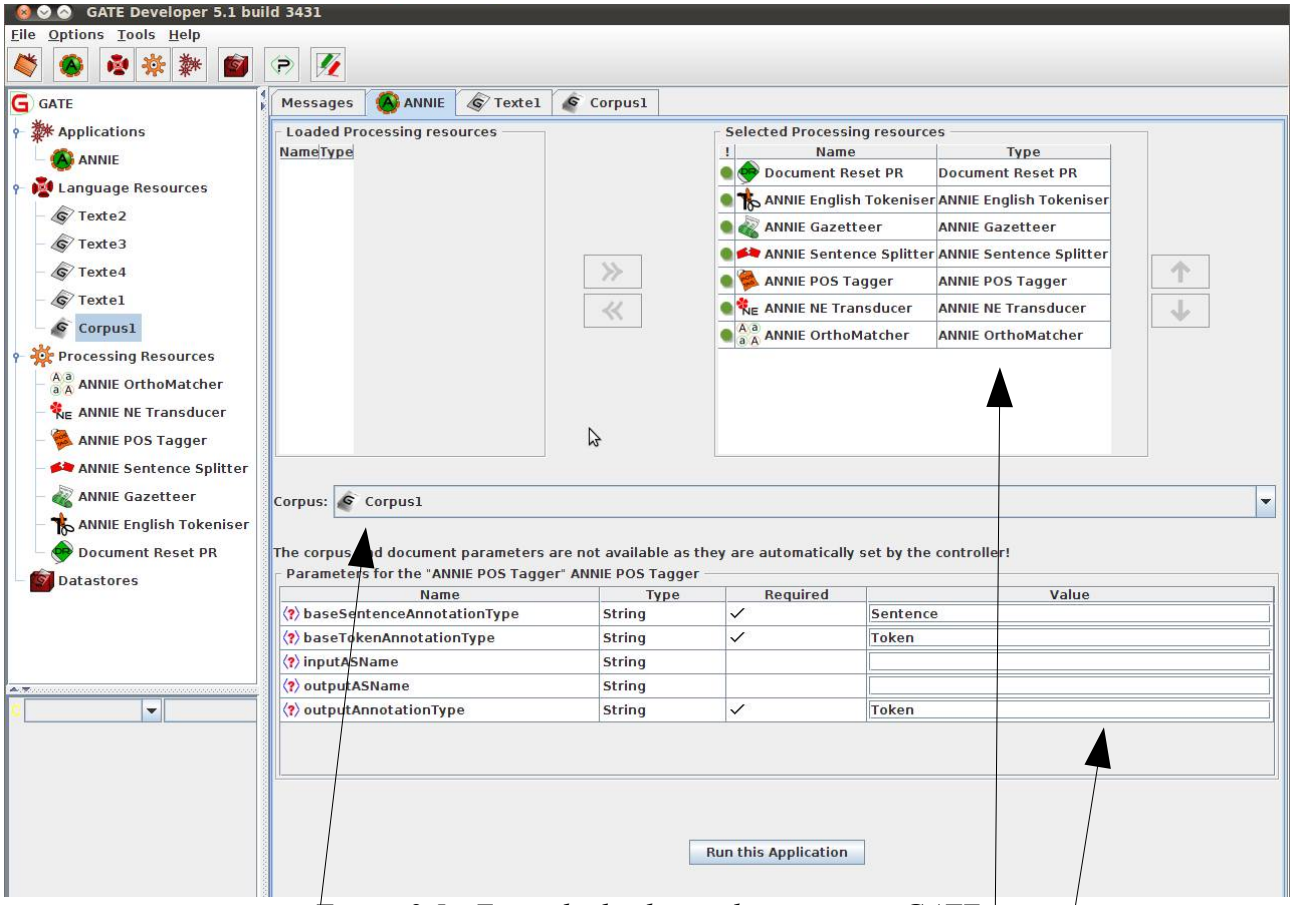


Figure 2.5 : Exemple de chaîne de traitement GATE

Corpus à traiter

Composition de la chaine de traitement

Paramètres du module sélectionné

18 Java Annotation Patterns Engine

2.2.4.2 - Le formalisme JAPE

Une partie des différents modules proposés dans GATE est basée sur JAPE (Java Annotation Patterns Engine), un transducteur à états finis permettant de reconnaître des expressions régulières sur les annotations. Ce système s'avère très utile en extraction d'information car il permet de définir les contextes d'apparition des éléments à extraire pour ensuite les repérer et les annoter. Le principe est de combiner différentes annotations « basiques » (*tokens*, syntagmes, relations syntaxiques, etc.) pour en créer de nouvelles plus complexes (entités nommées, relations, évènements, etc.) : cela revient à l'écriture de règles de production et donc à l'élaboration d'une grammaire régulière.

Une grammaire JAPE se décompose en plusieurs phases exécutées consécutivement et formant une cascade d'automates à états finis. Chaque phase correspond à un fichier « .jape » et peut être constituée d'une ou plusieurs règle(s) écrite(s) selon le formalisme associé à JAPE (dérivé de CPSL[19]). Classiquement, ces règles sont divisées en deux blocs : une partie gauche (« Left Hand Side » ou LHS) définissant un motif d'annotations à repérer et une partie droite (« Right Hand Side » ou RHS) contenant les opérations à effectuer sur ce motif. Le lien entre ces deux parties se fait par l'attribution d'une étiquette au motif (ou à ses constituants) en LHS et par sa réutilisation en RHS pour y appliquer les opérations nécessaires. Pour plus de clarté, prenons l'exemple d'une règle simple :

```
1.  Rule: OrgAcronym
2.  ((
3.   {Organization}
4.   {Token.string == "("}
5.   ({Token.orth == "allCaps"}):org
6.   {Token.string == ")"}
7.  )
8.  -->
9.  :org.Organization = {rule = "OrgAcronym"}
```

L'objectif de celle-ci est d'annoter en tant qu'organisation les acronymes entre parenthèses positionnés après une annotation de type « Organization ». Tout d'abord, l'on commence par donner un nom à la règle (ligne 1). Les lignes 2 à 7 définissent le motif à repérer dans le texte. Le signe « `- - >` » (ligne 8) sert de séparateur entre LHS et RHS. Enfin, la dernière ligne (ligne 9) exprime l'opération souhaitée. Précisons quelques règles syntaxiques de base du formalisme JAPE :

- La partie gauche de la règle est toujours entre parenthèses,
- La partie droite débute par le signe « --> »,
- Les types d'annotation sont encadrés par des accolades (ex : {Organization}),
- « Token.string » permet d'obtenir la valeur de la propriété « string » associée à l'annotation « Token »,
- « :org » permet d'étiqueter une partie du motif en LHS pour l'utiliser en RHS,

19 Common Pattern Specification Language

- La ligne 9 attribue une annotation de type « Organization » au segment étiqueté « org » en LHS ; l'ajout de la propriété « rule » à cette annotation permet d'indiquer quelle règle en est à l'origine.

La liste des annotations utilisées en LHS de la règle est déclarée en début de phase grâce à l'attribut « Input » : par exemple, « Input : Lookup, Token, Person ». Par ailleurs, l'attribut « Control » permet de définir l'ordre d'exécution des différentes règles d'une phase et « Debug » d'obtenir un affichage des éventuels conflits rencontrés entre règles.

Pour finir, précisons qu'un système de macros est également disponible : une macro permet de définir et de nommer une séquence d'annotations afin de la réutiliser plus rapidement par la suite.

2.2.4.3 - Quelques plugins intéressants

Dans cette partie, nous présentons différents plugins disponibles dans GATE qui nous ont été utiles pour réaliser les différentes tâches d'extraction.

Tout d'abord, nous devons parler d'ANNIE (A Nearly-New Information Extraction system), chaine complète dédiée à l'extraction d'information et plus particulièrement à l'extraction d'entités nommées pour l'anglais. Ce plugin fourni par les créateurs de GATE est basé sur le langage JAPE et comprend différents modules d'annotation :

- *Document Reset* : supprime toutes les annotations précédentes mises sur le document.
- *English Tokenizer* : découpe en mots (« tokens ») un texte anglais.
- *Gazetteer* : cherche les éléments d'une liste (appelée aussi « gazetteer ») dans le texte et les annote en tant que « Lookup ».
- *Sentence Splitter* : découpe le texte en phrases.
- *POS Tagger* : ajoute à l'annotation *Token* mise par le « tokenizer » une propriété « category » indiquant la catégorie morpho-syntaxique du mot en question. Il s'agit, ici, d'une version modifiée de l'étiqueteur Brill [Brill, 1992].
- *NE Transducer* : transducteur JAPE permettant de repérer un certain nombre d'entités nommées (Person, Organization, Location, Date, URL, Phone, Mail, Adress, etc.).
- *OrthoMatcher* : ajoute des relations d'identité référentielle (cf. glossaire) entre les entités nommées annotées précédemment.

ANNIE étant spécialisée dans le traitement de textes anglais, d'autres modules se sont avérés nécessaires pour l'extraction d'information en français. Pour cela, nous avons utilisé le plugin *Lang_French* et, en particulier, les modules de découpage en mots et d'étiquetage morpho-syntaxique adaptés pour la langue française (« French Tokenizer » et TreeTagger).

La phase d'extraction d'événements nous a également permis d'explorer d'autres outils fournis par GATE. L'analyseur syntaxique Stanford constitue l'un des plugins essentiels de cette étape : celui-ci fournit une analyse syntaxique en constituants et en dépendance des phrases du texte à traiter. La section 5.2.1.1 donnera une description plus détaillée de celui-ci. Par ailleurs, nos travaux nécessitant une analyse de la phrase en groupes nominaux et verbaux, les modules « NP

Chunker » et « VP Chunker » ont répondu à ce besoin. Pour finir, mentionnons le plugin nommé « Tools » qui fournit divers modules d'analyse linguistique tels qu'un analyseur morphologique (cf. glossaire), un « Flexible Gazetteer » ou encore un visualiseur d'arbre syntaxique.

3 - Création d'une ontologie de domaine

Le but premier de ce stage étant de répondre à un besoin de partage des informations dans le domaine du renseignement militaire, l'utilisation d'une ontologie s'est avérée essentielle. Nous donnons ici quelques éléments de définition et décrivons les différentes étapes de sa construction.

3.1 - Qu'est-ce qu'une ontologie ?

Le terme « ontologie » trouve son origine dans une branche de la philosophie nommée « métaphysique » qui étudie les principes de la réalité et a pour objet de définir la nature de l'être et du monde. Les ontologies ont été par la suite reprises dans le domaine de l'intelligence artificielle pour les systèmes à base de connaissances puis adaptées aux problématiques d'extraction de l'information. Dans ce contexte, plusieurs définitions ont été proposées dont celles de T. Gruber [Gruber, 1993] et R. Neches [Neches et al., 1991]. Gruber donne une définition très abstraite de l'ontologie, largement utilisée dans la littérature : «Une ontologie est une spécification explicite d'une conceptualisation»[20]. Le second la définit ainsi : « Une ontologie définit les termes et les relations de base du vocabulaire d'un domaine ainsi que les règles qui permettent de combiner les termes et les relations afin de pouvoir étendre le vocabulaire »[21]. La première définition fait référence à une caractéristique principale des ontologies à savoir l'élaboration d'un modèle abstrait de l'existant (conceptualisation) et sa formalisation en vue d'une exploitation par des machines. Le type d'ontologie varie selon la portée de cette modélisation. Dans le cadre de nos travaux, nous nous intéresserons exclusivement aux ontologies dites « de domaine » qui représentent le savoir d'un domaine particulier (le renseignement militaire dans notre cas).

Une ontologie de ce type recense les termes utilisés pour décrire et représenter un domaine de connaissances. Ceux-ci correspondent à des concepts plus ou moins génériques, également appelés « classes ». L'organisation entre ses classes est déterminée par des relations hiérarchiques ou taxonomiques (subsomption) : les classes supérieures correspondent à des concepts généraux et les classes inférieures (« sous-classes ») représentent des concepts plus spécifiques héritant des propriétés de leur « classe-mère ». Ces relations sont distinguées des propriétés qui peuvent être des relations entre deux concepts (« object properties ») ou des attributs associés à un concept (« data properties »).

Pour résumer, nous pouvons dire qu'une ontologie de domaine définit sémantiquement un ensemble de concepts utilisés par les experts de ce domaine (langue de spécialité) et leurs relations de manière non-ambigüe.

3.2 - Modélisation d'une ontologie

Pour permettre à un ordinateur d'exploiter cette représentation, des langages informatiques spécifiques aux ontologies ont été développés. Nous décrirons les principaux par la suite, ainsi que

20 Traduction de « An ontology is an explicit specification of a conceptualization »

21 Traduction de « An ontology defines the basic terms and relations comprising the vocabulary of a topic area as well as the rules for combining terms and relations to define extensions to the vocabulary »

quelques outils permettant de modéliser une ontologie.

3.2.1 - Formats

Un certain nombre de langages de spécification d'ontologie ont été créés dans le cadre de travaux sur le Web Sémantique. Parmi les plus connus nous pouvons citer N3 (Notation3), Common Logic, DOGMA[22], KIF[23], F-Logic, OIL[24] ou encore OWL[25] [W3C, 2004a].

Ce dernier reste le standard le plus renommé et le plus utilisé dans le domaine des ontologies. Comme la plupart des autres langages, OWL est basé sur les principes de la logique de premier ordre et en particulier sur les assertions du type « sujet-prédicat-objet ». Ce langage a été développé en 2002 par un groupe de travail du W3C[26] afin de modéliser des ontologies utilisables et échangeables sur le Web. Celui-ci s'inspire de langages comme DAML+OIL[27] et constitue une extension du standard RDF [W3C, 2004b]. RDF est un formalisme pour la représentation de la connaissance fondé sur la notion de triplet. Sur le modèle des assertions « sujet-prédicat-objet » mentionnées plus haut, un triplet est une relation « ressource-propriété-valeur ». RDF, tout comme son extension OWL, sont des langages basés sur la syntaxe XML. On trouvera en annexe un extrait de notre ontologie de domaine afin d'illustrer la syntaxe du format OWL[28].

3.2.2 - Outils existants

Bien que l'intérêt pour les ontologies en informatique soit relativement récent, de nombreux outils ont été développés dans le but de modéliser et de manipuler des ontologies. Le principal atout de ces logiciels est la possibilité de gérer une ontologie dans l'un des formats cités précédemment sans avoir à modifier manuellement le code sous-jacent. Nous mentionnerons ici uniquement les logiciels distribués librement les plus utilisés.

Swoop est un éditeur d'ontologies open source, développé par l'université du Maryland. Implémenté en Java, cet outil a été conçu pour traiter les formats RDF et OWL sous leurs différentes syntaxes. Parallèlement à ses fonctions d'édition, Swoop permet d'effectuer des raisonnements et propose un service de recherche des ontologies existantes.

Par ailleurs, nous pouvons citer l'outil gratuit KMGen dédié aux ontologies basées sur le formalisme KM (Knowledge Machine). Celui-ci intègre l'éditeur Emacs afin de créer et modifier les éléments de l'ontologie. De plus, KMGen peut être utilisé en mode multi-utilisateurs et à distance (réseau local ou internet).

Similaire sur ces derniers aspects, OntoWiki [AKSW, 2007] est une application Web de type « wiki » permettant de développer une ontologie de manière simple et collaborative. Cet outil est développé par le groupe de recherche AKSW[29] de l'université de Leipzig, également connu pour

22 Developing Ontology-Grounded Methods and Applications
23 Knowledge Interchange Format
24 Ontology Inference Layer
25 Web Ontology Language
26 World Wide Web Consortium
27 Darpa Agent Markup Language + Ontology Inference Layer
28 cf. Annexe 1
29 Agile Knowledge engineering and Semantic Web

leur projet DBPedia. Après quelques tests, cet outil s'avère plus orienté vers la population d'une ontologie que vers sa modélisation.

Terminons par l'éditeur d'ontologies le plus renommé et le plus utilisé parmi les experts en représentation des connaissances : l'environnement *Protégé* [Stanford, 2000]. Créé par les chercheurs de l'université de Stanford, *Protégé* est développé en Java, gratuit et open source. Il s'agit d'une plateforme d'aide à la création, la visualisation et la manipulation d'ontologies dans divers formats de représentation (RDF, RDFS, OWL, etc.). Ce logiciel peut également être utilisé en combinaison avec un moteur d'inférence (tel que RacerPro ou Fact) afin d'effectuer des raisonnements et d'obtenir de nouvelles assertions. De plus, de par la flexibilité de son architecture, *Protégé* est facilement configurable et extensible par les plugins développés au sein d'autres projets. Enfin, les créateurs de cet outil mettent l'accent sur l'aspect collaboratif dans la modélisation d'ontologies en proposant *Collaborative Protégé* et *WebProtégé*. Le premier est une extension intégrée à *Protégé* permettant à plusieurs utilisateurs d'éditer la même ontologie et de commenter les modifications effectuées par chacun. Un système de vote rend également possible la concertation sur tel ou tel changement. *WebProtégé* est une application Web légère et open source reprenant les principes de *Collaborative Protégé* dans le contexte du Web. Elle permet une édition d'ontologies collaborative et à distance.

Après observation des différents formats et outils précédents, nous avons fait le choix, pour notre stage, d'utiliser le logiciel *Protégé* ainsi que *WebProtégé* afin de modéliser notre ontologie au format OWL. En effet, ce format est le plus utilisé actuellement dans le domaine des ontologies et permet à notre ontologie d'être à la fois transportable d'un système à un autre, adaptée aux besoins du Web sémantique mais également compatible avec les standards définis par le W3C. Par ailleurs, l'environnement *Protégé* s'est révélé le mieux adapté à notre besoin pour plusieurs raisons. Tout d'abord, celui-ci supporte le format OWL que nous avons choisi d'utiliser. De plus, de par sa popularité, *Protégé* met à disposition de ses utilisateurs plusieurs possibilités de support telles que des listes de diffusion, un wiki, des cours personnalisés, des cours et tutoriels en ligne, ainsi qu'un système d'affiliation. Le wiki est accessible pour tous les niveaux et nous a été tout particulièrement utile pour l'installation et l'utilisation de *Collaborative Protégé* et *WebProtégé*. L'éditeur *Protégé* s'est avéré simple d'utilisation pour les tâches « classiques » de création d'une taxonomie de classes et d'ajout de propriétés. Nous avons également apprécié la possibilité d'y associer l'outil de visualisation GraphViz, afin d'obtenir une représentation sous forme d'arbre de notre ontologie. Enfin, sur les aspects « collaboration » et « utilisation à distance », nous avons privilégié *WebProtégé* pour son système de commentaires sur les différentes actions, son interface simple et personnalisable, sa gestion des comptes-utilisateur, etc.

3.3 - Méthodologie adoptée

Afin de construire notre ontologie du renseignement militaire, nous nous sommes reportés à la littérature du domaine et plus particulièrement aux articles détaillant les méthodes les plus courantes de modélisation [Noy et al., 2001] [Mizoguchi, 2003, 2004]. Cela nous a permis d'organiser notre travail en plusieurs étapes que nous détaillons ci-après.

3.3.1 - Tour d'horizon des ontologies disponibles

Notre objectif étant de développer une petite ontologie du renseignement militaire, nous avons tout d'abord mené quelques recherches pour faire le point sur les ontologies existantes. En effet, il nous est apparu intéressant de pouvoir éventuellement reprendre tout ou partie d'une modélisation déjà disponible. Nous avons pour cela observé des ontologies générales, dites « de haut niveau » mais également des ontologies du domaine militaire trouvées sur le Web grâce, notamment, au moteur de recherche sémantique Swoogle. Nous vous présentons, ici, celles qui se sont avérées intéressantes pour nos travaux.

Nous avons commencé par examiner les ontologies générales les plus connues et utilisées telles que SUMO[30], PROTON[31] ou encore COSMO[32].

SUMO [SUMO, 2004] a été modélisée en 2000 par la société californienne Teknowledge Corporation et reste l'une des ontologies privilégiées par l'organisation IEEE-SA[33] pour devenir un standard. Cette ontologie contient des concepts de haut niveau ou « meta-concepts » et peut servir de base à l'organisation d'encyclopédies par exemple. Celle-ci est à l'origine décrite en langage KIF mais a été récemment convertie en OWL.

L'ontologie PROTON [SEKT, 2005] est développée dans le cadre du projet de recherche européen SEKT[34] et comprend 4 modules : PROTON System (protons), PROTON Top (protont), PROTON Upper (protonu) et PROTON Knowledge Management (protonkm). Ceux-ci correspondent à différents niveaux de granularité et permettent une variété d'utilisation. Le module Protonu s'est avéré le plus intéressant pour nos travaux.

Un projet du groupe de recherche « Ontology and Taxonomy Coordinating » a donné naissance à l'ontologie COSMO. Celle-ci a été créée de façon collaborative et a pour but de représenter les concepts de base nécessaires aux diverses applications du domaine. Cette modélisation est disponible au format OWL et s'inspire de plusieurs autres ontologies de haut niveau telles que OpenCyc, SUMO, BFO[35] ou encore DOLCE[36]. La modélisation des lieux et organisations s'est avérée particulièrement intéressante pour le développement de notre ontologie.

OntoSem et SHOE[37], bien que moins renommées, constituent des ontologies générales intéressantes. La première a été conçue par Hakia, une société new-yorkaise spécialisée dans la recherche sémantique. L'ensemble d'ontologies Shoe, proposé au sein du projet Mindswap[38], est constitué de modélisations de divers niveaux dont nous avons retenu la plus générale et celle centrée sur les organisations.

