Institut National des Langues et Civilisations Orientales

Département Textes, Informatique, Multilinguisme

« Extraction de citations dans le domaine de la presse avec la résolution d'anaphores »

MASTER

TRAITEMENT AUTOMATIQUE DES LANGUES

Spécialité Ingénierie Multilingue

Par

Lucille Blanchard

Directeur de mémoire :

Frédérique Segond

Encadrant de stage :

Jacques Steinlin

Année universitaire 2015-2016

Table des matières

Remerciements

J'aimerais d'abord remercier toutes les personnes qui m'ont permis d'arriver jusqu'ici.
Un grand merci à ma directrice de mémoire, Frédérique Segond pour ses conseils, sa compréhension, et pour le temps qu'elle m'a consacré et un grand merci également à mon tuteur de stage, Jacques Steinlin, pour m'avoir guidé pendant ces 6 mois, pour tout ce qu'il m'a appris et ses encouragements tout au long de mon stage.
Merci également à tous les professeurs de l'Inalco, Paris 3 et Nanterre pour leurs enseignements et leur dévouement pendant ces deux années.
Merci à ma famille et mes amis pour m'avoir supporté et encouragé dans les moments de stress intense et à ma cousine pour m'avoir offert un toit où travailler au calme.
Et enfin merci à tous mes camarades de master pour ces deux années d'entraide, d'échanges et de rires avec qui j'ai beaucoup appris.

Introduction au sujet

« Le 49.3 est une brutalité, le 49.3 est un déni de démocratie, le 49.3 est une manière de freiner ou d'empêcher le débat parlementaire. »

Ce sont les propos du président François Hollande en 2006, alors premier secrétaire du PS et à l'époque contre le recours au 49.3 pour le projet de contrat première embauche (CPE). Malgré ces propos plutôt clairs sur la position de François Hollande vis-à-vis du 49.3, cette année le gouvernement y a eu recours pour faire passer la loi Travail.

Les députés qui sont contre la loi Travail n'ont pas hésité à la citer à l'Assemblée Nationale pour questionner le gouvernement. Par le biais de la presse, c'est aussi une façon d'interpeler la population et de montrer l'incohérence ou l'hypocrisie du gouvernement.
Comment les opposants à la loi Travail ont-ils retrouvés cette citation de François Hollande datant de 10 ans ? Peut-être ont-ils tout simplement une très bonne mémoire des propos dits par François Hollande. Même si c'est peut-être le cas pour les politiciens qui vivent de la politique, il est plus probable que les journalistes l'aient retrouvée grâce à des recherches dans les archives des journaux car il est légitime de ne pas pouvoir retenir tout ce que les politiciens disent. C'est donc un travail de relecture assez laborieux qu'ont les journalistes. Mais avec l'avancée des techniques dans le traitement automatique des langues et plus particulièrement celui du Big Data textuel, il est désormais possible de retrouver tous les propos de François Hollande, restitués mot pour mot. C'est le but de ce mémoire d'expliquer une des solutions pour le faire.

Dans le domaine du Traitement automatique des langues, l'extraction d'informations est un domaine très courant et un des plus vieux après la traduction automatique. L'extraction d'informations ou « IE » (information extraction) en anglais, est le processus d'extraction des informations pertinentes à partir de données non structurées. Il est apparu avec l'expansion d'Internet dans les foyers, qui génère depuis une production exponentielle de données en tout genre, produites par les internautes eux-mêmes. Désormais, l'internet est plein de données non structurées qui ne demandent qu'à être exploités par les entreprises, pour des finalités marketing, ou d'amélioration des services ou encore par les centres de recherche pour obtenir et exploiter les informations supplémentaires produites par les internautes. Ces informations peuvent être extraites à partir des e-mails, des réseaux sociaux, des forums, des formulaires, etc…

L'IE a différentes finalités : elle permet d'automatiser la lecture d'un document, d'en extraire ses mots-clés, de capter le sens global ou encore de générer un résumé automatique. C'est une pratique développée soit à base de grammaires locales, plus précisément de règles, soit à l'aide de calculs statistiques. Dans les deux cas, elle nécessite une connaissance métier des informations à extraire et du domaine, qu'il soit juridique, financier, ou littéraire.

L'IE se fait essentiellement sur des données textuelles mais l'extraction sur d'autres types de données (orales, manuscrites) commencent à émerger.
Le type d'informations que l'on peut vouloir relever est très variable, il dépend des besoins de l'entreprise et de ce qu'elle souhaite tirer de son corpus de données. Mais en général on commence par extraire les entités nommées car ce sont des informations essentielles pour établir le contexte.