Dans un second temps, nous nous sommes intéressés aux ontologies disponibles pour le domaine militaire. Tout d'abord, l'ontologie nommée « swint-terrorism », modélisée également dans le cadre de Mindswap et reprenant les concepts principaux nécessaires au domaine du terrorisme.

30 Suggested Upper Merged Ontology
31 PROTo ONtology
32 Common Semantic MOdel
33 Institute of Electrical and Electronics Engineers Standards Association
34 Semantically Enabled Knowledge Technology
35 Basic Formal Ontology
36 Descriptive Ontology for Linguistic and Cognitive Engineering
37 Simple HTML Ontology Extensions
38 Maryland Information and Network Dynamics Lab Semantic Web Agents Project

Puis, l'ontologie AKTiveSA [Smart, 2007], conçue lors du projet du même nom par le DIF DTC[39]. Au format OWL, celle-ci est dédiée à la description des contextes opérationnels militaires autres que la guerre.

Pour finir, d'autres modélisations plus spécifiques ont pu nous être utiles, telles que des ontologies spécialisées dans la description des évènements (SEM[40], Event Ontology) ou des entités géographiques (Geofile Ontology, MobiLife Space Ontology) [Minard, 2008].

Ces différentes observations nous ont permis de voir qu'aucune des ontologies trouvées ne correspondaient parfaitement au modèle de connaissances voulu et qu'il n'était donc pas adéquat de reprendre une de ces représentations en l'état. Toutefois, nous avons pu réutilisés certains éléments de ces ontologies dans les étapes qui suivent (cf. sections 3.3.2 et 3.3.3).

3.3.2 - Construction d'une taxonomie simple

Une seconde étape consiste à modéliser, pour commencer, une simple hiérarchie de classes. Pour cela, nous nous sommes aidés de l'organisation des différentes ontologies présentées plus haut mais également des recommandations de plusieurs standards OTAN[41] dits STANAG[42] [NATO, 2010].

Ceux-ci sont des accords de normalisation ratifiés par les pays de l'alliance définissant des normes pour permettre les interactions entre les différentes armées. Nous accordons une attention particulière aux catégories de l'intelligence définies par le STANAG 2433, connues sous le nom de « pentagramme du renseignement » (cf. Figure 3.1). Ce pentagramme reprend les éléments centraux du domaine du renseignement militaire et les définit comme ci-dessous.

La classe « Units » y est définie comme tout type de rassemblement humain partageant un même objectif, pouvant être hiérarchiquement structurée et divisée en sous-divisions. Il s'agit à la fois des organisations militaires, civiles, criminelles, terroristes, religieuses, etc.

La catégorie « Equipment » désigne toute sorte de matériel destiné à équiper une personne, une organisation ou un lieu pour remplir son rôle. Il peut s'agir d'équipement militaire ou civil, terrestre, aérien, spatial ou sous-marin.

L'élément « Places » regroupe les points ou espaces terrestres ou spatiaux, naturels ou construits par l'homme, pouvant être désignés par un ensemble de coordonnées géographiques.

La classe « Biographics » désigne les individus et décrit un certain nombre de propriétés associées telles que des éléments d'identification, des informations sur la vie sociale et privée, un ensemble de relations avec d'autres individus ou organisations, etc.

La catégorie « Event » décrit toute occurrence d'un élément considéré comme ayant de l'importance. Un évènement peut être divisé en plusieurs sous-évènements.

Notre objectif étant de développer une ontologie adaptée à ce domaine, nous avons fait le choix de nous concentrer sur ces 5 concepts et de ne pas modéliser les classes de plus haut niveau.

39 Data and Information Fusion Defence Technology Centre
40 Simple Event Model
41 Organisation du Traité de l'Atlantique Nord
42 STANdardization AGreement

Toutefois, ce standard reste trop généraliste et technique : il ne détaille pas les sous-classes du pentagramme et les diverses propriétés de classes évoquées doivent être triées et réorganisées.

Nous avons donc développé notre ontologie de haut en bas, en partant des concepts plus généraux vers les plus spécifiques. Nous avons effectué cette spécialisation en conservant les classes intéressantes des autres ontologies observées mais aussi en discutant avec des ingénieurs du département IPCC. Cet aspect collaboratif (grâce notamment à *WebProtégé*) constitue un point important car il nous a permis de mieux cerner les besoins des opérationnels du renseignement et de confronter nos idées à plusieurs avis.

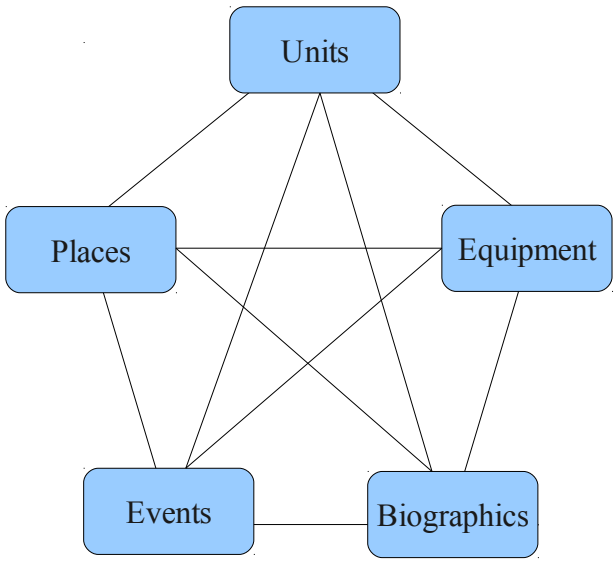


Figure 3.1 : Le pentagramme du renseignement

3.3.3 - Affinage selon les besoins du domaine

Une fois la taxonomie de base modélisée, la troisième étape a été d'affiner notre représentation du domaine : tout d'abord, en sous-spécifiant les classes supérieures[43], puis en déterminant les attributs[44] et relations[45] entre ces classes.

La création de sous-classes a été guidée par le contexte du renseignement militaire et ses éléments d'intérêt. Pour la classe « Equipment », nous nous sommes limités à décrire les différents types de véhicules et d'armes sans aller trop loin dans la spécification. La classe « Person » n'a pas nécessité plus de précision. En ce qui concerne le concept « Unit », nous avons choisi de distinguer deux sous-classes « Group » et « Organization » afin de différencier les groupements de personnes, point important dans le domaine visé. La classe « Place » a également été sous-typée en prenant en compte les besoins militaires, notamment par l'aspect stratégique des sous-classes du concept « Infrastructure ». Enfin, la modélisation de la classe « Event » s'est avérée une tâche essentielle compte tenu de l'importance de ces entités dans le renseignement et la veille militaire. Nous avons

43 cf. Annexe 2
44 cf. Annexe 3
45 cf. Annexe 4

pour cela réservé plus de temps à la spécification de la classe « MilitaryEvent », c'est-à-dire au choix et à l'organisation des différentes sous-classes en prenant en compte les observations préalables.

Afin de représenter au mieux le savoir du domaine, nous avons attribué, à chaque classe, un ensemble de propriétés permettant de les définir plus précisément en leur associant une valeur particulière. Cette valeur peut être une chaine de caractères, un nombre, une date, un booléen, etc. Comme nous l'avons déjà précisé plus haut, ces attributs sont héréditaires : ceux de la classe-mère sont automatiquement transmis aux classes-filles. Les 5 classes de plus haut niveau possèdent toutes les propriétés « name » et « alias ». Pour la classe « Event », nous avons précisé les dates de début et fin, la durée ainsi que le nombre de victimes. La classe « Person » possède un certain nombre d'attributs utiles pour le renseignement tels que la nationalité, la profession, l'âge, etc.

Enfin, parallèlement aux relations hiérarchiques de base, les concepts de l'ontologie sont liés par des relations sémantiques (« object properties »). Il s'agit de propriétés ayant pour valeur une classe de l'ontologie. Celles-ci ont également été choisies en fonction des besoins du renseignement militaire et en concertation avec les membres de notre équipe. De plus, nous avons fait le choix, pour plus de simplicité, de ne pas représenter les relations symétriques correspondantes : « isAffectedBy » pour « affects », par exemple. Ainsi par exemple, la classe « Event » entretient des relations sémantiques avec tous les autres éléments du pentagramme : un événement implique des participants de type « Person » ou « Unit » (« affects », « hasAgent ») ainsi qu'un instrument appartenant à la classe « Equipment » (« hasInstrument ») et se déroule dans un lieu associé à la classe « Place » (« takesPlace »). Un événement peut également être relié à d'autres évènements par des relations d'antécédence, succession, cause, conséquence, etc.

L'ontologie que nous venons de décrire constitue le modèle de connaissances qui servira de guide aux différentes étapes d'extraction d'information associées. En effet, une fois la structure de notre base de connaissances établie, il nous faut développer les outils qui permettront d'extraire d'un texte les différentes informations qui nous intéressent pour peupler cette ontologie (instances des différentes classes et liens existants entre ces instances). Comme nous l'avons déjà annoncé en introduction, notre travail d'extraction s'est organisé en 3 étapes :

- l'extraction d'entités nommées,
- l'extraction d'évènements,
- et l'extraction de relations entre entités.

4 - Extraction d'entités nommées

La phase d'extraction d'entités nommées consiste à mettre en place un système de détection et de typage des entités d'intérêt dans un texte. Notre objectif, ici, se limite à l'extraction des entités de type « Person », « Organization », « Date » et « Location », dans des textes écrits en anglais et en français. Nous détaillons par la suite notre façon de procéder pour réaliser cet objectif.

4.1 - Constitution de corpus

Le développement d'un système d'extraction d'entités nommées nécessite, au préalable, de rassembler un nombre suffisant de textes qui serviront non seulement de corpus d'observation (pour construire les règles) mais également de corpus de test. Le contexte de ce stage nous a naturellement guidé vers la collecte de corpus les plus proches possibles (au niveau de la forme et du contenu) de ceux utilisés par les experts du renseignement militaire. Nous nous sommes donc orientés vers des textes en anglais et en français, accessibles en sources ouvertes et dont le sujet est relatif au domaine militaire.

Une partie des campagnes d'évaluation présentées précédemment mettant leurs données à disposition gratuitement, nous avons réutilisé celles dont le thème et le format nous paraissaient les plus appropriés à notre tâche. Toutefois, les campagnes d'évaluation pour le français étant plus rares, nous n'avons eu accès qu'à des corpus de textes anglais.

Tout d'abord, nous avons utilisé les jeux de données de la campagne MUC 4 et plus particulièrement le corpus nommé « TST4-mixed-case ». Celui-ci correspond à une centaine de dépêches de presse (de taille plutôt courte) traitant toutes du terrorisme en Amérique Latine. Par ailleurs, la spécification TimeML (évoquée plus haut) a donné lieu à la création de nombreux corpus dont certains sont accessibles librement. Le premier que nous avons réutilisé se nomme TimeBank1.1 et comprend environ 200 articles de presse provenant d'une grande variété de sources et de domaines (données des campagnes DUC[46] et ACE, émissions transcrites des chaînes ABC et CNN, dépêches des journaux Wall Street, Associated Press et New York Times). Le second ensemble de textes, TempEvalTraining, a été conçu lors de la campagne Sem-Eval 2007 et correspond à une version modifiée du TimeBank. Enfin, nous avons récupéré le corpus AQUAINT[47], créé pour la campagne du même nom et composé d'une centaine d'articles de journaux. Celui-ci s'est avéré bien adapté à nos travaux car il reprend plusieurs sujets liés au domaine militaire (les bombardements au Kenya et en Tanzanie, l'affaire Elian Gonzalez à Cuba, les interventions de l'OTAN en Europe de l'Est, etc.).

Nous avons, d'autre part, pu compléter ces jeux de données par des textes collectés manuellement sur des sites internet liés au renseignement militaire : le site de l'ISAF[48] (force de sécurité de l'OTAN), celui de l'engagement du Canada en Afghanistan, mais encore tout site web de presse comme ceux de l'AFP[49], Reuters, etc.

46 Document Understanding Conference
47 Advanced QUestion Answering for INTelligence
48 International Security Assistance Force
49 Agence France-Presse

En ce qui concerne le traitement du français, nous avons eu de grandes difficultés à trouver des corpus disponibles gratuitement. Nous avons donc décidé de récolter manuellement quelques textes afin de constituer un petit corpus exploitable pour la création de notre système. Pour cela, nous avons privilégié les sites d'actualité et orienté nos recherches par des termes comme « Afghanistan », « Irak », « OTAN », etc. Les dépêches de l'AFP nous ont beaucoup servi de par leur format facilement exploitable (peu de mise en forme et de photos) mais aussi pour leur contenu très informatif (beaucoup d'entités nommées et d'évènements à extraire).

Pour finir, précisons que la manipulation de larges corpus dans l'environnement GATE nous a été facilitée par le système de « datastore ». Celui-ci permet, en optimisant la gestion de la mémoire, de manipuler facilement et rapidement une grande quantité de textes mais aussi de sauvegarder les modifications éventuelles faites sur ces données. Un « datastore » peut être créé à partir d'un dossier existant mais l'on peut également y stocké soit un corpus soit un texte seul depuis l'interface GATE. L'ensemble des textes contenus dans le « datastore » pourront ensuite être ouverts simultanément en un simple clic.

4.2 - ANNIE : observation et amélioration des résultats

Le système ANNIE, tel que nous l'avons déjà présenté (cf. section 2.2.4.3), est une chaine d'extraction d'entités nommées pour l'anglais. Nous avons donc choisi de la réutiliser dans le développement de notre outil d'extraction, tout en y apportant quelques modifications pour l'adapter à notre problématique et à notre représentation des connaissances. Pour cela, nous avons divisé notre travail en 3 phases : une première étape d'observation du fonctionnement et des résultats d'ANNIE, une seconde de modification et d'adaptation des règles, et enfin, une phase de comparaison des résultats des deux systèmes.

Dans un premier temps, nous nous sommes penchés sur le fonctionnement d'ANNIE et sur le module d'extraction qu'elle comprend, à savoir les règles linguistiques sous-jacentes et le transducteur JAPE associé. Nous présentons ici quelques limites de ce système et les points qui peuvent être améliorés. Précisons, pour commencer, qu'ANNIE permet d'annoter une quinzaine de types d'entités dans un texte : Phone, IP, Percent, Email, Organization, URL, Person, Money, Date, FirstPerson, Unknown, JobTitle, Title, Location. Nous pouvons d'ores et déjà discuter le statut d'entité nommée de certaines annotations telles que « Title » ou encore « FirstPerson ».

De plus, comme cela a déjà été dit, ce système d'extraction est basé sur l'association de listes de mots (« gazetteers ») et de règles JAPE. ANNIE comprend plus d'une centaine de « gazetteers » dont le plus long contient presque 24 000 mots. Le temps et le coût élevés de construction de ces lexiques nous amènent à nous interroger sur la nécessité de listes d'un telle taille et en si grand nombre [Mikheev et al., 1999]. Par ailleurs, nous avons pu observer que le transducteur est divisé en un nombre important de phases (fichiers .jape), chacune d'entre elles comprenant beaucoup de règles, dont certaines pourraient être, selon nous, regroupées. Il faut également noter que les règles d'ANNIE usent à plusieurs reprises d'annotations temporaires (« TempDate », « TempPerson », etc.) et nécessitent par conséquent l'utilisation répétée de code Java pour supprimer ces annotations. Nous pensons que ce point contribue à complexifier et obscurcir le fonctionnement du système.

Nous avons également examiné plus en détail les performances d'ANNIE sur nos corpus et n'avons pu que confirmer l'une des observations faites par les créateurs de ce système : l'extraction devient moins efficace lorsque le domaine du corpus traité s'éloigne de celui du corpus d'entrainement. Le tableau 4.1 reprend les observations faites par [Maynard et al., 2003] lorsqu'ils comparent l'extraction d'ANNIE sur des articles de presse économique (type de texte sur lequel ce système a été développé) et sur les corpus de la campagne ACE (textes de tous types).

	Précision	Rappel	F-mesure
Business news	89.0	91.8	90.4
Corpus ACE	55.8	59.7	57.8

Tableau 4.1 : Performances d'ANNIE

Présentons maintenant quelques exemples d'annotations erronées révélatrices des limites de ce système. Tout d'abord, l'extraction des entités de type « date » (en jaune ci-dessous) présente certaines anomalies :

- certains nombres à 4 chiffres sont annotés à tort : par exemple « No 3856 » ou encore « reference to 1966 graduates » ;
- certains suffixes numéraux posent problème : dans l'expression « By June 23rd 2009 », seul le mot « June » est repéré comme étant une date ;
- « 69 Jan 2007 » est annoté malgré l'anomalie au niveau du jour ;
- l'expression « June 30, 2009 » est annotée partiellement.

De plus, la détection des noms de personne (en bleu ci-dessous) n'est pas parfaite :

- le nom propre « Randolph » est étiqueté « Person » dans « in Randolph, Mass. » alors qu'il s'agit ici d'un lieu ;
- la même erreur est reproduite dans le syntagme « the San Pedro prison » ;
- l'on observe également des problèmes de portée des annotations : « Robert M. Gates » ou « Gen. Stanley A. McChrystal » ne sont pas annotés entièrement ;

Enfin, illustrons d'autres erreurs dans l'annotation des organisations (en vert ci-dessous) :

- dans l'expression suivante, l'adjectif « independent » est annoté à tort en tant qu'organisation : « There was no independent confirmation » ;
- même erreur pour le terme « TNT » qui est en réalité un type d'explosif : « 10 units of TNT » ;
- « The Economy and Finance Ministry » : ici, l'expression est bien une organisation, mais l'annotation est entravée par la conjonction « and ».

Cet ensemble de limites nous amène à une seconde phase de modification et d'adaptation

d'ANNIE selon les objectifs de notre stage.

Tout d'abord, comme nous l'avons déjà précisé, nous souhaitons extraire 4 types d'entités nommées, à savoir les noms de personne, d'organisation, de lieu et les dates. En ce qui concerne ce dernier type d'entité, nous pouvons distinguer les dates dites « absolues » des dates dites « relatives ». Les premières permettent à elles seules de se repérer sur un axe temporel (par exemple, « le 9 janvier 2002 » ou « l'année 2010 ») alors que les expressions de temps relatives (« hier matin », « le 5 mars », etc.) nécessitent des informations complémentaires provenant du texte ou du contexte extra-linguistique(cf. glossaire). Dans le cadre de ce stage, nous avons fait le choix de nous intéresser exclusivement aux dates absolues et de laisser de côté les dates relatives. En effet, l'extraction de ces dernières nécessite des traitements plus complexes, notamment en vue d'une normalisation. La normalisation des dates consiste à les convertir en un format standard facilitant la communication entre différents systèmes. Nous avons choisi ici la norme internationale ISO 8601[50] définissant une représentation numérique standard de la date et de l'heure. Nous avons effectué cette standardisation grâce à une règle JAPE : celle-ci récupère séparément les différents composants de la date (année, mois et jour), reconstitue la date complète au format ISO et ajoute celle-ci à l'annotation de type « Date ». Le tableau 4.2 ci-dessous rassemble quelques exemples de dates extraites et normalisées par notre outil dans des textes en anglais.

Date extraite	Date normalisée
1999-03-12	1999-03-12
1948	1948-01-01/12-31
April 4, 1949	1949-04-04
July 1997	1997-07-01/31
03-12-99	1999-12-03

Tableau 4.2 : Normalisation des dates

Un autre choix important a guidé le développement de notre système d'extraction : privilégier la précision par rapport au rappel. Il nous est apparu plus important dans le contexte du renseignement militaire d'éviter ce que l'on appelle plus communément le « bruit », c'est-à-dire l'extraction d'informations erronées. Nous avons donc orienté nos travaux pour maximiser la précision, c'est-à-dire l'extraction d'informations pertinentes pour mieux répondre à l'attente des opérationnels du domaine. Concrètement, cela se traduit par la construction de règles linguistiques dont les résultats sont plus sûrs et la mise à l'écart de règles pouvant entrainer de fausses annotations.

Une première étape a consisté à mettre de côté les règles et *gazetteers* d'ANNIE concernant des entités que nous ne souhaitons pas extraire (« URL », « Id », « Phone », etc.). Nous avons ainsi divisé par deux le nombre de *gazetteers*, simplifiant alors le système d'extraction et son éventuelle modification par une tierce personne. Cette diminution s'applique également au nombre de phases et de règles linguistiques ; pour l'extraction des dates, par exemple, nous n'avons conservé que les

50 http://www.iso.org/iso/fr/support/faqs/faqs_widely_used_standards/widely_used_standards_other/date_and_time_format.htm

règles détectant les dates absolues. Notre adaptation a nécessité par ailleurs la création de nouvelles listes de mots utiles aux nouvelles règles développées : une liste de grades militaires (pour la détection de personnes), une anti-liste contenant tous les termes pouvant être confondus avec des lieux, une liste de noms complets de personnes dont le prénom n'est pas commun et ne peut pas aider à la détection (ex : « Pol Pot ») ou encore une liste des noms de décennies (« nineties », « forties », etc.). Nous avons également dû supprimer certains termes des *gazetteers* existants comme « independent » qui, dans les exemples précédents, était annoté à tort comme une organisation.

Dans un deuxième temps, nous avons modifié les règles qui généraient du bruit et réorganisé l'ensemble des phases du transducteur. Nous présentons ci-dessous (cf. Figure 4.1) un exemple d'adaptation d'une règle présente dans la chaine ANNIE : nous avons (ligne 16) élargi la représentation d'une organisation par la possibilité d'avoir plusieurs noms en majuscule et une marque de possession associée. Remarquons également l'utilisation de la macro « UPPER » désignant de manière raccourcie tout nom comportant une majuscule.

ANNIE

```
1. Rule:OrgContext1
2. Priority: 1
3. // company X
4. // company called X
5.
6. (
7.  {Token.string == "company"}
8.  (
9.   (
10.      {Token.string == "called"}|
11.      {Token.string == "dubbed"}|
12.      {Token.string == "named"}
13.    )
14.   )?
15.  )
16.  ({Unknown.kind == PN}):org
```

ANNIE modifiée

```
1.   Rule: OrgCalled
2.   Priority: 70
3.   // company X
4.   // company called X
5.
6.   (
7.   {Token.string == "company"}
8.   (
9.      (
10.       {Token.string == "called"} |
11.       {Token.string == "dubbed"} |
12.       {Token.string == "named"}
13.    )
14.  )?
15.  )
16.  ((UPPER (POSS)?)[1,4]):org
```

Figure 4.1 : Adaptation d'une règle d'ANNIE

En ce qui concerne l'organisation des différentes phases d'extraction, nous nous sommes référés à l'article de [Mikheev et al., 1999] afin de parer à d'éventuelles ambiguïtés entre les entités « Person », « Organization » et « Location ». Le principe suggéré par ces auteurs est le suivant : afin d'éviter toute ambiguïté, il est conseillé d'exécuter en début de chaine les règles les plus sûres (basées sur le contexte), puis de typer les entités encore inconnues grâce aux *gazetteers* et de lever les ambiguïtés restantes dans une phase finale[51]. Nos propres observations nous ont amenées à choisir, en outre, un ordre de détection entre les 4 entités-cibles : nous typons, tout d'abord, les dates qui ne sont généralement pas confondues avec les autres entités ; puis, vient une phase de détection des organisations, suivie par les entités de type « Person » et, enfin, les noms de lieux (cf. Figure 1.1). En effet, nous avons pu, tout d'abord, observer que les entités de type « Organization » peuvent inclure des noms de personne ou de lieu et doivent donc être repérées en priorité afin d'écarter l'ambiguïté. Sur le même principe, certains prénoms présentant une homonymie avec des noms de

51 Méthode fréquente en Traitement Automatique des Langues

lieux, les entités de type « Person » doivent être extraites avant celles de type « Location ».

```
Phases:
first             //déclaration de macros générales
date              //détection des dates absolues
date2
organization      //détection des organisations par le contexte
organization2
person            //détection des personnes par le contexte
location          //détection des lieux par le contexte
unknown           //repérage des entités encore inconnues
org_gaz           //détection des organisations par gazetteers
person_gaz        //détection des personnes par gazetteers
loc_gaz           //détection des lieux par gazetteers
final             //phase de désambiguïsation finale
```

Figure 4.2 : Organisation des phases d'extraction d'ENs pour l'anglais

Pour finir, nous vous présentons un exemple comparatif des améliorations que nous avons apportées au système ANNIE (cf. figures 4.3 et 4.4). Nous pouvons y noter qu'ANNIE détecte plus de dates que notre propre système car il s'agit de dates relatives que nous avons choisies d'ignorer. Les erreurs d'ANNIE auxquelles nous avons apporté des solutions y sont entourées en rouge.

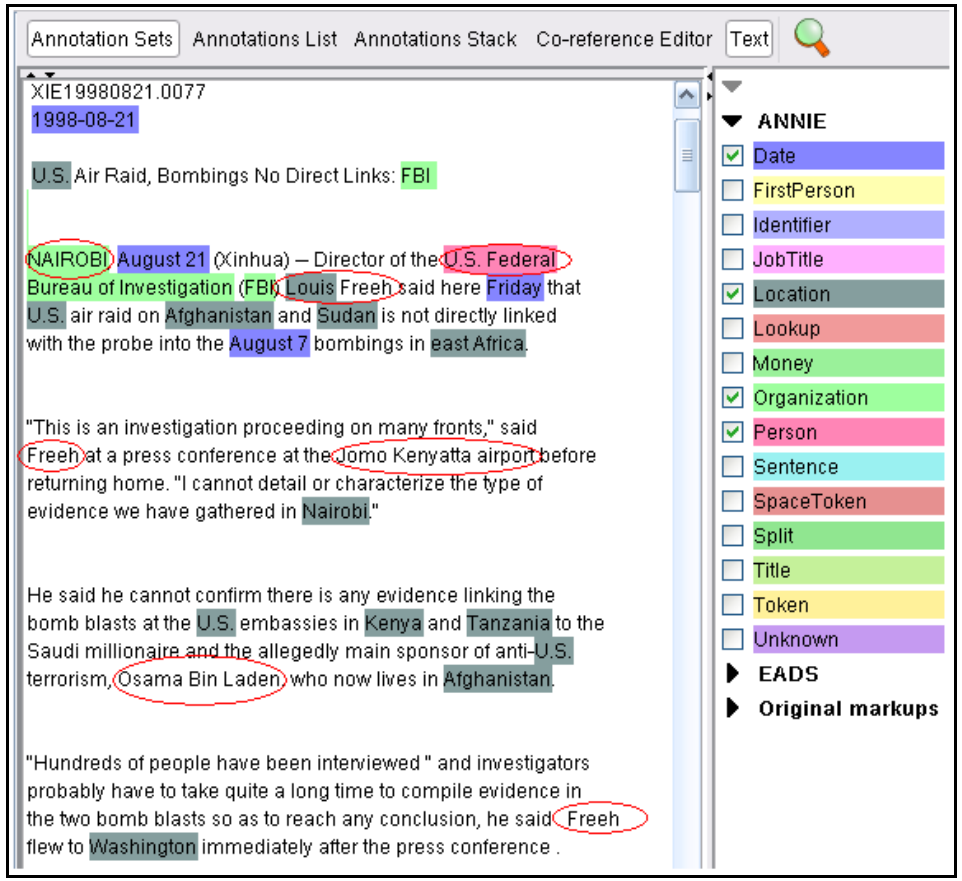


Figure 4.3 : Extraction d'ENs en anglais : ANNIE

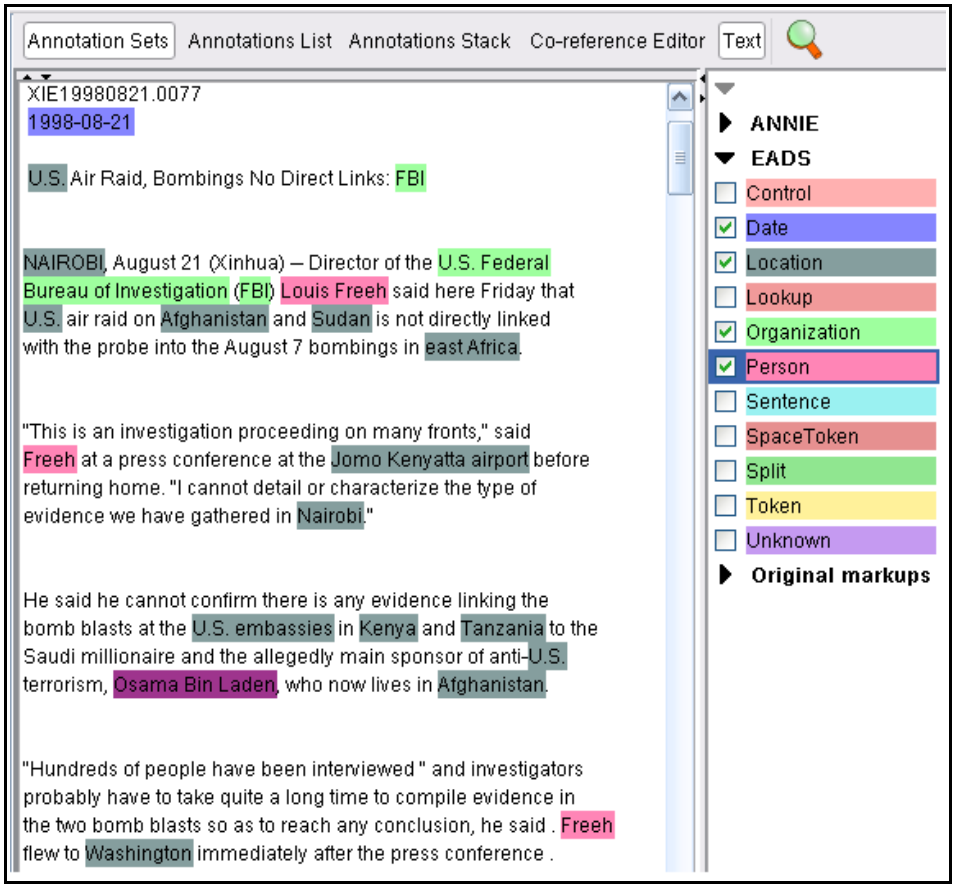


Figure 4.4 : Extraction d'ENs en anglais : ANNIE modifiée

4.3 - Traitement du français

L'environnement GATE propose une chaine d'extraction dédiée au traitement de la langue française, que nous avons choisi d'utiliser dans le cadre de nos travaux. Toutefois, nous verrons que l'extraction d'entités nommées pour le français a nécessité plus de travail car, à l'heure actuelle, le transducteur et les grammaires inclus dans cette chaine ne sont pas adaptés à cette langue.

Commençons par décrire brièvement les différents modules de cette chaine. Celle-ci est organisée de la même façon qu'ANNIE (cf. section 2.2.4.3) mais comprend quelques modules spécifiques au français. Y sont adaptés le découpage en mots (« tokenizer ») ainsi que l'analyse morpho-syntaxique. Cette dernière tâche est assignée à l'outil TreeTagger [Schmid, 1994] et nécessite donc son installation avec les paramètres adéquats. Cet outil a été développé par Helmut Schmid de l'université de Stuttgart et permet d'annoter une dizaine de langues. Les annotations obtenues indiquent pour chaque mot (« Token ») sa catégorie grammaticale ainsi que son lemme, informations essentielles pour le repérage des entités nommées.

La réutilisation de cette chaine a nécessité de créer notre propre transducteur basé sur nos règles linguistiques pour le traitement du français. En effet, une simple traduction des règles anglaises n'aurait pas suffi de par les nombreuses différences syntaxiques et typographiques entre ces deux langues. La première étape a consisté à déterminer les indices d'apparition des entités à extraire. Pour cela, nous nous sommes aidés de la littérature du domaine et notamment d'articles scientifiques tels que [McDonald, 1996]. Ici, McDonald divise ces indices en deux catégories : les indices internes et les indices externes. Les premiers font partie intégrante de l'entité nommée comme le prénom pour une entité « Personne » ou encore une abréviation indiquant le statut juridique d'une organisation (« SA », « SAS », etc.). Les seconds correspondent au contexte textuel de l'entité pouvant indiquer son type (dits « mots déclencheurs » ou « trigger words ») : « l'année » pour une date, « M. » pour une personne, etc. La découverte de ces indices s'est faite par observation des corpus décrits précédemment mais également par analogie avec le système anglais.

Notre méthode de construction des règles a été la suivante : après observation des différents contextes d'une entité, nous déterminons les différentes particularités des éléments de ce contexte (catégorie grammaticale, typographie, appartenance à tel *gazetteer*, etc.) afin de dégager des caractéristiques communes et ainsi construire une règle la plus générale et englobante possible. Les règles ainsi obtenues ont été organisées de façon similaire à ce que nous avons mis en place pour le traitement de l'anglais (cf. Figure 4.5). Nous nous sommes également inspiré des *gazetteers* déjà présents dans ANNIE pour construire leurs équivalents en français. Par ailleurs, nous avons repris et adapté le système de normalisation des dates développé précédemment pour les textes anglais.

L'élaboration de ces règles linguistiques a été suivie d'une phase de test pour détecter d'éventuelles erreurs d'annotation et corriger les règles en conséquence. La figure 4.6 montre un exemple de texte français où les entités nommées de type « date », « lieu », « organisation » et « personne » ont été annotées par notre système.

```
Phases:
first          //déclaration de macros générales
date           //détection des dates absolues
date2
organisation   //détection des organisations par le contexte
organisation2
organisation3
person         //détection des personnes par le contexte
location       //détection des lieux par le contexte
location2
unknown        //repérage des entités encore inconnues
org_gaz        //détection des organisations par gazetteers
person_gaz     //détection des personnes par gazetteers
loc_gaz        //détection des lieux par gazetteers
final          //phase de désambiguïsation finale
```

Figure 4.5 : Organisation des phases d'extraction pour le français

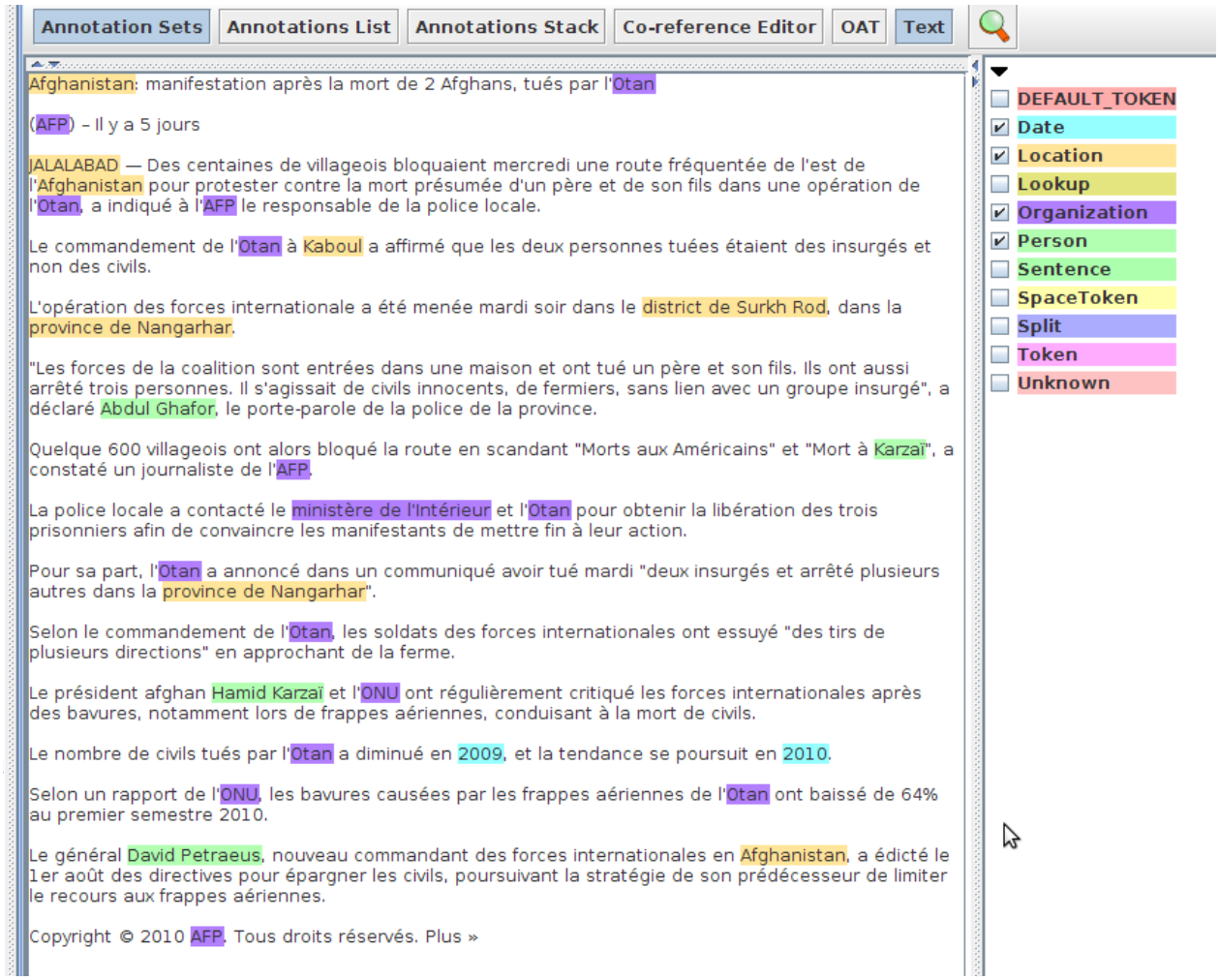


Figure 4.6 : Extraction d'ENs en français

5 - Extraction d'évènements

Dans le contexte de ce stage, le repérage des évènements reste une tâche essentielle bien qu'assez complexe car l'évènement constitue une entité de notre ontologie bien particulière. Nous pouvons la définir de façon générale comme une action (un « process ») à laquelle sont reliés un ou plusieurs participants ou circonstants. La spécification de la classe « Event » de notre ontologie présente les différents types d'évènements d'intérêt dans le cadre de nos travaux. Nous détaillons ci-dessous les étapes que nous avons suivies pour réaliser l'extraction de ces évènements dans des textes anglais et français.

5.1 - Définition de la méthode

5.1.1 - Observation des travaux existants

Afin de mettre en place notre méthode d'extraction, nous avons parcouru quelques publications traitant de l'extraction d'évènements et avons retenu quelques approches intéressantes.

Tout d'abord, nous avons pu constater deux grands courants principaux dans la définition du concept d'évènement : l'approche TimeML [Pustejovsky et al., 2003] et la vision de la campagne ACE [ACE, 2005]. Dans le cadre de la spécification TimeML, un évènement est défini comme tout élément pouvant être situé dans le temps. Le modèle ACE décrit, lui, un évènement comme une structure complexe impliquant différents arguments. Il en résulte que la première approche vise à annoter tous les évènements d'un texte, alors que la seconde a pour cibles uniquement les évènements d'intérêt pour une application donnée.

Le projet TARSQI[52] respectant la spécification TimeML a donné lieu au développement du système Evita[53], un outil de détection d'évènements. [Sauri et al., 2005] présente succinctement les principes théoriques sur lesquels repose cet outil ainsi que son fonctionnement général. Plusieurs points nous ont intéressée, dont, pour commencer, la description des éléments textuels pouvant être considérés comme des évènements : à savoir, les verbes, noms et adjectifs. Ces trois catégories de mots sont considérées comme étant les plus porteuses de sens. Les auteurs détaillent par la suite les différentes méthodes d'extraction associées à chaque type de « déclencheur » et plus particulièrement les caractéristiques textuelles et grammaticales à prendre en compte. Ainsi, pour la détection d'évènements déclenchés par un verbe, Evita opère un découpage en syntagmes verbaux et détermine pour chacun sa tête ; puis, vient une phase de tri lexical pour écarter les têtes ne dénotant pas un événement (verbes d'état, etc.) ; l'on tient ensuite compte des traits grammaticaux du verbe tels que la voix, la polarité (positif/négatif), la modalité, etc. ; et une analyse syntaxique de surface vient aider à l'identification des différents participants de l'événement. Ces mécanismes sont également complétés par des techniques statistiques que nous ne détaillerons pas ici.

La deuxième approche retenue est décrite dans [Aone et al., 2000] : il s'agit du système

52 Temporal Awareness and Reasoning Systems for Question Interpretation
53 Events In Texts Analizer

REES[54] permettant l'extraction de relations et d'évènements à grande échelle. Cet outil repose sur l'utilisation combinée de lexiques et de patrons syntaxiques pour la détection d'évènements principalement basés sur des verbes. Ces lexiques correspondent à une description syntaxique et sémantique des arguments de chaque verbe déclencheur d'évènement. Ces informations sont par la suite réutilisées au sein des patrons syntaxiques décrivant les différents contextes d'apparition d'un événement.

Enfin, [Ahn, 2006] présente une approche statistique de détection d'évènements selon les recommandations de la campagne ACE. Celui-ci décompose l'extraction en quatre tâches dont les deux premières nous intéressent particulièrement : le repérage des déclencheurs d'évènement et l'attribution des arguments. Le système détermine tout d'abord si la catégorie d'un mot en fait un possible candidat (verbe, nom, adjectif, etc.) et classe ensuite celui-ci parmi les types d'évènements définis lors de la campagne. Le repérage des différents arguments se fait par détection de paires entre le déclencheur et les autres entités de la phrase. Ces paires sont déterminées grâce à différents indices tels que la catégorie morpho-syntaxique des deux éléments, le type d'évènement, le type d'entité ou encore les relations de dépendances au sein de la paire.

5.1.2 - Élaboration d'une approche

Inspirés par les différentes techniques précédentes, notre but est d'élaborer une approche d'extraction d'évènements qui soit la plus générale possible et ceci à différents niveaux : elle se veut, tout d'abord, applicable à l'analyse de textes en plusieurs langues (français et anglais) et plusieurs domaines, mais aussi à différentes plateformes et environnements de traitement de la langue. Elle devra, enfin, pouvoir être adaptée à l'utilisation de plusieurs analyseurs syntaxiques. Pour plus de clarté, nous décrivons ici notre approche du point de vue du domaine militaire, même si celle-ci nous paraît aisément transposable dans d'autres domaines.

Tout d'abord, nous considérons comme possibles déclencheurs d'évènement les verbes et les noms, éléments porteurs de sens. En effet, bien que certaines des approches décrites plus haut considèrent les adjectifs comme tels, nous les jugeons peu significatifs dans le cadre de nos travaux. Le repérage de ces déclencheurs se fait par l'utilisation de listes (*gazetteers*) contenant, d'une part, des lemmes verbaux pour les déclencheurs de type « verbe » et, d'autre part, des lemmes nominaux pour les déclencheurs de type « nom ». Chaque *gazetteer* est associé à un type d'évènement c'est-à-dire à une classe de l'ontologie.