Expert System France, anciennement Temis, est une entreprise spécialisée en extraction d'informations textuelles et c'est dans les bureaux de Paris, au sein du département Professional

Services, qu'a eu lieu le stage.

Les informations que l'on souhaite extraire ici sont les citations et plus précisément celles issues de la presse française. On traitera alors les citations avec ou sans guillemets et on résoudra les anaphores qui sont auteurs de citations.

Nous présenterons dans ce mémoire, le contexte dans lequel il a été réalisé et la problématique à laquelle nous tenterons de répondre. Il y aura une présentation des différents travaux et outils disponibles dans le domaine. Puis le travail d'analyse sur les corpus de presse et la typologie des citations, la mise en place de la solution et les différentes évaluations. Enfin les améliorations apportées, notamment la résolution d'anaphores et celles qui restent à développer.

1. Présentation de l'entreprise, du contexte et de la problématique

Nous présentons dans ce premier chapitre l'entreprise dans laquelle a eu lieu le stage, ainsi que le contexte du projet mis en place et par conséquent sa problématique.

1. PRESENTATION DES DEUX ENTREPRISES ET DE L'EQUIPE PS

Pour comprendre ce mémoire, il est nécessaire de le replacer dans son contexte. Pour ce faire, on présentera dans ce premier chapitre les deux entreprises, Expert System et Temis, le contexte du projet d'extraction des citations et la problématique qui suit ce mémoire.

Expert System France
Expert System est une entreprise italienne basée à Modène, créée en 1989 et qui propose des solutions sémantiques. Elle a racheté en 2015 Temis, Text Mining Solutions, une entreprise fondée en 2000 par les employés du département Text Mining d'IBM et donc spécialiste dans l'analyse et l'extraction de données textuelles non structurées.
Avec la fusion récente d'Expert System et Temis, le travail effectué et, par conséquent, ce mémoire utilisera les technologies des deux entreprises ainsi que les solutions développées pour faire la passerelle entre les différents outils.
La filiale Expert System France comprend plusieurs équipes dont celle des Professional Services (PS). Le PS est consacré aux clients, principalement du secteur privé, qui ont acheté ou souhaitent acheter la solution. Il est composé de 7 consultants, réparti également entre linguistes informaticiens et ingénieurs informaticiens. Chaque consultant doit gérer plusieurs projets et est souvent amené à se rendre chez le client.
Les autres équipes qui la composent sont celles de la Recherche et du Développement (R&D), qui développent les outils, les Projets Innovants qui comme le PS se consacre aux projets des clients mais du secteur public, les commerciaux qui vendent les solutions, ainsi que la direction, les

ressources humaines et l'IT qui gèrent l'entreprise.

Expert System
La technologie d'Expert System repose sur Cogito Studio. Cogito Studio est un environnement de développement pour créer et déployer des applications sémantiques. Il permet de développer des projets de fouille de textes fondés sur une analyse sémantique pour répondre à des besoins d'extraction, de catégorisation ou de désambiguïsation. Il possède un éditeur de développement de règles, un éditeur de taxonomie, des panneaux pour visualiser les sorties (d'extraction, de désambiguïsation et de catégorisation) ainsi que des assistants pour créer des corpus de tests, spécifier la catégorisation attendue pour chaque document, analyser les résultats et accéder à des rapports d'analyse. Les règles développées dans Cogito Studio s'appuient sur l'analyse du désambigüiseur et sur l'analyse sémantique du Sensigrafo, le réseau sémantique multilingue développé par Expert System. Les règles développées dans Cogito Studio se font dans des langages propriétaires : C pour catégorisation, D pour domaine et E pour extraction. Les documents en entrée doivent être au format CogitoStudio et les sorties sont sous format XML.