Après une phase de découpage en mots, un analyseur morphologique attribue à chaque « token » son lemme. Nous comparons ensuite chaque lemme aux listes de déclencheurs et, s'ils correspondent, le mot lemmatisé est annoté comme étant un déclencheur d'évènement. De plus, on lui associe la classe d'évènements qu'il représente. Ici, un déclencheur ayant plusieurs sens pourrait poser problème et provoquer une mauvaise classification de l'évènement repéré. Par exemple, le verbe « opérer » peut être un événement militaire mais peut aussi référer au domaine de la chirurgie. La durée de notre stage ne nous permettant pas de résoudre ces problèmes de polysémie, nous avons fait le choix, en amont, de développer des *gazetteers* de déclencheurs monosémiques.

Une fois les déclencheurs d'évènement repérés, il nous faut leur associer les différents participants impliqués. Il s'agit pour cela de repérer les relations entre le déclencheur et d'autres

54 Relation and Event Extraction System

entités de la phrase. Il nous paraît alors judicieux d'utiliser un analyseur syntaxique donnant les dépendances entre les différents éléments de la phrase. En effet, cela nous permet d'obtenir une meilleure précision par rapport à l'utilisation d'un simple découpage en syntagme (« chunking ») et de règles JAPE associées. Ici, nous faisons le choix de n'utiliser que les dépendances « principales » à savoir les relations « sujet », « objet », « préposition » et « modifieur de nom ». En effet, après observation de corpus annotés en dépendance, ces relations nous apparaissent suffisantes pour extraire les principaux participants.

Notre objectif ici étant de décrire une méthode générique d'annotation d'évènements, il nous paraît intéressant de fournir des moyens méthodologiques applicables à différents analyseurs syntaxiques. Pour cela, nous avons observé les structures de sortie de quelques analyseurs en dépendance (Stanford parser, Syntex, XIP[55]) afin d'en déduire une structure générale sur laquelle baser notre méthodologie. Tout d'abord, les analyses que nous avons pu observer s'appuient toutes sur une structure dite « prédicat-argument », c'est-à-dire une représentation des dépendances comme une relation entre un élément recteur et un élément dépendant. La plupart des « parsers » observés fournissent une représentation au format XML de leurs analyses mais certains proposent aussi leur propre formalisme. Il est toutefois possible de déduire une façon commune d'annoter les dépendances : la relation est mise entre les éléments centraux du syntagme-recteur et du syntagme-dépendant, plus communément appelés « têtes de syntagme » (cf. Figure 5.1). Les participants de l'évènement correspondent aux syntagmes associés à ces « têtes ». Ceux-ci peuvent être extraits par une analyse en constituants (souvent fournie avec l'analyse en dépendance) délimitant les différents syntagmes d'une phrase : syntagmes nominaux (SN ou « NP » en anglais), verbaux (SV ou « VP » en anglais), prépositionnels (SP ou « PP » en anglais) ou adjectivaux (SA ou « AP » en anglais).

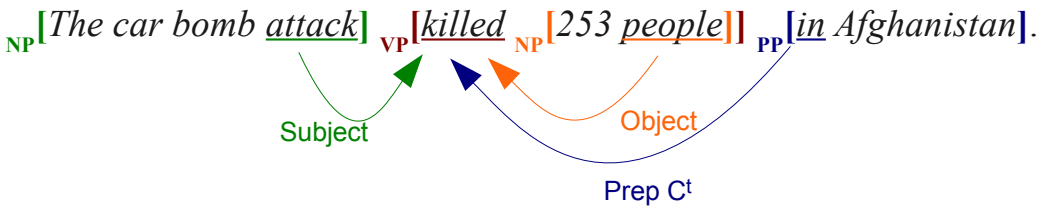


Figure 5.1 : Exemple d'analyse syntaxique

Une fois ces syntagmes rattachés à l'élément déclencheur par les relations de dépendance, il nous faut leur attribuer un rôle sémantique (« agent », « instrument », etc.). Cela est rendu possible par une étude de la structure argumentale du verbe ou du nom déclencheur : c'est-à-dire sa valence (nombre d'arguments) et les rôles sémantiques de ses différents actants. Si nous prenons l'exemple des verbes «tuer» et « mourir », nous remarquons qu'ils ont des valences différentes (2 et 1 respectivement) et que leurs sujets n'ont pas le même rôle sémantique : le premier sera « agent » et le second « patient ».

L'attribution de ces rôles sémantiques nécessite d'étudier, en amont, la construction des lemmes présents dans nos *gazetteers* [François et al., 2007]. Pour cela, nous avons choisi de constituer des classes argumentales, chacune d'elles correspondant à un type de construction verbale ou nominale. Cette indication sera conservée lors du repérage des éléments déclencheurs dans le

55 Xerox Incremental Parser

texte et réutilisée lors de l'extraction des participants. Il s'agit, par exemple, d'indiquer que pour le verbe « attaquer » le sujet syntaxique correspondra à l'agent alors que l'objet direct sera le patient de l'évènement. Cette analyse sémantique s'avère plutôt simple en ce qui concerne le sujet et l'objet d'un déclencheur mais nécessite quelques adaptations pour le traitement des syntagmes prépositionnels. Ceux-ci, qu'ils soient dépendants d'un verbe ou d'un nom, peuvent recouvrir différents rôles sémantiques pour une classe donnée. Il est alors nécessaire de prendre en considération la préposition à laquelle ils sont associés : ainsi, les prépositions « dans » ou « within » en anglais, indiqueront que le complément prépositionnel est de type « lieu » alors que « par » ou « by » précèderont un « agent ».

Notre extraction pourra être améliorée ultérieurement en prenant en compte d'autres paramètres tels que la voix (passive ou active) du déclencheur, la polarité de la phrase (négative ou positive), la modalité mais aussi les phénomènes de valence multiple.

5.2 - Implémentation dans GATE

Présentons maintenant une réalisation de notre approche d'extraction d'évènements dans l'environnement GATE. Précisons, au préalable, que pour chaque événement détecté, nous nous fixons pour objectif d'extraire, s'ils sont présents, les participants/circonstants suivants : la date, le lieu, l'agent, le patient et l'instrument.

5.2.1 - Traitement de l'anglais

Dans les sections suivantes, nous décrivons les différentes étapes de la construction de notre outil d'extraction d'évènements pour des textes de langue anglaise.

5.2.1.1 - Installation des plugins GATE nécessaires

La mise en œuvre de notre approche dans GATE a nécessité l'utilisation d'outils supplémentaires proposés dans cet environnement et donc l'installation de nouveaux plugins.

Tout d'abord, l'analyseur morphologique (« GATE Morphological Analyser ») proposé au sein du plugin « Tools » nous a permis d'obtenir, pour chaque mot du texte traité, le lemme associé (« root »). Rappelons que cette information nous est nécessaire pour comparer chaque mot lemmatisé avec les termes présents dans les *gazetteers*.

Cela est rendu possible par le système de « Flexible Gazetteer » : ce type de ressource (également fourni par le plugin « Tools ») permet de comparer une liste de mots, non plus à une simple chaine du texte, mais également à certaines propriétés du mot et donc, entre autres, à son lemme.

D'autre part, comme nous l'avons précisé auparavant, le repérage des participants de l'évènement nécessite un découpage de la phrase en syntagmes. Pour cela, nous avons utilisé les ressources « NP Chunker » et « VP Chunker » disponibles respectivement dans les plugins « Tagger NP Chunking » et « Tools ». Le premier permet une analyse de la phrase en groupes nominaux. Le second fournit les groupes verbaux en indiquant quelques informations supplémentaires telles que le temps et la voix du verbe-recteur, le type de syntagme (fini, non-fini, participe, etc.), la polarité, etc.

Enfin, un composant essentiel de notre chaine d'extraction est l'analyseur syntaxique en dépendance. Après avoir examiné les différentes solutions proposées dans GATE, nous avons opté pour l'utilisation du « Stanford parser » [De Marneffe et al., 2008]. Celui-ci propose deux types d'analyse syntaxique : en constituants et en dépendance. Basé sur des techniques probabilistes, il a été développé à l'université Stanford en Californie et permet de traiter l'arabe, le chinois, l'anglais et l'allemand. Cet analyseur peut être paramétré pour choisir, entre autres, le type d'analyse voulu.

5.2.1.2 - Constitution des *gazetteers*

Nous avons, pour commencer, listé pour chaque type d'évènement militaire (i.e. pour chaque sous-classe de la classe « Event ») l'ensemble des lemmes susceptibles de le réaliser dans un texte. Nous avons fait cela à la fois pour les lemmes verbaux et nominaux. Concrètement cela se traduit par la création dans notre chaine de deux « flexible gazetteers » : le premier pour les déclencheurs de type « verbe » et le second pour les déclencheurs de type « nom ». Nous avons, pour construire ces listes, observé plusieurs textes du domaine militaire pour repérer les différents verbes et noms associés à chaque type. Nous nous sommes également aidée de dictionnaires de synonymes tels que celui du CNRTL[56] ou du CRISCO[57] à l'université de Caen. De plus, pour pouvoir déterminer les rôles sémantiques de chaque argument de l'évènement, nous avons associé à chaque lemme verbal une indication sur sa structure argumentale. Nous avons pour cela déterminé 5 classes argumentales et associé à chacune d'elles une numérotation (cf. Tableau 5.1). La figure 5.2 constitue un exemple de liste de lemmes verbaux associés à la classe « DamageEvent » et des classes argumentales attribuées à chaque élément de cette liste. L'exécution de ces deux types de *gazetteers* donnent lieu à des annotations de type « Lookup » indiquant le type d'évènement déclenché ainsi que, pour les lemmes verbaux, la classe argumentale à laquelle ils appartiennent.

	Voix active	Voix Passive
Classe 1	Sujet = Agent Objet = Patient	Sujet = Patient Ct Prep « by »[58] = Agent
Classe 2	Sujet = Agent Objet = Instrument	Sujet = Instrument Ct Prep « by » = Agent
Classe 3	Sujet = Instrument Objet = Patient	Sujet = Patient Ct Prep « by » = Instrument
Classe 4	Sujet = Agent Objet = Lieu	Sujet = Lieu Ct Prep « by » = Agent
Classe 5	Sujet = Patient	Ø

Tableau 5.1 : Classes argumentales

56 Centre National de Ressources Textuelles et Lexicales
57 Centre de Recherches Inter-langues sur la Signification en COntexte
58 Complément prépositionnel introduit par la préposition « by »

```
destroy§struct=4 // § est un séparateur
blow§struct=1
break§struct=1
crush§struct=4
hit§struct=1
damage§struct=4
shatter§struct=4
```

Figure 5.2 : Gazetteer associé à la classe « DamageEvent »

5.2.1.3 - Développement des règles linguistiques

Une fois les déclencheurs d'évènements repérés, il nous faut construire un transducteur JAPE exécutant un ensemble de règles linguistiques permettant d'extraire les différents arguments de l'évènement. Ces règles exploitent les informations fournies préalablement par les autres modules de la chaine d'extraction (cf. Figure 5.3) et sont organisées en plusieurs phases (cf. Figure 5.5).

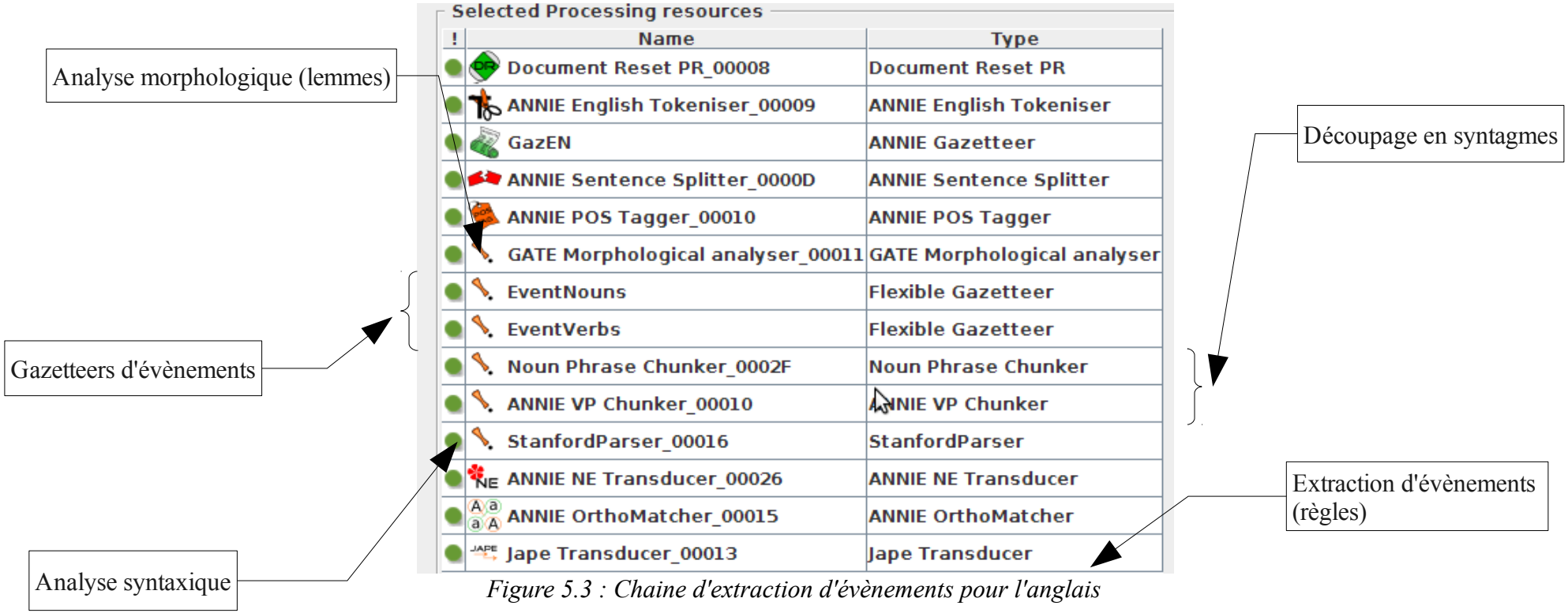


Figure 5.3 : Chaine d'extraction d'évènements pour l'anglais

```
Phases:
preProcess  //pré-traitements sur les annotations des modules précédents
NPVerbs1    //repérage des groupes nominaux associés au déclencheur verbal
NPVerbs2
NPNouns     //repérage des groupes nominaux associés au déclencheur nominal
PP1         //repérage des groupes prépositionnels associés aux déclencheurs
PP2
SemRoles    //attribution des rôles sémantiques aux arguments du déclencheur
Final       //nettoyage des annotations temporaires
```

Figure 5.4 : Organisation des phases d'extraction d'évènements pour l'anglais

La première phase opère plusieurs traitements pour adapter les annotations mises précédemment aux phases d'extraction qui vont suivre. Les groupes nominaux sont triés pour ne conserver que ceux qui présentent un intérêt pour l'extraction, à savoir ceux étant impliqués dans une relation « sujet », « objet » ou « modifieur » avec un élément déclencheur. Les groupes prépositionnels sont modifiés pour indiquer l'identifiant de leur tête ainsi que la préposition auxquels ils se rattachent. Enfin, les déclencheurs verbaux et nominaux sont annotés avec des propriétés nécessaires pour la suite : pour les premiers, la voix, la classe argumentale, la classe sémantique[59] et l'ontologie ; pour les seconds, la classe sémantique et l'ontologie.

Dans un deuxième temps, trois phases associent à chaque déclencheur les groupes nominaux avec lesquels il entretient une dépendance. On obtient alors une annotation de type « Event » sur le déclencheur, ayant pour propriétés les différents arguments repérés jusqu'à présent et indiquant pour chacun d'eux s'il s'agit d'une entité nommée de type « Person », « Organization », « Date » ou « Location ».

Par la suite, on associe chaque déclencheur à ses compléments prépositionnels en précisant pour chacun la préposition qui les régit et s'il s'agit d'une entité nommée. Ces informations sont ajoutées aux annotations de type « Event » mises par les phases précédentes.

Pour finir, une phase sert à l'attribution de rôles sémantiques aux différents participants/circonstants de l'évènement. A l'aide de la structure argumentale donnée par les *gazetteers* et des informations reprises par le tableau 5.1, l'on ajoute à l'annotation « Event » les propriétés « agent », « patient », etc. ayant pour valeur les syntagmes correspondants. De plus, lorsqu'un syntagme-participant correspond à une entité nommée, nous pouvons prendre en compte son type pour déterminer le rôle sémantique du syntagme : un syntagme de type « Date » se verra attribuer le rôle de circonstant temporel de l'évènement. Enfin, les informations temporaires nécessaires à l'extraction sont supprimés lors d'une dernière étape de nettoyage.

Nous présentons ci-dessous un exemple de texte anglais où les évènements militaires ont été annotés par notre outil (cf. Figure 5.5), ainsi que la vue détaillée des différentes caractéristiques de l'évènement « killed » (entouré en rouge dans le texte) (cf. Figure 5.6).

59 Type d'évènement défini dans l'ontologie

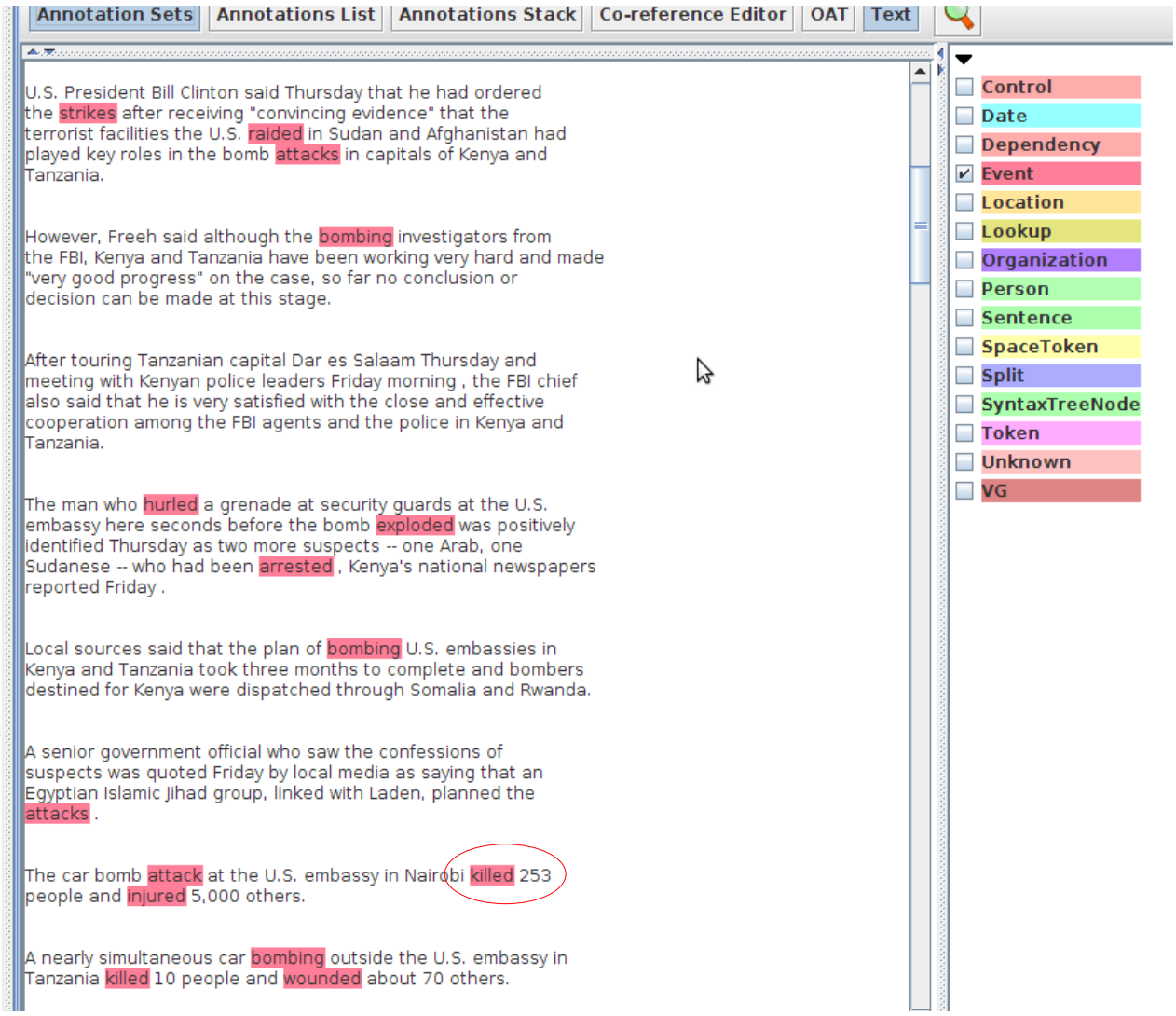


Figure 5.5 : Extraction d'évènements en anglais

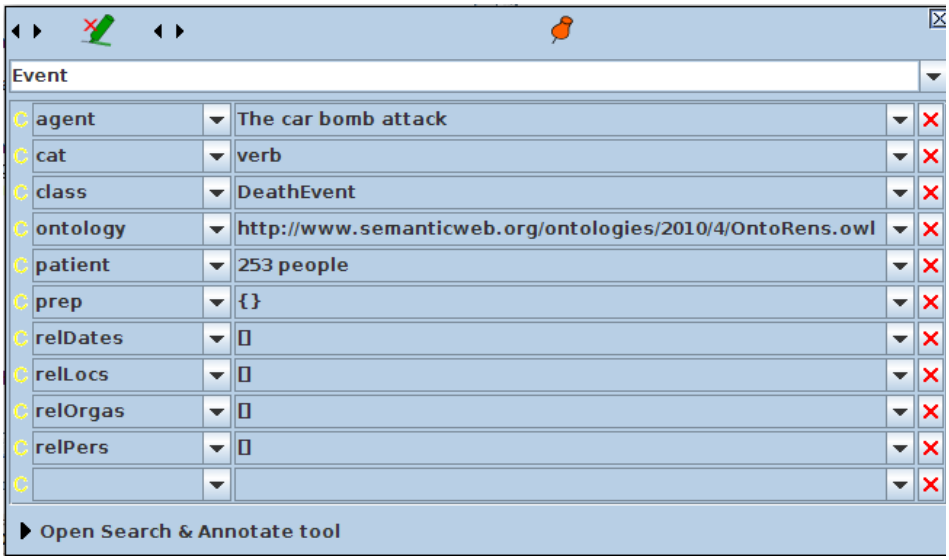


Figure 5.6 : Exemple d'une annotation de type "Event"

5.2.2 - Traitement du français

L'analyse syntaxique constituant une composante essentielle à notre approche d'extraction, le traitement du français repose sur la disponibilité d'un analyseur en dépendance pour cette langue. Nos recherches dans ce sens n'ayant pas abouti, nous n'avons pu implémenter dans GATE un système d'extraction d'évènements pour le français. Toutefois, nous tenons à présenter les différents analyseurs trouvés lors de nos investigations et les raisons pour lesquelles ceux-ci n'ont pu servir à la construction de notre outil.

Tout d'abord, précisons qu'à l'heure actuelle, l'environnement GATE ne propose pas de plugin d'analyse syntaxique dédié à la langue française. Nous nous sommes donc mis en quête d'un analyseur externe dont les résultats pourraient être réutilisés au sein de GATE, sans entraîner trop d'adaptations techniques.

Nos recherches nous ont mené vers des analyseurs syntaxiques développés par des groupes de recherche du domaine et distribués librement à des fins de recherche. Le premier auquel nous nous sommes intéressée se nomme LASAF[60] et est développé au sein du GREYC, laboratoire de Caen. Nous avons donc pris contact avec les chercheurs de l'équipe ISLanD[61] responsables de cet analyseur, mais celui-ci ne s'avère fonctionner que sur un système d'exploitation Mac OS, système que nous n'avons pas à notre disposition dans le cadre de ce stage. Une seconde solution est l'analyseur XLFG de l'équipe-projet ALPAGE[62] de l'INRIA[63]. Il s'agit d'un analyseur syntaxique pour le français qui repose sur le formalisme des grammaires LFG[64]. Toutefois, après installation de cet outil, son fonctionnement s'est vu entravé par de nombreux problèmes techniques.

60 Logiciel d'Analyse Syntaxique Automatique du Français
61 Interaction, Sémiotique : LANgue, Diagrammes
62 Analyse Linguistique Profonde A Grande Échelle
63 Institut National de Recherche en Informatique et en Automatique
64 Lexical Functional Grammar

Nous nous sommes alors orientée vers Syntex, analyseur syntaxique créé par Didier Bourigault au sein de l'ERSS[65] à Toulouse. Celui-ci nous a indiqué que cet analyseur n'est plus distribué, même pour la recherche, et nous a proposé de lui fournir un corpus afin de nous retourner quelques résultats. Nous avons donc constitué un petit corpus d'articles de l'AFP en français et obtenu l'analyse en dépendance au format spécifique à Syntex[66]. L'observation de ces résultats nous a permis d'élaborer l'approche générale décrite précédemment, toutefois, leur réutilisation dans GATE demandait trop d'efforts techniques pour pouvoir être effectuée durant ce stage.

Pour finir, nous avons également pu obtenir les résultats de l'analyseur XIP sur ce même corpus[67] grâce à une licence disponible à des fins de recherche au sein de notre université. Pour des raisons identiques à celles que nous venons d'exposer, ces résultats nous ont uniquement servi à découvrir un format général d'analyse en dépendance.

5.3 - Analyse qualitative et améliorations possibles

L'extraction d'évènements que nous venons de décrire comporte certaines limites et pourrait être, avec un peu plus de temps, améliorée par plusieurs moyens.

Le point essentiel reste l'implémentation de la méthode pour le traitement de textes en français. En effet, l'absence d'analyseur syntaxique disponible pour cette langue a constitué un réel obstacle au développement de notre outil. Face à cela, il est possible de développer soi-même un analyseur en dépendance (tâche longue et complexe) ou d'obtenir une licence payante.

Toutefois, même si un tel outil reste essentiel, notre extraction d'évènements en anglais montre que cela ne résout pas tous les problèmes. Premièrement, l'analyseur en dépendance Stanford n'est pas infaillible (environ 70 % de précision et 60% de rappel) et, face à des phrases aux structures complexes, ne permet pas d'annoter toutes les relations nécessaires à l'extraction. Ainsi, dans la phrase suivante, l'analyseur Stanford ne parvient pas à déterminer que « the sixth British soldier » est le sujet de « killed ».

On Wednesday, Britain's ministry of defence said one of its troops was shot dead in a firefight with Taliban-led insurgents in Helmand province, ***the sixth British soldier*** *killed in Afghanistan this month.*

Nous avons également remarqué une certaine limitation dans l'attribution de la voix (active, passive) aux différentes phrases d'un texte, élément indispensable pour attribuer les rôles sémantiques aux participants. Cela est particulièrement fréquent en ce qui concerne les titres : l'analyseur n'attribue pas de voix aux phrases telles que la suivante.

NATO soldier killed in southern Afghanistan

Même si un analyseur performant améliore grandement l'extraction d'évènements, notre système peut être amélioré par d'autres moyens. Tout d'abord, dans le cadre de ce stage, nous n'avons pas eu le temps d'exploiter toutes les dépendances trouvées par le *Stanford parser* et nous nous sommes limitée aux relations « sujet », « objet », « modifieur de nom » et « complément prépositionnel » (cf. section 5.1.2). Cependant, l'analyse en dépendance est plus complète et propose d'autres relations telles que « aux » ou « auxpass » (temps composé), « iobj » (objet indirect

65 Équipe de Recherche en Syntaxe et Sémantique
66 cf. Annexe 9
67 cf. Annexe 10

d'un verbe), « prt » (verbe à particule), « tmod » (modifieur temporel), etc.

Quant à l'attribution de rôles sémantiques aux compléments prépositionnels, elle pourrait être améliorée par une meilleure analyse des prépositions dans les deux langues grâce notamment à la reprise de travaux sur la sémantique prépositionnelle tels que ceux de [Cadiot, 2002].

Concernant la détection d'évènements déclenchés par des noms, il faudra étudier plus en détail leur structure argumentale afin de déterminer des classes sur le même principe que pour les verbes. Dans ce cas précis, l'analyse des prépositions sera précieuse car, en anglais et français, la majorité des arguments d'un prédicat nominal est introduite par une préposition. Nous devrons également améliorer l'analyse des modifieurs de noms auxquels nous n'avons pas encore attribué de rôles sémantiques.

Pour finir, comme nous l'avons déjà précisé plus haut, notre outil peut être raffiné en complétant les listes de déclencheurs (verbaux et nominaux) et en prenant en compte la négation et la modalité (interrogation, supposition, etc.) dans la détection des évènements.

6 - Extraction de relations

La dernière tâche à accomplir durant ce stage est la détection de relations entre entités nommées et plus particulièrement des relations « Personne-Personne » et « Personne-Organisation ». Nous détaillons ci-après la méthode employée ainsi que le fonctionnement du système développé.

6.1 - Méthode d'extraction

Notre outil d'extraction devant être en adéquation avec notre modèle de connaissances, nous avons commencé par recenser les différentes relations « Personne-Personne » et « Personne-Organisation » présentes dans notre ontologie de domaine. Ainsi, notre extraction sera centrée autour de quatre types de relations : lien de parenté (« isFamilyOf »), appartenance à une organisation (« isMemberOf »), direction d'une organisation (« isLeaderOf »), lien divers (« isLinkedTo »). Comme l'indique notre ontologie, la première est une relation entre deux personnes, les deux suivantes lient une personne à une organisation (« Unit ») et la dernière peut recouvrir les deux cas précédents.

Notre approche s'inspire des deux types d'extraction déjà menées, à savoir l'extraction d'entités nommées et l'extraction d'évènements. De la première nous avons gardé le principe de règles linguistiques contextuelles et de la seconde, la constitution de listes de lemmes pour chaque type de relation.

Notre méthode consiste à définir, par observation de corpus du domaine visé, un ensemble de mots déclencheurs pour chaque type de relation de l'ontologie. Pour un meilleur rappel dans l'extraction et un temps de développement plus court, nous faisons le choix de constituer des listes de lemmes plutôt que des listes de formes fléchies. Ceux-ci sont ensuite comparés aux lemmes des mots du texte à traiter qui, s'il y a correspondance, sont annotés en tant que possibles déclencheurs de relation.

Dans un deuxième temps, un ensemble de règles linguistiques teste le contexte de chaque déclencheur pour déterminer la présence d'une relation du type souhaité (« Personne-Personne » ou « Personne-Organisation »). Lorsque le contexte concorde, une annotation, dont le type est donné par l'élément déclencheur, est créée sur l'ensemble des éléments de la relation (arguments et lien).

6.2 - Implémentation dans GATE

L'approche d'extraction détaillée ci-dessus a été réalisée dans l'environnement GATE pour le traitement de textes anglais.

Notre chaine d'extraction de relations utilise pour base les différents modules employés dans l'extraction d'entités nommées. Nous y avons ajouté un analyseur morphologique, un « flexible gazetteer » ainsi que notre transducteur JAPE contenant les règles d'extraction (cf. Figure 6.1).

La première étape pour développer notre outil a été d'observer les corpus du domaine militaire afin de repérer les éléments lexicaux déclencheurs des relations de notre ontologie. Ainsi, pour chaque type de relation, nous avons créé une liste de lemmes susceptibles de l'annoncer[68]. Nous obtenons donc autant de *gazetteers* que de relations-cibles. Nous avons, par la suite, créé dans GATE les ressources nécessaires : à savoir l'analyseur morphologique pour obtenir le lemme de chaque mot du texte et le « flexible gazetteer » associé afin d'obtenir des annotations de type « Lookup » indiquant le type de relation.

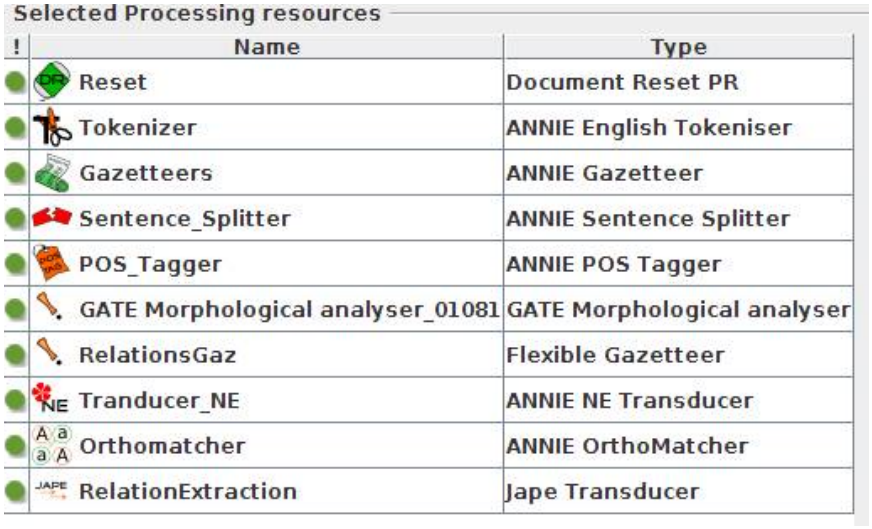


Figure 6.1 : Chaine d'extraction de relations pour l'anglais

Ces premières annotations nous ont permis de tourner notre attention vers les différentes réalisations linguistiques de ces relations. De la même façon que pour les entités nommées, ceci nous a permis de dégager des contextes généraux d'apparition de ces relations et ainsi construire les règles JAPE associées. Celles-ci sont comprises dans une seule phase d'extraction et font appel à du code Java en partie droite afin de récupérer diverses informations d'annotations préalables de type « Person », « Organization », « Lookup » ou encore « Token »[69].

L'annotation finale a pour type le nom de la relation dans notre ontologie, à savoir « isFamilyOf », « isMemberOf », « isLeaderOf » ou « isLinkedTo ». L'ajout de propriétés à cette annotation permet d'indiquer plusieurs informations au sujet de la relation détectée. La figure 6.2 nous montre qu'une annotation indique les identifiants des entités nommées impliquées ainsi que le mot déclencheur mais également la classe et l'ontologie associées.

68 cf. Annexe 5
69 cf. Annexe 6

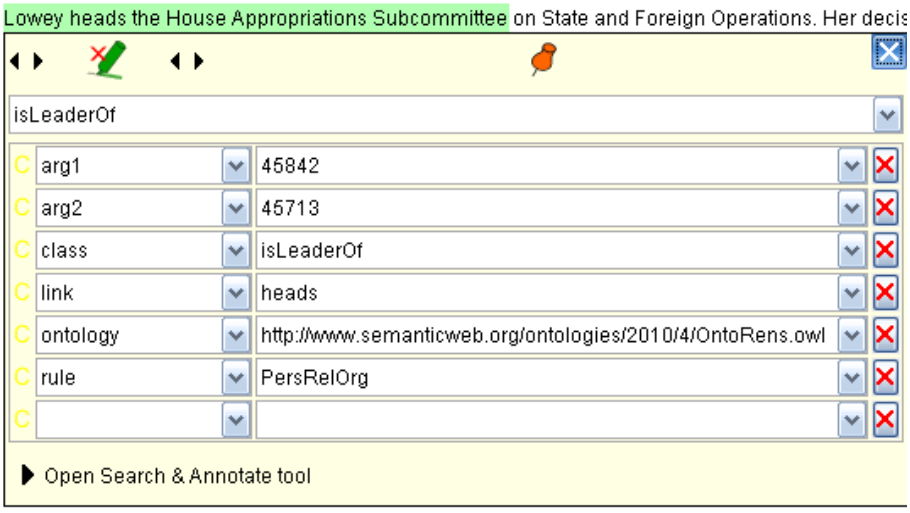


Figure 6.2 : Exemple d'annotation d'une relation en anglais

6.3 - Analyse qualitative et améliorations possibles

Notre extraction de relations n'est qu'à l'état de prototype, l'implémentation que nous avons réalisée n'a, pour l'instant, que vocation à montrer la faisabilité de notre approche. Notre système devra être amélioré sur plusieurs points et implémenté pour les deux langues en question.

Pour commencer, l'extraction d'information dépendant étroitement du modèle de connaissances associé (i.e. notre ontologie de domaine), nous devrons affiner notre représentation du domaine afin de définir précisément quelles relations il est nécessaire d'extraire. Il sera préférable d'interagir avec des opérationnels du renseignement militaire pour cerner avec précision leurs besoins et compléter notre ontologie en conséquence.

Par ailleurs, nous extrayons actuellement uniquement les relations entre personnes et organisations et nos résultats sont donc soumis aux performances de notre outil dans l'extraction des entités nommées. Une erreur dans la détection d'une entité de ce type peut entraîner une mauvaise extraction de relations. Dans l'exemple qui suit, une relation de type « isLeaderOf » est repérée entre la personne « Lowey » et l'organisation « House Appropriations Subcommittee ». La délimitation erronée de l'entité de type « Organization » provoque une extraction de relation partiellement correcte.

<PERS>Lowey</PERS> heads the <ORG>House Appropriations Subcommittee</ORG> on State and Foreign Operations.

Pour finir, notre système est basé sur de simples patrons linguistiques et pourra être complété par une analyse syntaxique en dépendance telle que celle employée pour l'extraction d'évènements. En effet, l'utilisation de simples règles ne permet pas d'extraire des relations « à distance » au sein de la phrase. [Nakamura-Delloye, 2010] propose une méthode d'extraction de relations entre entités nommées basée sur une analyse en dépendance. Son idée est de repérer les chemins syntaxiques entre deux entités (i.e. l'ensemble des relations de dépendance qu'il faut parcourir pour relier ces deux entités) afin de construire par généralisation un ensemble de patrons

de relations syntaxiques spécifique à tel type de relation sémantique. Cette approche nous paraît intéressante et envisageable pour améliorer les performances de notre extraction. L'extrait suivant constitue un exemple de relation plus complexe difficile à détecter uniquement grâce à une règle linguistique mais qu'une analyse syntaxique contribuerait à repérer.

<PERS>Senator Carl Levin</PERS>, a Democrat from Michigan who heads the <ORG>Armed Services Committee</ORG> [...]

7 - Évaluation des résultats

Une dernière étape du stage est l'évaluation des résultats de notre outil. Nous avons pour cela choisi de nous limiter à l'évaluation de l'extraction d'entités nommées en anglais et en français. En effet, cette extraction s'avère moins complexe à évaluer que les extractions d'évènements ou de relations et plus facilement comparable aux résultats d'autres systèmes. De plus, à l'heure actuelle, il s'agit de la tâche la plus aboutie de notre stage : l'extraction d'évènements et de relations étant encore à l'état de prototypes et implémentées uniquement pour l'anglais. Nous présentons, par la suite, les différentes phases de notre évaluation ainsi que les conclusions que nous avons pu faire au vu des résultats obtenus.

7.1 - Protocole d'évaluation

L'évaluation de notre travail a, tout d'abord, nécessité de faire plusieurs choix concernant le type d'évaluation à mettre en place. Deux solutions se sont alors présentées : réutiliser les données d'une campagne d'évaluation existante ou créer notre propre système d'évaluation. Le premier cas impliquait de trouver un corpus du domaine militaire, où les entités nommées ont été préalablement annotées, et accompagné de scripts de « scoring ». Nous avons, pour cela, examiné les données librement diffusées des campagnes d'évaluation décrites précédemment (cf. section 2.2.2). Nous n'avons, toutefois, pas trouvé de données satisfaisant les trois conditions nécessaires citées plus haut. En conséquence, nous avons opté pour la deuxième solution, c'est-à-dire développer notre système d'évaluation.

Nous avions, tout d'abord, besoin de corpus de textes du domaine militaire en anglais et en français. Dans le premier cas, nous avons choisi de réutiliser le corpus AQUAINT (cf. section 4.1) dont le domaine et la taille (une centaine de textes) correspondent à notre besoin. Pour le français, n'ayant pas trouvé de corpus adéquat, nous nous sommes créé un corpus de même taille et composé de dépêches AFP sur le thème de l'Afghanistan. Ces deux corpus ont, par la suite, été annotés grâce à notre outil de détection d'entités nommées.

Dans un deuxième temps, nous avons fait le choix d'une évaluation manuelle et proposé à quatre membres du département d'évaluer les résultats de notre outil. Les évaluateurs sont des ingénieurs spécialisés dans le « media mining », ayant déjà des connaissances dans le domaine de l'extraction d'information et des entités nommées. Nous leur avons fourni une trentaine de fiches d'évaluation[70] pour chacune des langues évaluées ainsi qu'une notice contenant quelques consignes pour évaluer les annotations[71].

Chaque fiche d'évaluation est composée d'un texte annoté par notre outil ainsi que d'un tableau d'évaluation. Les annotations se présentent sous la forme de balises ouvrantes et fermantes délimitant l'entité nommée reconnue et dont le type indique celui de l'entité (ORG, LOC, PERS ou DATE). Comme indiqué dans les consignes, le tableau reprend chaque annotation du texte et

70 cf. Annexe 7
71 cf. Annexe 8

propose, pour chacune d'elles, de juger son intérêt, son type et ses limites[72]. L'évaluateur a également la possibilité d'ajouter un commentaire pour chaque annotation.

7.2 - Analyse des résultats

Avant toute analyse des résultats de l'évaluation, nous avons examiné différents modes d'évaluation présentés par [Nadeau et al., 2007] et plus particulièrement ceux des campagnes MUC, ACE et ESTER (cf. 2.2.2). Cela nous a permis, dans un premier temps, de faire le point sur les différents types d'erreurs d'annotation existants afin de construire une fiche d'évaluation adéquate. Par ailleurs, nous nous sommes inspirée des consignes fournies par ces campagnes pour proposer nos propres indications concernant plusieurs cas d'annotation complexes à juger.

Dans un deuxième temps, nous avons défini les métriques qui nous permettront d'évaluer les résultats de nos travaux. La précision, le rappel et la F-mesure ont été choisis pour leur utilisation fréquente dans le domaine du TAL et dans un souci de comparaison future avec d'autres systèmes d'extraction. Rappelons brièvement les modes de calcul de ces métriques :

$$précision = \frac{nombre\ d'entités\ correctement\ étiquetées\ E}{nombre\ d'entités\ étiquetées\ E}$$

$$rappel = \frac{nombre\ d'entités\ correctement\ étiquetées\ E}{nombre\ d'entités\ E}$$

$$F_{mesure} = \frac{2 \cdot précision \cdot rappel}{précision + rappel}$$

Figure 7.1 : Métriques d'évaluation

Afin d'analyser les évaluations faites par notre équipe, nous avons établi 5 types d'annotation et décompté pour chacun d'eux le nombre d'occurences dans chaque corpus. Ces types sont les suivants :

- ***COR*** : annotation ayant un intérêt et dont le type et les limites sont corrects ;
- ***COR Lim*** : annotation ayant un intérêt, dont le type est bon mais mal délimitée ;
- ***INC*** : annotation sans intérêt (absence d'entité nommée) ;
- ***INC Type*** : annotation indiquant bien une entité nommée et bien délimitée mais dont le type est erroné ;
- ***MANQ*** : entité nommée n'ayant pas été annotée par le système.