Temis
De son côté, Temis s'est construit en développant la technologie Luxid qui se compose de plusieurs outils : le Skill Cartridge Builder (SCB), l'Annotation WorkBench (AWB), le WebStudio, et Luxid Information Analytics (LIA). Le SCB est un éditeur qui permet d'écrire des règles d'extraction dans des fichiers sources, qui une fois compilés sont utilisables et exportables. On appelle alors le résultat de la compilation, une cartouche de connaissance (« skill cartridge »). AWB est l'outil d'évaluation de Temis. Il permet d'importer des corpus sur lesquels on applique une cartouche pour évaluer sa qualité (précision, rappel, f-mesure). Le WebStudio est le service Web de Luxid qui permet de configurer et visualisation les résultats de cette cartouche. On importe dans le projet des documents, on configure les paramètres du plan d'annotation, qui peut être composé de cartouches, de convertisseurs, de scripts, etc… On peut aussi créér des ontologies et des thésaurus pour la catégorisation depuis le WebStudio. Il existe des cartouches génériques et donc réutilisables pour différents projets, comme la cartouche TM360, qui extrait les entités nommées et d'autres plus spécifiques, qui sont développées pour des demandes particulières, comme la cartouche Living-Quotes développée pour extraire les citations. LIA est une application web qui permet de faire des analyses sur le projet choisi. Les documents en entrée peuvent être dans n'importe quel format car ils sont convertis selon leur format d'origine, par des scripts de conversion vers le format LUX, un format propriétaire de type XML. Les cartouches sont écrites dans le langage propriétaire de Temis : Temis Script Language (TSL).

Ce mémoire et le stage se situent donc dans un contexte de transition pour fusionner ces deux entreprises et leurs solutions, c'est pourquoi tout au long du mémoire, les deux outils seront utilisés.

2. LES BESOINS DE L'ENTREPRISE

Actuellement, l'entreprise possède une cartouche (Luxid) qui permet d'extraire les citations en

français, en anglais et en allemand, la cartouche Living-Quotes. Mais après comparaison de cette cartouche avec une autre solution d'extraction des citations sur les mêmes documents, il s'est avéré que la cartouche extrayait moins de citations que son concurrent, notamment à cause des anaphores quand elles sont auteurs de la citation.
Le projet est donc né à la suite de ces résultats et à la demande d'un agent commercial de l'entreprise qui souhaite proposer à ses clients de la presse et de l'édition, une solution complète d'extraction des citations.

Au début du stage, il était prévu d'améliorer l'extraction des citations en repartant de la cartouche existante. Mais après avoir suivi deux sessions de formations sur l'outil Cogito Studio et étant donné le contexte de transition entre les deux entreprises, il a été jugé bon d'appliquer les connaissances acquises durant ces formations sur des projets concrets. Les règles d'extraction seront donc implémentées à partir de Cogito Studio.

Une discussion avec l'agent commercial concerné a permis de définir les objectifs. Dans un premier temps, il faut évaluer la cartouche Living-Quotes car il n'y a pas de métriques officielles. Ensuite, identifier ses limites et enfin l'améliorer, en autres, en ajoutant une résolution des anaphores. Dans l'idéal, on souhaite atteindre une précision de 80-90%, et un rappel de 75% et avoir en sortie, les extractions de l'auteur, la citation et la phrase complète.
L'extraction des citations dans la presse servirait, entre autres, aux journalistes, pour retrouver parmi les éditions précédentes, les propos cités de telle ou telle personnalité, sans avoir besoin de tous les relire. L'exemple d'utilisation très concrète que l'on m'a donné est celui de l' « affaire Cahuzac » datant de 2012 et concerne le ministre délégué au Budget Jérôme Cahuzac qui lutte alors activement contre la fraude fiscale. On apprend à cette période que lui-même fraude, malgré le fait qu'il ait affirmé « payer ses impôts » et « ne jamais avoir eu de compte en suisse ».
L'extraction automatique des citations est donc une façon d'avoir à disposition aisément les propos de quelqu'un afin de rappeler au public ou à la personne elle-même ce qu'elle a dit auparavant.

Mais ce n'est pas la seule application possible. L'extraction automatique des citations peut avoir d'autres utilisateurs et finalités. Par exemple dans le domaine d'analyse du discours, la citation peut être un critère déterminant pour évaluer l'influence d'une personne. Une personne influente est une personne que l'on cite de nombreuses fois. Autrement dit, selon le nombre de fois où ses mots auront été cités, on pourra déterminer si la personne en question est influente ou non. C'est le cas, par exemple, des politiciens. Ségolène Royal et sa « bravitude » qui a été repris maintes fois sur les réseaux sociaux et par les médias démontre que c'est une personne influente.
En ce qui concerne les réseaux sociaux, les citations peuvent permettre de détecter des influenceurs sur Internet. Twitter, qui facilite la communication et les échanges dispose du « retweet », principe de partage d'informations qui consiste à publier de nouveau le statut posté par un tiers pour montrer son agrément ou au contraire son désaccord ou encore pour réagir. Facebook intègre aussi un système de partage similaire. Dans les deux cas, ce sont d'autres types de citations. Tout d'abord, cela concernera plus des propos écrits qu'oraux et il n'y aura pas de guillemets pour identifier que ce sont les propos d'un autre, uniquement l'information que cela a été dit par un tiers et répéter par un autre. On peut ainsi exploiter les données des réseaux sociaux pour repérer des djihadistes et ceux qui les recrutent, en analysant qui reprend les mots de qui et dans quelle proportion.