72 cf. Annexe 8 pour de plus amples explications

A partir de ces différents types, nous avons estimé trois autres paramètres nécessaires pour le calcul des métriques (cf. Figure 7.1) :

- ***CORRECT*** : nombre d'annotations correctes mises par le système (*nombre d'entités correctement étiquetées E* dans les formules de précision et rappel) ;
- ***ANNOTÉ*** : nombre total d'annotations mises par le système (*nombre d'entités étiquetées E* dans la formule de précision) ;
- ***RÉEL*** : nombre total d'entités nommées existantes (*nombre d'entités E* dans la formule de rappel).

Ces trois paramètres ont été calculés ainsi :

- *CORRECT = COR* **ou** *CORRECT = COR + COR Lim*[73]
- *ANNOTÉ = COR + COR Lim + INC + INC Type*
- *RÉEL = COR + COR Lim + MANQ*

Les métriques d'évaluation qui nous intéressent peuvent donc être redéfinies de la façon suivante :

$$précision = \frac{CORRECT}{ANNOTÉ}$$

$$rappel = \frac{CORRECT}{RÉEL}$$

$$F_{mesure} = \frac{2 \cdot précision \cdot rappel}{précision + rappel}$$

Figure 7.2 : Métriques d'évaluation redéfinies

Les résultats de notre extraction d'entités nommées en anglais et en français sont présentés ci-dessous (cf. tableaux 7.1 et 7.2). Les trois métriques précédentes sont calculées pour chaque type d'entité (DATE, LOC, ORG, PERS) mais également pour l'ensemble des entités nommées (EN). Les lignes grisées correspondent aux calculs où l'on considère les annotations correctes comme ayant un type et des limites exacts (i.e. *CORRECT = COR*). Les lignes laissées blanches prennent également en compte les annotations mal délimitées dans le nombre d'annotations correctes (i.e. *CORRECT = COR + COR Lim*). Nous commenterons ces chiffres dans la section suivante.

73 Une annotation correcte peut être définie soit comme une annotation dont le type et les limites sont exacts, soit comme une annotation dont, au minimum, le type est correct. Cette distinction dépend de la « sévérité » avec laquelle on juge le système d'extraction. Nous présenterons, ici, les résultats de notre système dans les deux cas de figure pour plus de transparence dans notre évaluation.

	Précision	Rappel	F-mesure
DATE	0,99	0,73	0,84
	0,99	0,73	0,84
LOC	0,94	0,91	0,92
	0,98	0,93	0,95
ORG	0,72	0,66	0,69
	0,86	0,73	0,79
PERS	0,94	0,83	0,88
	0,98	0,86	0,92
EN	**0,90**	**0,83**	**0,86**
	0,96	**0,86**	**0,91**

Tableau 7.1 : Évaluation de l'extraction d'ENs en anglais

	Précision	Rappel	F-mesure
DATE	0,98	0,69	0,81
	1,00	0,70	0,83
LOC	0,97	0,89	0,93
	0,99	0,91	0,95
ORG	0,94	0,75	0,84
	0,98	0,78	0,87
PERS	0,73	0,71	0,72
	0,96	0,94	0,95
EN	**0,91**	**0,80**	**0,85**
	0,98	**0,85**	**0,91**

Tableau 7.2 : Évaluation de l'extraction d'ENs en français

7.3 - Observations et améliorations envisagées

Nous pouvons, tout d'abord, dire, au vu de la F-mesure globale (EN), que notre système d'extraction obtient de bons résultats en anglais et en français. En effet, la majorité des outils d'extraction d'entités nommées atteignent 90% de F-mesure sur des tâches similaires. La différence d'environ 6% entre les deux types d'annotations correctes révèle, d'ores et déjà, des améliorations nécessaires dans la portée des annotations.

Analysons maintenant ces métriques du point de vue de chaque type d'entité nommée. Premièrement, la reconnaissance des dates s'avère presque parfaite au niveau de la précision et cela s'explique par le fait qu'il n'existe généralement pas d'ambiguïté entre le type « date » et les autres types d'entité. Toutefois, le rappel demande à être amélioré, c'est-à-dire que notre outil n'a pas détecté toutes les dates présentes dans les corpus anglais et français (cf. Figure 7.3). Cela est dû à l'absence de certains mots déclencheurs de date dans les *gazetteers* anglais et français. Ce problème pourra être facilement corrigé pour augmenter de façon significative le rappel. Cependant, une limite importante est l'ambiguïté fréquente entre une date n'étant pas repérable par sa structure ou

par un mot déclencheur et un nombre quelconque à quatre chiffres. Il s'agira dans ce cas de trouver le juste équilibre entre bruit et silence. Par ailleurs, une meilleure extraction devra prendre en compte les dates relatives et permettre leur normalisation par la résolution de référence.

- Kopp has been arrested in several states since **1990** for protesting abortion.
- En hausse de 14% par rapport à **2008**, ce bilan est le plus lourd depuis le déclenchement de la guerre

Figure 7.3 : Exemples de dates non-détectées

Notre outil obtient de bons résultats dans la détection des noms de lieux dans les deux langues. On observe, tout de même, une légère différence selon la définition du paramètre *CORRECT*. Il nous faut également rappeler que nous n'avons pas considéré certaines ambiguïtés entre LOC et ORG et que, de ce fait, certaines entités annotées en tant que « Location » pourraient être aussi annotées en tant que « Organization » (cf. Figure 7.4). Deux solutions se présentent pour le futur : adopter, comme le fait la campagne ACE, un type d'entité réunissant les entités géopolitiques, ou bien, prendre en compte dans les règles linguistiques le contexte particulier de ce type d'ambiguïté (verbes à sujet humain, etc.).

- **Cuba** condems bomb blasts in Kenya.
- Ces détenus appartiennent au Mouvement islamique de l'Ouzbékistan, considéré par **Washington** comme une organisation terroriste.

Figure 7.4 : Exemples de lieux ambigus

L'extraction des noms d'organisation reste la moins performante au vu de nos résultats : cela concerne toutes les métriques en anglais et seulement le rappel en français. Le manque de précision dans le cas du corpus anglais peut avoir plusieurs causes. Tout d'abord, ayant repris les *gazetteers* déjà présents dans ANNIE, nous n'avons pas vérifié l'ensemble des termes qu'ils contiennent. Or, nous avons pu observer que certaines annotations incorrectes sont dues à la présence de termes ambigus voire inappropriés dans les listes de déclencheurs d'organisation (p. ex., « Base », « Palace », « Embassy », « clinic », « house », etc.). La suppression de ces termes ou leur marquage en tant que « terme ambigu » serait une solution à ce genre d'erreur. D'autre part, nous avons remarqué que certaines de nos règles limitent trop peu la structure des noms d'organisation et donnent lieu à des annotations erronées (cf. Figure 7.5). Afin d'améliorer la précision, il sera nécessaire de corriger ce type de règles générant du bruit. Enfin, un dernier type d'annotation incorrecte peut être du au module de résolution de coréférences nommé Orthomatcher. En effet, celui-ci présente quelques limites et peut être à l'origine d'un mauvais typage d'entité. Prenons pour exemple le cas de l'entité « Clinton's administration » : celle-ci est détectée, à raison, par notre système comme une organisation. Toutefois, lorsque l'entité « Clinton » apparaitra une nouvelle fois

dans le texte, l'Orthomatcher détectera une coréférence et annotera cette dernière, non en tant que personne, mais bien comme une organisation.

- State
- South Africa, police
- Cuban
- Communist
- Canadian
- High

Figure 7.5 : Exemples d'annotations incorrectes de type "Organization"

Dans les deux langues traitées, le rappel est inférieur à celui des autres entités : de nombreuses organisations ne sont pas détectées par notre outil. Ce taux plus bas que la moyenne met en avant une complexité plus élevée dans la reconnaissance de telles entités. En effet, les noms d'organisation peuvent, en premier lieu, inclure un nom de personne ou de lieu et être, par conséquent, mal typés (*INC Type*). Par ailleurs, contrairement aux entités de type « Location », une liste exhaustive d'organisations est peu envisageable au vu du nombre important de telles entités et de leur renouvellement constant. Nous pourrions améliorer nos résultats en complétant nos listes d'organisations, mais ce n'est pas la solution idéale compte tenu du temps important nécessaire. Pour finir, le décompte des annotations mal délimitées comme des annotations correctes entraîne des écarts importants dans les métriques de ce type d'entité ; cela montre encore la complexité de telles entités et, particulièrement en anglais, la difficulté de notre outil à bien cerner leurs frontières.

La détection des entités de type « Person » présente globalement de bons résultats. Nous devons toutefois souligner, dans le traitement de textes français, une diminution des performances lorsque l'on ne considère que les annotations dont le type et les limites sont exacts. Cela traduit un bon typage des entités mais des difficultés dans leur délimitation. Après observation des erreurs, nous avons constaté que, dans de nombreux cas, le titre ou la fonction était inclu dans l'entité « Person ». Il s'agit là d'une erreur dans la construction des règles linguistiques pour le français, qui, rapidement corrigée, permettra de rééquilibrer les métriques pour ce type d'entité.

Pour conclure, précisons que le choix d'une évaluation humaine comporte une part de variabilité provenant des différences d'évaluation entre évaluateurs. En effet, malgré des consignes précises, il existe toujours des cas particulièrement complexes à juger (ambiguïtés) que les annotateurs ne perçoivent pas de façon identique. Pour un meilleur aperçu de cette variabilité, nous avons demandé à deux des annotateurs d'évaluer un même échantillon du corpus anglais (cf. Tableau 7.3 pour quelques exemples de ces différences).

Évaluateur 1	Évaluateur 2
<ORG>Defence Ministry</ORG>	Ø
<ORG>Soviet Union</ORG>	<ORG>Soviet Union</ORG> => type discutable
<LOC>Europe</LOC>	<LOC>Europe's East</LOC>
<ORG>NATO</ORG>	<ORG>member of NATO</ORG>
<DATE>1968</DATE>	Ø
Ø	<ORG>anti-communist Solidarity Forces</ORG>

Tableau 7.3 : Différences inter-annotateurs

Conclusion

Pour conclure ce mémoire, nous pouvons, tout d'abord faire le point sur le travail que nous avons réalisé tout au long de ce stage. Le tableau suivant synthétise l'ensemble de nos réalisations pour l'anglais et le français du point de vue de nos objectifs de départ.

	Anglais		**Français**	
	Réalisé	*À réaliser*	*Réalisé*	*À réaliser*
Ontologie de domaine	• Modélisation • Vocabulaire		• Modélisation	• Vocabulaire
Entités nommées	• Adaptation d'ANNIE		• Création des règles et gazetteers	
Évènements	• Méthodologie d'extraction • Création des règles et gazetteers		• Méthodologie d'extraction	• Implémentation avec analyseur en dépendance
Relations entre entités	• Méthodologie d'extraction • Création des règles et gazetteers		• Méthodologie d'extraction	• Création des règles et gazetteers

Tableau 1 : Synthèse des réalisations

Nous avons, pour le traitement des textes anglais, atteint nos objectifs et réalisé les différentes tâches prévues. Celles-ci n'ont toutefois pu être entièrement achevées pour la langue française en raison de plusieurs problèmes rencontrés. Comme nous l'avons expliqué, l'absence d'analyseur syntaxique dédié au français a constitué un obstacle majeur à l'extraction d'évènements dans cette langue. Par ailleurs, le développement du vocabulaire français pour notre ontologie de domaine ainsi que l'implémentation de notre méthode d'extraction de relations n'ont pu être effectués par manque de temps.

Notre ontologie de domaine, ayant été développée en concertation avec les membres de notre équipe, répond bien à leurs attentes et aux besoins des éventuelles applications. Cette modélisation pourrait être améliorée par une rencontre avec un « opérationnel » du renseignement militaire afin de mieux définir les classes et propriétés d'intérêt. De plus, cette ontologie devra être peuplée par les différents éléments extraits (instances) afin de créer une véritable base de connaissances dédiée au partage des informations entre les différents acteurs du domaine. L'évaluation de notre extraction d'entités nommées (précision, rappel et F-mesure) a révélé de bons résultats d'ensemble en anglais et en français. Cependant, ces mesures s'avèrent assez hétéroclites selon le type d'entité considéré : la détection des lieux et des personnes obtient les meilleurs scores ;

le rappel doit être amélioré en ce qui concerne les dates ; l'extraction des organisations s'avère être la tâche la plus complexe et montre les moins bons résultats en anglais. Cette évaluation nous a permis de percevoir plus globalement la qualité de notre extraction mais aussi de mettre en avant certaines limites qui devront être dépassées dans le futur. Les deux derniers types d'extraction (évènements et relations) n'ayant pu être réalisées en français, nous avons mis l'accent sur la mise en œuvre de méthodologies générales afin de permettre une implémentation ultérieure. Nous sommes convaincus qu'une analyse syntaxique (constituants et dépendances) est indispensable pour ces tâches d'extraction. Celles-ci nécessitent, toutefois, un analyseur relativement fiable et robuste et nous avons pu constater que de nombreux efforts sont encore à mener dans ce domaine en termes de performances et de langues traitées.

Plus généralement, si la durée du stage l'avait permis, nous aurions souhaité explorer les différentes techniques d'apprentissage en TAL. La clé de bons résultats en extraction d'information réside, à notre avis, dans la combinaison d'approches linguistiques et statistiques.

A un niveau plus personnel, ce stage au sein d'EADS et du département IPCC a été une expérience extrêmement enrichissante sur plusieurs points.

Sur le plan technique, j'ai pu, durant ces six mois, mettre en application les diverses connaissances acquises tout au long de ma formation en linguistique et informatique. Mes deux années de master TAL ont su m'apporter le savoir théorique nécessaire pour bien cerner les différentes problématiques associées à l'extraction de l'information ; notamment par l'acquisition de concepts et de techniques fondamentaux, tels que les entités nommées, les ontologies ou encore les différents niveaux d'analyse de texte (*tokenization*, analyse morphologique et syntaxique, etc.). Ce stage m'a également permis de réutiliser une partie de mes connaissances en linguistique générale, syntaxe, sémantique ou encore lexicologie. J'ai aussi réalisé qu'il a manqué, dans ma formation, un enseignement des différents outils et formalismes couramment utilisés en extraction d'information (GATE, Protégé, OWL, etc.). D'autre part, mes bases en informatique ont facilité la communication avec les ingénieurs du département ainsi que mon adaptation aux différents outils utilisés. Une expérience préalable du langage Java m'aurait beaucoup servi dans l'élaboration des règles linguistiques, mais j'ai su m'adapter et réutiliser mes connaissances dans d'autres langages pour appréhender celui-ci.

Pour finir, ce stage m'a beaucoup apporté professionnellement et humainement. Mon immersion durant 6 mois au sein d'EADS m'a permis de compléter ma vision du monde de l'entreprise, de son organisation, de ses atouts et contraintes. J'ai apprécié le fait de travailler au sein d'une équipe dont chaque membre possède son champ de compétences et où celles-ci sont mises en commun afin de réaliser un projet et répondre au besoin concret d'un client. Réaliser un outil dans son ensemble, de l'analyse de l'existant jusqu'à son évaluation, en respectant un ensemble d'objectifs et de contraintes, m'a également beaucoup plu et a su entretenir ma motivation. C'est principalement pour l'intérêt que je porte à ce sujet et le dynamisme de cette équipe que j'ai accepté la proposition de continuer l'expérience en thèse CIFRE (en attente d'acceptation). En collaboration avec le GREYC, celle-ci s'intitule « Vers une capitalisation des connaissances orientée utilisateur : extraction et structuration automatique de l'information issue de sources ouvertes » et me permettra d'approfondir les différentes problématiques abordées au cours de ce stage.

Bibliographie – Sitographie

[1] 2005. The ACE 2005 (ACE05) evaluation plan. http://www.nist.gov/speech/tests/ace/ace05/doc/ace05-evalplan.v3.pdf (consulté le 11/08/2010).

[2] Ahn, D. (2006). The stages of event extraction. In *ARTE '06: Proceedings of the Workshop on Annotating and Reasoning about Time and Events (Sydney, Australia),* pp.1-8. Morristown, USA : ACL.

[3] AKSW. University of Leipzig. OntoWiki. http://ontowiki.net/Projects/OntoWiki (consulté le 17/08/2010).

[4] Alias-i. LingPipe. http://alias-i.com/lingpipe/ (consulté le 17/08/2010).

[5] Aone, C. & Ramos-Santacruz, M. (2000). REES: a large-scale relation and event extraction system. In *Proceedings of the Sixth Conference on Applied Natural Language Processing (Seattle, Washington),* pp.76-83. Morristown, USA : ACL.

[6] Bird, S. NLTK. http://www.nltk.org/ (consulté le 17/08/2010).

[7] Brill E. (1992). A Simple Rule-Based Part of Speech Tagger. In Proceedings of the third conference on Applied natural language processing, pp. 152-155. Morristown, USA : ACL.

[8] Cadiot, P. (2002). Schémas et motifs en sémantique prépositionnelle : vers une description renouvelée des prépositions dites « spatiales », In *Travaux de linguistique, volume 1, number 44,* pp. 9-24.

[9] Cunningham, H. et al. (2002). GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*. Philadelphia.

[10] De Marneffe, M.-C. & Manning, C. D. (2008). Stanford typed dependencies manual. Technical report, Stanford University, USA.

[11] François, J., Le Pesant, D. & Leeman, D. (2007). Présentation de la classification des Verbes Français de Jean Dubois et Françoise Dubois-Charlier. In J. François, D. Le Pesant & D. Leeman (dir.), *Le classement syntactico-sémantique des verbes français*, p.153. Paris : Larousse.

[12] Gruber, T. (1993). A translation approach to portable ontology specifications. In *Knowledge Acquisition, volume 5, number 2,* pp. 199–220. London : Academic Press Ltd.

[13] Ireson, N. & Ciravegna, F. (2005). Pascal Challenge The Evaluation of Machine Learning for Information Extraction. In *Proceedings of Dagstuhl Seminar Machine Learning for the Semantic Web.*

[14] Maynard, D., Bontcheva, K. & Cunningham, H.(2003). Towards a semantic extraction of named entities. In *Recent Advances in Natural Language Processing.* Bulgaria.

[15] McCallum, A. (2005). Information Extraction: Distilling Structured Data from Unstructured Text. In *Queue, volume 3, number 9*, 48-57. New York : ACM.

[16] McCallum, A. MALLET. http://mallet.cs.umass.edu/ (consulté le 17/08/2010).

[17] McDonald, David D. (1996). Internal and external evidence in the identification and semantic categorization of proper names. In *Corpus Processing For Lexical Acquisition*, pp.21-39. Cambridge, USA : MIT Press.

[18] Mikheev, A., Moens, M., & Grover, C. (1999). Named Entity recognition without gazetteers. In *Proceedings of the Ninth Conference on European Chapter of the Association For Computational Linguistics*. Morristown, USA : ACL.

[19] Minard, A.-L. (2008). *Recherche et analyse de ressources terminologiques liées à la topographie*. Rapport de stage de Master 1 « Traitement automatique des langues », Université Lille 3, Lille, France.

[20] Mizoguchi, R. (2004). Tutorial on ontological engineering - Part 2: Ontology development, tools and languages. In *New Generation Computing, Vol.22, No.1*, pp.61-96, OhmSha&Springer.

[21] Mizoguchi, R. (2003). Tutorial on ontological engineering - Part 1: Introduction to Ontological Engineering. In *New Generation Computing, Vol.21, No.4*, pp.365-384, OhmSha&Springer.

[22] Nadeau, D. & Sekine, S. (2007). A survey of named entity recognition and classification. In John Benjamins Publishing Company (Pub.), *Journal of Linguisticae Investigationes, vol.30, number 1*, pp.3-26.

[23] Nakamura-Delloye, Y. (2010). Extraction des chemins entre deux entités nommées en vue de l'acquisition des patrons de relations. In *IC 2010*, Nîmes, France.

[24] NATO. STANAGs STANdard Agreements. http://www.nato.int/cps/en/natolive/stanag.htm (consulté le 17/08/2010).

[25] Neches, R. et al. (1991). Enabling technology for knowledge sharing. In *AI Magazine, volume 12, number 3*, pp. 36-56. Menlo Park, USA : American Association for Artificial Intelligence.

[26] Noy, N. F. & McGuinness, D. L. (2001). Ontology Development 101 : a guide to creating your first ontology. In *Stanford Knowledge Systems Laboratory,* Technical Report. Stanford, USA : Stanford Medical Informatics.

[27] Ontotext. http://www.ontotext.com/ (consulté le 17/08/2010).

[28] OpenNLP. http://opennlp.sourceforge.net/ (consulté le 17/08/2010).

[29] Poibeau, T. (2003). Extraction automatique d'information. Du texte brut au web sémantique. Paris : Hermès.

[30] Pustejovsky, J. et al. (2003). TimeML: Robust Specification of Event and Temporal Expressions in Text. In M. T. Maybury (Ed.), *New Directions in Question Answering*, pp.28-34. AAAI Press.

[31] Sauri, R. et al. (2005). Evita: A Robust Event Recognizer For QA Systems. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing,* pp.700-707. Morristown, USA : ACL.

[32] Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing.* Manchester, UK.

[33] SEKT. PROTON PROTo ONtology. http://proton.semanticweb.org/ (consulté le 17/08/2010).

[34] Smart, P. R., Russell, A. & Shadbolt, N. R. (2007). AKTiveSA: Supporting Civil-Military Information Integration in Military Operations Other than War. In *4th International Conference on Knowledge Systems for Coalition Operations (KSCO).* Waltham, USA.

[35] Stanford Center for Biomedical Informatics Research. Protégé. http://protege.stanford.edu/ (consulté le 17/08/2010).

[36] Stanford NLP Group. http://nlp.stanford.edu/ (consulté le 17/08/2010).

[37] SUMO. http://www.ontologyportal.org/ (consulté le 17/08/2010).

[38] Thomson Reuters. OpenCalais. http://www.opencalais.com/ (consulté le 17/08/2010).

[39] W3C. OWL Web Ontology Language. http://www.w3.org/TR/owl-features/ (consulté le 17/08/2010).

[40] W3C. RDF Resource Description Framework. http://www.w3.org/RDF/ (consulté le 17/08/2010).

Table des figures

Figure 1.1 : Divisions du groupe EADS....10
Figure 1.2 : Organisation de la division Defence & Security....11
Figure 1.3 : La plateforme WebLab....13
Figure 2.1 : Métriques d'évaluation....17
Figure 2.2 : Exemple d'annotation dans GATE....20
Figure 2.3 : Exemple de corpus dans GATE....20
Figure 2.4 : Exemple de texte annoté dans GATE....21
Figure 2.5 : Exemple de chaîne de traitement GATE....22
Figure 3.1 : Le pentagramme du renseignement....31
Figure 4.1 : Adaptation d'une règle d'ANNIE....38
Figure 4.2 : Organisation des phases d'extraction d'ENs pour l'anglais....39
Figure 4.3 : Extraction d'ENs en anglais : ANNIE....40
Figure 4.4 : Extraction d'ENs en anglais : ANNIE modifiée....40
Figure 4.5 : Organisation des phases d'extraction pour le français....42
Figure 4.6 : Extraction d'ENs en français....42
Figure 5.1 : Exemple d'analyse syntaxique....46
Figure 5.2 : Gazetteer associé à la classe « DamageEvent »....49
Figure 5.3 : Chaine d'extraction d'évènements pour l'anglais....50
Figure 5.4 : Organisation des phases d'extraction d'évènements pour l'anglais....51
Figure 5.5 : Extraction d'évènements en anglais....52
Figure 5.6 : Exemple d'une annotation de type "Event"....53
Figure 6.1 : Chaine d'extraction de relations pour l'anglais....57
Figure 6.2 : Exemple d'annotation d'une relation en anglais....58
Figure 7.1 : Métriques d'évaluation....61
Figure 7.2 : Métriques d'évaluation redéfinies....62
Figure 7.3 : Exemples de dates non-détectées....64
Figure 7.4 : Exemples de lieux ambigus....64
Figure 7.5 : Exemples d'annotations incorrectes de type "Organization"....65

Table des tableaux

Tableau 4.1 : Performances d'ANNIE....36
Tableau 4.2 : Normalisation des dates....37
Tableau 5.1 : Classes argumentales....48
Tableau 7.1 : Évaluation de l'extraction d'ENs en anglais....63
Tableau 7.2 : Évaluation de l'extraction d'ENs en français....63
Tableau 7.3 : Différences inter-annotateurs....66
Tableau 1 : Synthèse des réalisations....68

Glossaire

Analyse morphologique

Étude de la formation des mots (flexion, dérivation, composition) et de leur décomposition en morphèmes (racine et affixes). Une analyse morphologique indique également la catégorie grammaticale du mot, son lemme, etc.

Analyse syntaxique

Étude des relations entre les mots au sein de la phrase. L'analyse syntaxique traite de l'ordre des mots dans la phrase, des phénomènes de rection entre ces unités mais également des fonctions grammaticales que les mots peuvent remplir.

Apprentissage supervisé

Méthode d'apprentissage automatique où l'on utilise un échantillon de données dont la classe est connue au préalable afin d'apprendre des règles permettant de classer de nouvelles données.

Contexte extra-linguistique

Désigne la situation (en dehors de la langue) dans laquelle un énoncé a été prononcé ou écrit. Les élements de ce contexte (entités extra-linguistiques) permettent de résoudre les phénomènes de référence et d'anaphore.

Identité référentielle

Désigne le cas où deux expressions ont le même référent, désignent le même objet du monde (« Nicolas Sarkozy » et « le Président de la République Française »).

Lemmatisation

Associer à un mot (forme fléchie) son lemme, c'est-à-dire la forme canonique, non-fléchie qui lui correspond. Par exemple, « soufflaient » a pour lemme « souffler » et « intelligentes » a pour lemme « intelligent ».

Recteur

Qui dirige, qui est à la tête de.

Table des abréviations

ACE : Automatic Content Extraction
ACL : Association of Computational Linguistics
AFP : Agence France-Presse
AG : Aktiengesellschaft
AKSW : Agile Knowledge engineering and Semantic Web
ALPAGE : Analyse Linguistique Profonde A Grande Échelle
ANNIE : A Nearly-New Information Extraction
ANR : Agence Nationale pour la Recherche
AP : Adjectival Phrase
AQUAINT : Advanced QUestion Answering for INTelligence
BFO : Basic Formal Ontology
CASA : Construcciones Aeronáuticas Sociedad Anónima
CEA-List : Commissariat à l'Énergie Atomique – Laboratoire d'Intégration des Systèmes et des Technologies
CIFRE : Conventions Industrielles de Formation par la REcherche
CNRTL : Centre National de Ressources Textuelles et Lexicales
CoNLL : Conference on Computational Natural Language Learning
COSMO : Common Semantic MOdel
COTS : Commercial Off-The-Shelf
CPSL : Common Pattern Specification Language
CRF : Conditional Random Fields
CRISCO : Centre de Recherches Inter-langues sur la Signification en COntexte
DAML : Darpa Agent Markup Language
DARPA : Defence Advanced Research Projects Agency
DASA : Daimler Chrysler Aerospace AG
DIF DTC : Data and Information Fusion Defence Technology Centre
DOGMA : Developing Ontology-Grounded Methods and Applications
DOLCE : Descriptive Ontology for Linguistic and Cognitive Engineering
DS : Defence & Security
DUC : Document Understanding Conference
EADS : European Aeronautic Defence and Space Company
ECTS : European Credit Transfer and Accumulation System
ELRA : European Language Resources Association
ERSS : Équipe de Recherche en Syntaxe et Sémantique
ESTER : Évaluation des Systèmes de Transcription Enrichie d'Émissions Radiophoniques
Evita : Events In Texts Analizer
GATE : General Architecture for Text Engineering
GMBH : Gesellschaft mit beschränkter Haftung
GREYC : Groupe de Recherche en Informatique, Image, Automatique et Instrumentation de Caen
HMM : Hidden Markov Model
IA : Intelligence Artificielle
IEEE-SA : Institute of Electrical and Electronics Engineers Standards Association
INRIA : Institut National de Recherche en Informatique et en Automatique
IPCC : Information Processing Control & Cognition
ISAF : International Security Assistance Force

ISLanD : Interaction, Sémiotique : LANgue, Diagrammes
ISO : International Organization of Standardization
JAPE : Java Annotation Patterns Engine
KIF : Knowledge Interchange Format
KIM : Knowledge and Information Management
KM : Knowledge Machine
LASAF : Logiciel d'Analyse Syntaxique Automatique du Français
LFG : Lexical Functional Grammar
LHS : Left-Hand Side
LIP6 : Laboratoire d'Informatique de Paris 6
LITIS : Laboratoire d'Informatique, du Traitement de l'Information et des Systèmes
LR : Language Resource
LREC : Language Resources Evaluation Conference
MALLET : Machine Learning for LanguagE Toolkit
Mindswap : Maryland Information and Network Dynamics Lab Semantic Web Agents Project
MBDA : Matra Bae Dynamics Aérospatiale
ML : Machine Learning
MRT : Machine Reading Program
MUC : Message Understanding Conference
N3 : Notation 3
NE : Named Entity
NIST : National Institute of Standards and Technology
NLP : Natural Language Processing
NLTK : Natural Language ToolKit
NP : Noun Phrase
OIL : Ontology Inference Layer
OSINT : Open Source Intelligence
OTAN : Organisation du Traité de l'Atlantique Nord
OWL : Web Ontology Language
POS : Part-Of-Speech
PP : Prepositional Phrase
PR : Processing Resource
PROTON : PROTo ONtology
R&D : Recherche & Développement
R&T : Recherche & Technologie
RDF : Resource Description Framework
RDFS : Resource Description Framework Schema
REES : Relation and Event Extraction System
RHS : Right-Hand Side
SA : Syntagme Adjectival
SCDE : Studies Concept Development & Experimentation
SDC : System Design Center
SEKT : Semantically Enabled Knowledge Technology
SEM : Simple Event Model
SHOE : Simple HTML Ontology Extensions
SN : Syntagme Nominal
SOA : Service Oriented Architecture
SP : Syntagme prépositionnel
STANAG : STANdard AGreement
SUMO : Suggested Upper Merged Ontology
SV : Syntagme Verbal

TAC : Text Analysis Conference
TAL : Traitement Automatique des Langues / du Langage
TALN : Traitement Automatique du Langage Naturel
TARSQI : Temporal Awareness and Reasoning Systems for Question Interpretation
VP : Verb Phrase
XIP : Xerox Incremental Parser
XML : eXtensible Markup Language

Table des annexes

Annexe 1
Exemple d'ontologie au format OWL.....85
Annexe 2
Ontologie du renseignement militaire : hiérarchie des classes.....88
Annexe 3
Ontologie du renseignement militaire : Data Properties.....90
Annexe 4
Ontologie du renseignement militaire : Object Properties.....91
Annexe 5
Liste de lemmes associés à la relation « isMemberOf ».....92
Annexe 6
Exemple d'une règle JAPE pour l'extraction d'une relation
« Personne-Organisation ».....93
Annexe 7
Exemple de fiche d'évaluation pour l'anglais.....95
Annexe 8
Consignes pour l'évaluation en anglais.....96
Annexe 9
Exemple d'analyse en dépendance par Syntex.....98
Annexe 10
Exemple d'analyse en dépendance par XIP.....99

Annexe 1
Exemple d'ontologie au format OWL

```
<?xml version="1.0"?>
<!DOCTYPE rdf:RDF [
    <!ENTITY owl "http://www.w3.org/2002/07/owl#" >
    <!ENTITY swrl "http://www.w3.org/2003/11/swrl#" >
    <!ENTITY swrlb "http://www.w3.org/2003/11/swrlb#" >
    <!ENTITY xsd "http://www.w3.org/2001/XMLSchema#" >
    <!ENTITY owl2xml "http://www.w3.org/2006/12/owl2-xml#" >
    <!ENTITY rdfs "http://www.w3.org/2000/01/rdf-schema#" >
    <!ENTITY rdf "http://www.w3.org/1999/02/22-rdf-syntax-ns#" >
    <!ENTITY protege "http://protege.stanford.edu/plugins/owl/protege#" >
    <!ENTITY xsp "http://www.owl-ontologies.com/2005/08/07/xsp.owl#" >
    <!ENTITY OntoRens
"http://www.semanticweb.org/ontologies/2010/4/OntoRens.owl#" >
]>
<rdf:RDF xmlns="http://www.semanticweb.org/ontologies/2010/4/OntoRens.owl#"
     xml:base="http://www.semanticweb.org/ontologies/2010/4/OntoRens.owl"
     xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
     xmlns:swrl="http://www.w3.org/2003/11/swrl#"
     xmlns:protege="http://protege.stanford.edu/plugins/owl/protege#"
     xmlns:owl2xml="http://www.w3.org/2006/12/owl2-xml#"
     xmlns:xsp="http://www.owl-ontologies.com/2005/08/07/xsp.owl#"
     xmlns:owl="http://www.w3.org/2002/07/owl#"
     xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
     xmlns:OntoRens="http://www.semanticweb.org/ontologies/2010/4/OntoRens.owl#
"
     xmlns:swrlb="http://www.w3.org/2003/11/swrlb#"
     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
    <owl:Ontology rdf:about=""/>

    <!--
    ///////////////////////////////////////////////////////////////////////////
    //
    // Object Properties
    //
    ///////////////////////////////////////////////////////////////////////////
     -->

    <!-- http://www.semanticweb.org/ontologies/2010/4/OntoRens.owl#affects -->
    <owl:ObjectProperty rdf:about="#affects">
        <rdfs:range rdf:resource="#Person"/>
        <rdfs:range rdf:resource="#Place"/>
        <rdfs:range rdf:resource="#Unit"/>
        <rdfs:subPropertyOf rdf:resource="#eventProperties"/>
    </owl:ObjectProperty>
    http://www.nato.int/cps/en/natolive/stanag.htm
    <!-- http://www.semanticweb.org/ontologies/2010/4/OntoRens.owl#bornIn -->
    <owl:ObjectProperty rdf:about="#bornIn">
        <rdfs:range rdf:resource="#Place"/>
        <rdfs:subPropertyOf rdf:resource="#personProperties"/>
    </owl:ObjectProperty>
```

```
    <!--
    ///////////////////////////////////////////////////////////////////////
    //
    // Data properties
    //
    ///////////////////////////////////////////////////////////////////////
     -->

    <!-- http://www.semanticweb.org/ontologies/2010/4/OntoRens.owl#address -->
    <owl:DatatypeProperty rdf:about="#address">
        <rdfs:subPropertyOf rdf:resource="#personDProperties"/>
        <rdfs:range rdf:resource="&xsd;string"/>
    </owl:DatatypeProperty>

    <!-- http://www.semanticweb.org/ontologies/2010/4/OntoRens.owl#age -->
    <owl:DatatypeProperty rdf:about="#age">
        <rdfs:subPropertyOf rdf:resource="#personDProperties"/>
        <rdfs:range rdf:resource="&xsd;integer"/>
    </owl:DatatypeProperty>

    <!--
    ///////////////////////////////////////////////////////////////////////
    //
    // Classes
    //
    ///////////////////////////////////////////////////////////////////////
     -->

    <!--
http://www.semanticweb.org/ontologies/2010/4/OntoRens.owl#AdministrativePlace
-->
    <owl:Class rdf:about="#AdministrativePlace">
        <rdfs:subClassOf rdf:resource="#Place"/>
    </owl:Class>

    <!-- http://www.semanticweb.org/ontologies/2010/4/OntoRens.owl#AirVehicle
-->
    <owl:Class rdf:about="#AirVehicle">
        <rdfs:subClassOf rdf:resource="#Vehicle"/>
    </owl:Class>

    <!--
http://www.semanticweb.org/ontologies/2010/4/OntoRens.owl#ArrestOperation -->
    <owl:Class rdf:about="#ArrestOperation">
        <rdfs:subClassOf rdf:resource="#MilitaryOperation"/>
    </owl:Class>

    <!-- http://www.semanticweb.org/ontologies/2010/4/OntoRens.owl#AttackEvent
-->
    <owl:Class rdf:about="#AttackEvent">
        <rdfs:subClassOf rdf:resource="#MilitaryEvent"/>
```

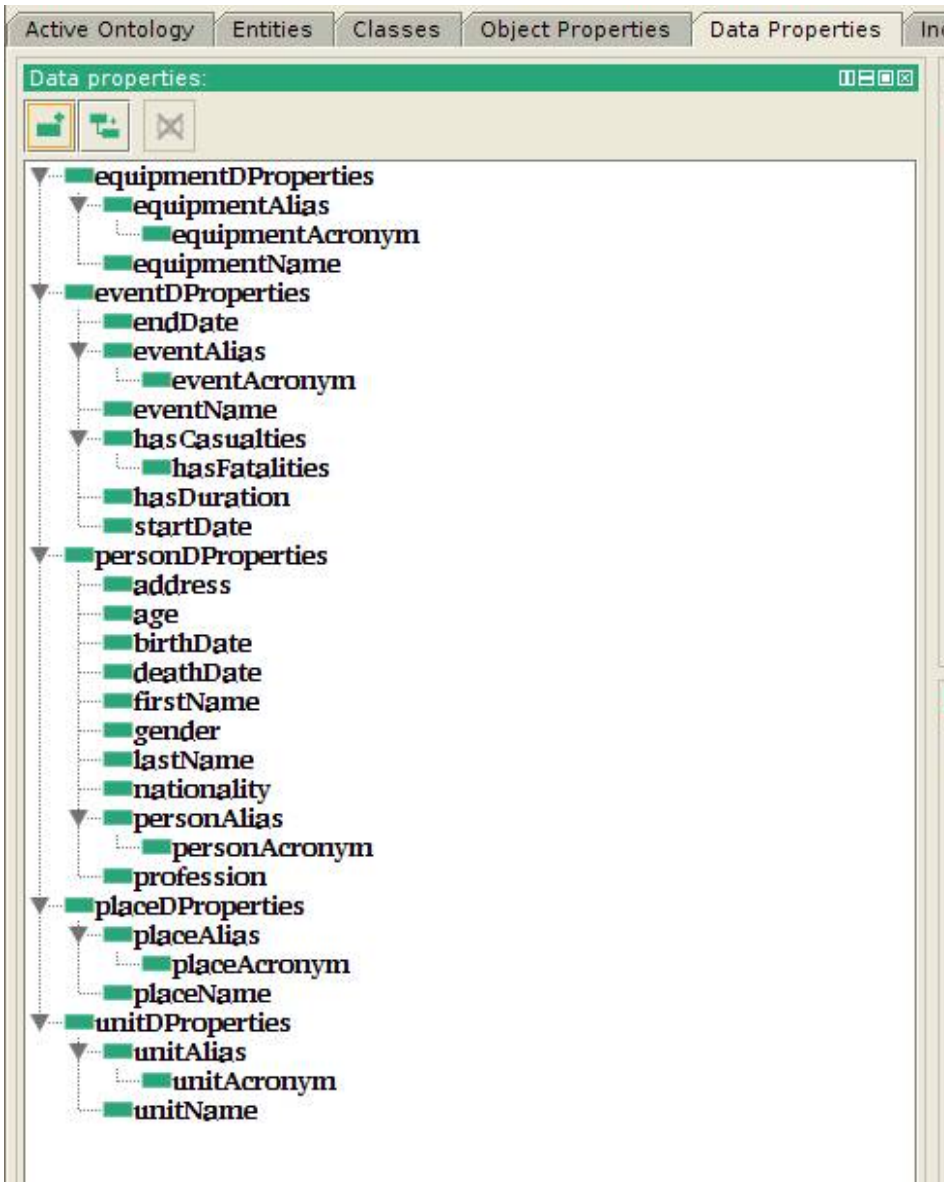
```
    </owl:Class>
</rdf:RDF>
<!-- Generated by the OWL API (version 2.2.1.1138)
http://owlapi.sourceforge.net -->
```

Annexe 2
Ontologie du renseignement militaire : hiérarchie des classes

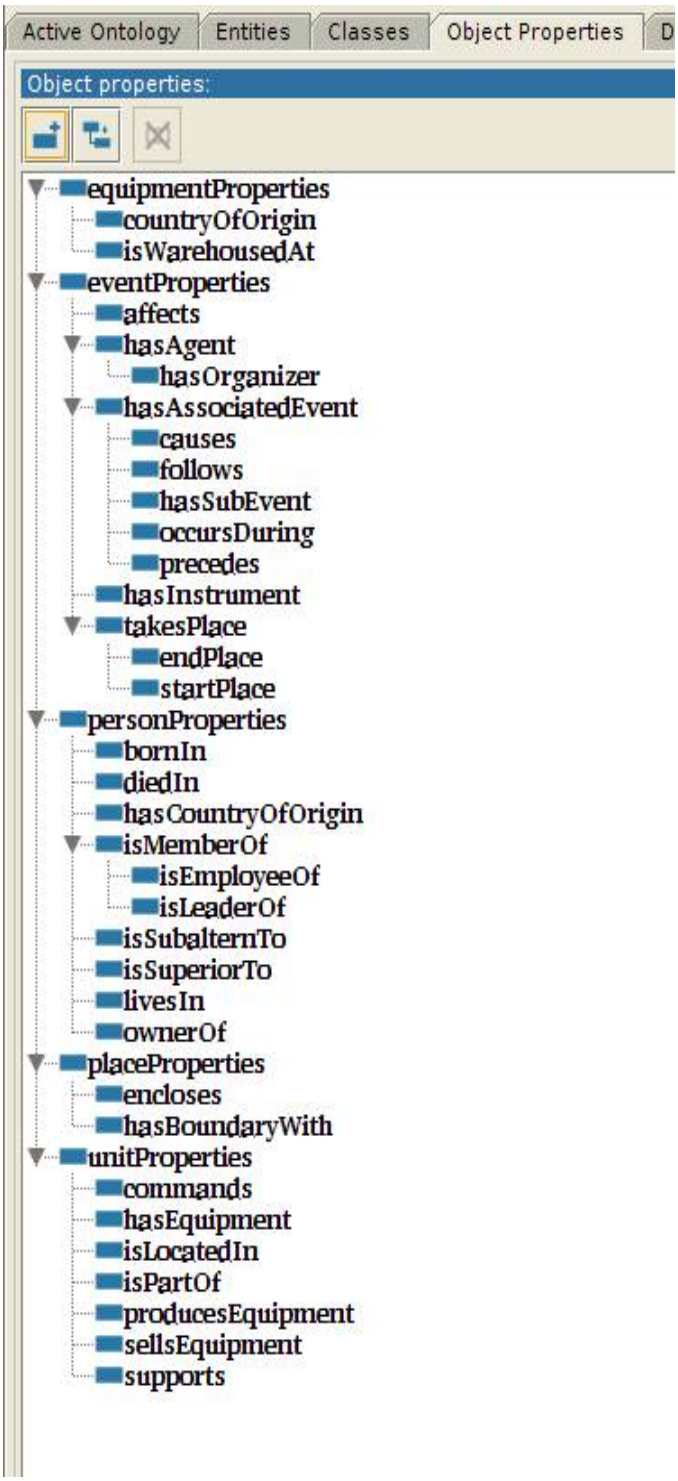
```
Thing
    Equipment
        Vehicle
            AirVehicle
            LandVehicle
            WaterVehicle
        Weapon
            CommonWeapon
            MassDestructionWeapon
            SmallWeapon
    Event
        EconomicEvent
        HumanitarianEvent
        MilitaryEvent
            AttackEvent
                BombingEvent
                ShootingEvent
            CrashEvent
            DamageEvent
            DeathEvent
            FightingEvent
            InjureEvent
            KidnappingEvent
            MilitaryOperation
                ArrestOperation
                HelpOperation
                PeaceKeepingOperation
                SearchOperation
                SurveillanceOperation
                TrainingOperation
                TroopMovementOperation
            NuclearEvent
            TrafficEvent
        NaturalEvent
        PersonalEvent
        PoliticalEvent
        SecurityEvent
    Person
    Place
        AdministrativePlace
            City
            Continent
            Country
        Infrastructure
            Building
            CommunicationInfrastructure
            PowerInstallation
        NaturalPlace
    Unit
        Group
```

```
            BeliefGroup
                 ReligiousGroup
                 Sect
            CriminalGroup
                 Gang
                 TerroristGroup
            EthnicGroup
       Organization
            CommercialOrg
            MilitaryOrg
            NonGovOrg
            ParamilitaryOrg
            PoliticalOrg
            SecurityOrg
```

Annexe 3
Ontologie du renseignement militaire : Data Properties



Annexe 4
Ontologie du renseignement militaire : Object Properties



Annexe 5
Liste de lemmes associés à la relation « isMemberOf »

```
brigadier
Capt.
Captain
captain
col
Col.
colonel
commander
commissioner
Flight Lieutenant
Flt Lt
Gen.
general
governor
inspector
Lieutenant Colonel
Lt.
Lt. Col.
maj.
Maj.
major
member
professor
Secretary General
Secretary-General
sergeant
spokesman
spokesperson
Tech. Sgt.
Warrant Officer
```

Annexe 6
Exemple d'une règle JAPE pour l'extraction d'une relation « Personne-Organisation »

```
Rule: PersOrgRel
(
  ({Person}):arg1
  (COMMA)?
  (DET)?
  ({Organization}):arg2
  (REL):relType
):rel
-->
{
  gate.AnnotationSet relTypeSet = (gate.AnnotationSet) bindings.get("relType");
  gate.Annotation relTypeAnn = relTypeSet.iterator().next();

  gate.AnnotationSet relSet = (gate.AnnotationSet) bindings.get("rel");
  gate.Annotation relAnn = relSet.iterator().next();

  gate.AnnotationSet arg1Set = (gate.AnnotationSet) bindings.get("arg1");
  gate.Annotation arg1Ann = arg1Set.iterator().next();

  gate.AnnotationSet arg2Set = (gate.AnnotationSet) bindings.get("arg2");
  gate.Annotation arg2Ann = arg2Set.iterator().next();

  String annot = "";
  String relTxt = "";

  gate.AnnotationSet lookupSet = (gate.AnnotationSet) inputAS.get("Lookup",
relTypeAnn.getStartNode().getOffset(), relTypeAnn.getEndNode().getOffset());
  if(lookupSet.size() == 1) {
    gate.Annotation lookupAnn = lookupSet.iterator().next();
    annot = (String) lookupAnn.getFeatures().get("minorType");
    try {
      relTxt =
doc.getContent().getContent(relTypeAnn.getStartNode().getOffset(),
relTypeAnn.getEndNode().getOffset()).toString();
    } catch(InvalidOffsetException ioe) {
      ioe.printStackTrace();
    }
  }else if(lookupSet.size() > 1) {
    Iterator it = lookupSet.iterator();
    while(it.hasNext()) {
      gate.Annotation lookupAnn = (gate.Annotation) it.next();
      if(lookupAnn.getFeatures().get("majorType").equals("relation")) {
        annot = (String) lookupAnn.getFeatures().get("minorType");
        try {
          relTxt =
doc.getContent().getContent(relTypeAnn.getStartNode().getOffset(),
relTypeAnn.getEndNode().getOffset()).toString();
```

```
          } catch(InvalidOffsetException ioe) {
            ioe.printStackTrace();
          }
        }
      }
    }

    String classOnto = annot;

    gate.FeatureMap features = Factory.newFeatureMap();
    features.put("class", classOnto);
    features.put("ontology",
"http://www.semanticweb.org/ontologies/2010/4/OntoRens.owl");
    features.put("arg1", arg1Ann.getId());
    features.put("arg2", arg2Ann.getId());
    features.put("link", relTxt);
    features.put("rule", "PersOrgRel");

    outputAS.add(relSet.firstNode(), relSet.lastNode(), annot, features);
}
```

Annexe 7
Exemple de fiche d'évaluation pour l'anglais

APW19990122.0193
<DATE>1999-01-22</DATE> 13:06:18
usa

DNA Tests Link Slain Doctor <PERS>Suspect</PERS>

<LOC>WASHINGTON</LOC> (<ORG>AP</ORG>) -- Preliminary DNA tests link a missing anti-abortion activist to a strand of hair found near where a sniper shot and killed a <LOC>Buffalo</LOC>, <LOC>N.Y.</LOC>, doctor who performed abortions, a law enforcement official said Friday. The first round of DNA tests on the hair at the <ORG>FBI Laboratory</ORG> here established a high probability it came from the same person as a hair found in a <LOC>New Jersey</LOC> home where <PERS>James C. Kopp</PERS>, a 44-year-old anti-abortion protester, lived last year, the official said. The first DNA tests did not exclude a match between the two strands. <ORG>Kopp</ORG> has eluded authorities since they obtained a warrant for him as a material witness in the Oct. 23 sniper shooting of Dr. Barnett <PERS>Slepian</PERS>, a 52-year-old obstetrician-gynecologist who performed abortions. The search for <ORG>Kopp</ORG> was recently extended to <LOC>Mexico</LOC>. Meantime, <ORG>FBI</ORG> agents and <ORG>Metropolitan Police</ORG> officers assigned to a joint terrorism task force here scanned the crowd of anti-abortion protesters at the annual March for Life on Capitol Hill, because <ORG>Kopp</ORG> has been either a participant in or arrested at this march in each of the last three years, according to another law enforcement official, Both officials requested anonymity. The <ORG>FBI</ORG> is conducting further DNA tests of the hair found outside Dr. <PERS>Slepian</PERS>'s home.

Annotation	Intérêt	Type	Limites	Commentaire
<DATE>1999-01-22				
<PERS>Suspect				
<LOC>WASHINGTON				
<ORG>AP				
<LOC>Buffalo				
<LOC>N.Y.				
<ORG>FBI Laboratory				
<LOC>New Jersey				
<PERS>James C. Kopp				
<ORG>Kopp				
<PERS>Slepian				
<ORG>Kopp				
<LOC>Mexico				
<ORG>FBI				
<ORG>Metropolitan Police				
<ORG>Kopp				
<ORG>FBI				
<PERS>Slepian				

Annexe 8
Consignes pour l'évaluation en anglais

Chaque fiche d'évaluation est divisée en 2 parties :

1. le texte annoté
2. un tableau d'évaluation

1. Quatre types d'entités sont annotées dans le texte : Date, Organization, Location et Person. Les annotations se présentent sous la forme de balises, dont le type et la couleur correspondent au type d'entité.

- Date → <DATE></DATE>
- Organization → <ORG></ORG>
- Location → <LOC></LOC>
- Person → <PERS></PERS>

2. Le tableau d'évaluation présente en ligne chaque annotation du texte. Pour chacune d'entre elles, il vous faut évaluer trois critères :

- Intérêt : ce critère indique si l'annotation présente un intérêt, à savoir si l'élément annoté est bien une entité nommée.
- Type : cette colonne permet d'évaluer si l'entité annotée correspond au type indiqué (Person, Date, Organization, Location).
- Limites : ce critère évalue la portée de l'annotation, c'est-à-dire si elle recouvre de façon correcte l'entité nommée.

Pour chacun de ces critères, il vous est demandé d'indiquer par un signe quelconque une évaluation négative. Lorsque le « type » de l'annotation est erroné, il vous sera demandé d'indiquer le bon « type » dans la case correspondante. Dans la colonne « Limites », vous inscrirez un signe « + » pour une annotation trop longue, un signe « - » pour une annotation trop courte et vous laisserez la case vide lorsque la portée de l'annotation est correcte.

Par ailleurs, lorsque vous repérez dans le texte des entités nommées n'ayant pas été annotées, surlignez-les ou entourez-les dans le texte (en respectant si possible le code couleur) et reportez-les dans les lignes vides du tableau prévues à cet effet. Vous indiquerez également dans la colonne « type » le type de la nouvelle annotation.

Enfin, une dernière colonne « commentaire » est à votre disposition pour d'éventuelles remarques sur votre évaluation.

Cas particuliers :

- Dans le cas d'une double annotation de même type (par exemple, <PERS>Fidel <PERS>Castro</PERS></PERS>), seule l'annotation la plus englobante sera évaluée. Dans l'exemple précédent, l'annotation sera donc considérée comme ayant un intérêt, un type correct et des limites exactes.
- Les noms de pays annotés comme « Location » et référant, dans le contexte de la phrase, au gouvernement de ce pays seront considérés comme bien annotés. On ajoutera toutefois une autre annotation de type « Organization ». Ex : « <LOC>Cuba</LOC> fears [...] ».
- « U.S. » annoté comme « Location » et référant en réalité à une nationalité (« U.S. Secretary of State »), sera jugé comme une annotation sans intérêt.
- Seules les lieux à référence absolue seront considérés comme tels : « Slepian's home » sera ignoré.
- Lorsqu'un lieu vient en préciser un autre, comme dans « Amherst, N.Y. », l'on annote chacun des lieux séparément : « <LOC>Amherst</LOC>, <LOC>N.Y.</LOC> ». Toutefois, cela diffère pour le cas suivant : « Newark, N.J. airport ». Ici, nous avons 3 annotations de type « Location » : « <LOC><LOC>Newark</LOC>, <LOC>N.J.</LOC> airport</LOC> ».
- Les entités de type « Person » n'englobent pas les termes de civilité ou de fonction : l'annotation suivante est donc correctement délimitée : « Dr.<PERS>Barnett Slepian</PERS> ».
- La mention d'un prénom seul est considéré comme une entité de type « Person ».
- Seules les dates absolues sont annotées, c'est-à-dire les dates permettant à elles-seules de situer le moment sur un axe temporel :
  - année
  - mois + année
  - jour + mois + année
- Les heures et fuseaux horaires ne sont pas compris dans les annotations de type « Date ».
- Les annotations de type « Organization » doivent englober les mentions juridiques telles que « SA », « Ltd », etc.

Annexe 9
Exemple d'analyse en dépendance par Syntex

```
<SEQ id="T_8">
    <TXT>La Norvège maintient environ 500 hommes dans le pays , principalement à Kaboul
et dans le nord .</TXT>
    <tokens>
      <t i="1" l="le" f="La" c="Det??" p="D"/>
      <t i="2" l="Norvège" f="Norvège" c="NomPrXXInc" p="NP"/>
      <t i="3" l="maintenir" f="maintient" c="VCONJS" p="V"/>
      <t i="4" l="environ" f="environ" c="Prep" p="O"/>
      <t i="5" l="500" f="500" c="DetNum" p="D"/>
      <t i="6" l="homme" f="hommes" c="Nom?P" p="N"/>
      <t i="7" l="dans" f="dans" c="Prep" p="O"/>
      <t i="8" l="le" f="le" c="Det??" p="D"/>
      <t i="9" l="pays" f="pays" c="NomMS" p="N"/>
      <t i="10" l="," f="," c="TypoCoordPrep" p="T"/>
      <t i="11" l="principalement" f="principalement" c="Adv" p="R"/>
      <t i="12" l="à" f="à" c="Prep" p="O"/>
      <t i="13" l="Kaboul" f="Kaboul" c="NomPrXXInc" p="NP"/>
      <t i="14" l="et" f="et" c="CCoordPrep" p="Cc"/>
      <t i="15" l="dans" f="dans" c="Prep" p="O"/>
      <t i="16" l="le" f="le" c="Det??" p="D"/>
      <t i="17" l="nord" f="nord" c="NomMS" p="N"/>
      <t i="18" l="." f="." c="Typo" p="T"/>
    </tokens>
    <dependances>
      <d r="DET" s="1" c="2"/>
      <g r="DET" s="2" c="1"/>
      <d r="SUJ" s="2" c="3"/>
      <g r="SUJ" s="3" c="2"/>
      <g r="PREP" s="3" c="4"/>
      <g r="PREP" s="3" c="14"/>
      <g r="NOMPREP" s="4" c="6"/>
      <d r="PREP" s="4" c="3"/>
      <d r="DET" s="5" c="6"/>
      <g r="DET" s="6" c="5"/>
      <d r="NOMPREP" s="6" c="4"/>
      <g r="NOMPREP" s="7" c="9"/>
      <d r="CC" s="7" c="14"/>
      <d r="DET" s="8" c="9"/>
      <g r="DET" s="9" c="8"/>
      <d r="NOMPREP" s="9" c="7"/>
      <g r="NOMPREP" s="12" c="13"/>
      <d r="CC" s="12" c="14"/>
      <d r="NOMPREP" s="13" c="12"/>
      <g r="CC" s="14" c="7"/>
      <g r="CC" s="14" c="12"/>
      <g r="CC" s="14" c="15"/>
      <d r="PREP" s="14" c="3"/>
      <g r="NOMPREP" s="15" c="17"/>
      <d r="CC" s="15" c="14"/>
      <d r="DET" s="16" c="17"/>
      <g r="DET" s="17" c="16"/>
      <d r="NOMPREP" s="17" c="15"/>
    </dependances>
  </SEQ>
```

Annexe 10
Exemple d'analyse en dépendance par XIP

Phrase analysée :
La Norvège maintient environ 500 hommes dans le pays, principalement à Kaboul et dans le nord.

```
<DEPENDENCY name="SUBJ">
  <PARAMETER ind="0" num="4" word="maintient"/>
  <PARAMETER ind="1" num="2" word="Norvège"/>
</DEPENDENCY>
<DEPENDENCY name="OBJ">
  <PARAMETER ind="0" num="4" word="maintient"/>
  <PARAMETER ind="1" num="10" word="hommes"/>
</DEPENDENCY>
<DEPENDENCY name="VMOD">
  <FEATURE attribute="POSIT1" value="+"/>
  <PARAMETER ind="0" num="4" word="maintient"/>
  <PARAMETER ind="1" num="20" word="principalement"/>
</DEPENDENCY>
<DEPENDENCY name="VMOD">
  <FEATURE attribute="POSIT1" value="+"/>
  <PARAMETER ind="0" num="4" word="maintient"/>
  <PARAMETER ind="1" num="16" word="pays"/>
</DEPENDENCY>
<DEPENDENCY name="VMOD">
  <FEATURE attribute="POSIT2" value="+"/>
  <PARAMETER ind="0" num="4" word="maintient"/>
  <PARAMETER ind="1" num="24" word="Kaboul"/>
</DEPENDENCY>
<DEPENDENCY name="COORDITEMS">
  <PARAMETER ind="0" num="24" word="Kaboul"/>
  <PARAMETER ind="1" num="32" word="nord"/>
</DEPENDENCY>
<DEPENDENCY name="PREPOBJ">
  <PARAMETER ind="0" num="32" word="nord"/>
  <PARAMETER ind="1" num="28" word="dans"/>
</DEPENDENCY>
<DEPENDENCY name="PREPOBJ">
  <PARAMETER ind="0" num="24" word="Kaboul"/>
  <PARAMETER ind="1" num="22" word="à"/>
</DEPENDENCY>
<DEPENDENCY name="PREPOBJ">
  <PARAMETER ind="0" num="16" word="pays"/>
  <PARAMETER ind="1" num="12" word="dans"/>
</DEPENDENCY>
<DEPENDENCY name="DETERM">
  <FEATURE attribute="NUM" value="+"/>
  <PARAMETER ind="0" num="10" word="hommes"/>
  <PARAMETER ind="1" num="8" word="500"/>
</DEPENDENCY>
<DEPENDENCY name="DETERM">
  <FEATURE attribute="DEF" value="+"/>
  <PARAMETER ind="0" num="32" word="nord"/>
```

```
  <PARAMETER ind="1" num="30" word="le"/>
</DEPENDENCY>
<DEPENDENCY name="DETERM">
  <FEATURE attribute="DEF" value="+"/>
  <PARAMETER ind="0" num="16" word="pays"/>
  <PARAMETER ind="1" num="14" word="le"/>
</DEPENDENCY>
<DEPENDENCY name="DETERM">
  <FEATURE attribute="DEF" value="+"/>
  <PARAMETER ind="0" num="2" word="Norvège"/>
  <PARAMETER ind="1" num="0" word="La"/>
</DEPENDENCY>
<DEPENDENCY name="LIEU">
  <FEATURE attribute="PAYS" value="+"/>
  <PARAMETER ind="0" num="2" word="Norvège"/>
</DEPENDENCY>
<DEPENDENCY name="LIEU">
  <FEATURE attribute="VILLE" value="+"/>
  <PARAMETER ind="0" num="24" word="Kaboul"/>
</DEPENDENCY>
```

Table des matières

Remerciements....4
Sommaire....6
Introduction....8
1 - Présentation de l'entreprise....10
1.1 - EADS - European Aeronautic Defence and Space....10
1.2 - La division DS : Defence & Security....11
1.3 - Le département IPCC : Information Processing, Control & Cognition....12
1.3.1 - Présentation générale....12
1.3.2 - La plateforme WebLab....12
2 - Présentation du stage....14
2.1 - Contexte et besoins de l'entreprise....14
2.2 - L'extraction d'information : état de l'art....14
2.2.1 - Un sous-domaine du TAL....14
2.2.2 - Les campagnes d'évaluation....16
2.2.3 - Les outils existants....18
2.2.4 - GATE : General Architecture for Text Engineering....19
2.2.4.1 - Fonctionnement général....19
2.2.4.2 - Le formalisme JAPE....23
2.2.4.3 - Quelques plugins intéressants....24
3 - Création d'une ontologie de domaine....26
3.1 - Qu'est-ce qu'une ontologie ?....26
3.2 - Modélisation d'une ontologie....26
3.2.1 - Formats....27
3.2.2 - Outils existants....27
3.3 - Méthodologie adoptée....28
3.3.1 - Tour d'horizon des ontologies disponibles....29
3.3.2 - Construction d'une taxonomie simple....30
3.3.3 - Affinage selon les besoins du domaine....31
4 - Extraction d'entités nommées....34
4.1 - Constitution de corpus....34
4.2 - ANNIE : observation et amélioration des résultats....35
4.3 - Traitement du français....41
5 - Extraction d'évènements....44
5.1 - Définition de la méthode....44
5.1.1 - Observation des travaux existants....44
5.1.2 - Élaboration d'une approche....45
5.2 - Implémentation dans GATE....47

5.2.1 - Traitement de l'anglais....47
5.2.1.1 - Installation des plugins GATE nécessaires....47
5.2.1.2 - Constitution des gazetteers....48
5.2.1.3 - Développement des règles linguistiques....49
5.2.2 - Traitement du français....53
5.3 - Analyse qualitative et améliorations possibles....54
6 - Extraction de relations....56
6.1 - Méthode d'extraction....56
6.2 - Implémentation dans GATE....56
6.3 - Analyse qualitative et améliorations possibles....58
7 - Évaluation des résultats....60
7.1 - Protocole d'évaluation....60
7.2 - Analyse des résultats....61
7.3 - Observations et améliorations envisagées....63
Conclusion....68
Bibliographie – Sitographie....70
Table des figures....74
Table des tableaux....76
Glossaire....78
Table des abréviations....80
Table des annexes....84
Annexes....85
Résumé....104
Abstract....104

Mots-clés : Traitement Automatique de la Langue (TAL), extraction d'information, fouille de textes, ontologie, entités nommées

Résumé

Aujourd'hui, l'abondance des sources d'information publiques (sites internet, presse, radio, télévision, etc.) a fait émergé le besoin de « fouiller » cette masse de documents afin d'en extraire des connaissances pertinentes dans un but donné. L'équipe IPCC, au sein d'EADS Defence & Security, est chargée de l'innovation en matière de traitement de l'information. Ce mémoire présente une ontologie de domaine et les outils d'extraction de l'information associés pour l'anglais et le français. Après une brève analyse des outils et techniques existants en modélisation d'ontologie et extraction d'information, nous présentons les différents travaux réalisés durant notre stage. Nous avons modélisé, grâce au logiciel Protégé, une petite ontologie de domaine au format OWL, dédiée au renseignement militaire. Afin de repérer dans un texte les différents éléments d'intérêt, nous avons développé, grâce à l'environnement GATE, un outil d'extraction d'entités nommées, évènements et relations. Nous détaillons ici la méthode choisie, les étapes de réalisation ainsi que l'évaluation quantitative et qualitative des résultats obtenus.

Keywords : Natural Language Processing (NLP), information extraction, text mining, ontology, named entities

Abstract

Nowadays, the increasing of information sources (websites, newspapers, radio, TV, etc.) has led to “dig” these documents in order to extract relevant knowledge considering a set purpose. The IPCC team at EADS Defence & Security is responsible for invention in media mining. This thesis introduces a domain ontology and the associated tools for information extraction in English and French texts. After a brief analysis of existing tools and techniques in ontology development and information extraction, we present the work done during our training course. First, we used the Protégé tool to create a small OWL domain ontology dedicated to military intelligence. In order to recognize the elements of interest, we have build, through GATE architecture, a system to extract named entities, events and relations. We present here our methodology, the different stages of implementation as well as a quantitative and qualitative evaluation of our results.