Les applications sont nombreuses mais la problématique définie pour ce mémoire se situe d'abord dans le domaine de la presse.

3. PROBLEMATIQUE

L'objectif de ce mémoire est d'être capable, dans un article de presse, de savoir QUI a dit QUOI. C'est la question à laquelle on doit pouvoir répondre de manière automatisée à la fin de ce mémoire.
Les principaux objectifs du projet sont d'évaluer la cartouche actuelle et d'améliorer l'extraction des citations, notamment à l'aide de la résolution des anaphores dans les cas où elles sont auteurs d'une citation. Par conséquent, on doit chercher des réponses aux questions suivantes :

- Qu'est-ce qu'une citation ? Quelles sont les différentes définitions ? Présentation. Choisir une définition ou définir sa propre définition et s'y tenir du début à la fin du projet.
- Définir les contours d'une citation : qu'est-ce qui est une citation, qu'est-ce qui n'en est pas une ?
- Quels sont les éléments qui la composent ?
- Quels sont les auteurs de citations qui nous intéressent ?
- Comment extraire les citations ?
- Qu'est-ce qu'une anaphore ? Quels sont les différents types ? Lesquelles nous intéressent ?

On s'efforcera dans ce mémoire de répondre à la problématique posée en présentant les différentes étapes qui ont permis d'y arriver.

Ce projet d'extraction des citations arrive dans un contexte assez particulier de fusion entre deux entreprises expertes dans leurs domaines respectifs et les deux technologies seront présentes tout au long de ce mémoire, comme elles le furent pendant le stage. Elles se complètent afin de mener à bien ce projet.

2. État de l'art et progrès

Pour répondre au mieux à la problématique de ce mémoire, il est indispensable d'avoir le maximum de connaissances dans les domaines abordés. C'est pourquoi nous présenterons ici un état de l'art sur les différents domaines que sont l'extraction d'informations, l'extraction de citations et la résolution d'anaphores, avec leurs théories et leurs méthodes comparées et résumées.

1. L'EXTRACTION D'INFORMATIONS DANS LA PRESSE

On présentera dans cette partie, les différentes informations qui peuvent être intéressantes de relever dans la presse, autres que les citations.

Un exemple d'informations recherchées sont les évènements, les entités nommées ou encore les relations (Doddington et al., 2004). Dans le programme ACE (Automatic Content Extraction), ils cherchent à relever ces trois éléments : les entités mentionnées dans le texte et pas seulement les noms, avec l'Entity Detection & Tracking (EDT), les relations entre ces entités grâce au Relation Detection & Charaterization (RDC) et enfin les évènements aux lesquels ces entités participent à l'aide de l'Event Detection & Charaterization (EDC). Le programme ACE extrait ces informations à partir de données audio, image et textuelles en anglais, chinois et arabe.
Une autre information qui est jugée importante est de pouvoir remettre les évènements dans leur contexte temporel (Setzer et Gaizauskas, 2000) en annotant les évènements et les dates dans des corpus de presse. En effet, ils jugent qu'avoir l'information d'un évènement sans connaître sa date est inutile. Ils cherchent donc à extraire des contextes temporels. En ce qui concerne l'extraction des évènements, ils citent des exemples d'évènements comme les attaques terroristes ou les lancements de fusées.
De manière plus générale, on peut extraire dans le domaine des affaires, les acquisitions, les profits et les gains d'une entreprise ou dans le domaine médical, les interactions entre les médicaments ou l'impact de certaines protéines (Christopher Manning, Stanford University). Evidemment, l'extraction des entités nommées (NER) est courante dans le domaine de la presse, autant que dans les autres domaines.
L'extraction d'informations peut également servir à la compréhension des articles de presse, notamment en extrayant les sujets et les objets directs autour des verbes comme assassiner, bombarder ou kidnapper (Riloff, 1999) pour mieux comprendre ce qui s'est passé. Cette méthode s'appuie sur une connaissance du monde (World Knowledge).

4. L'EXTRACTION DE CITATIONS

Une citation est le propos d'une personne rapporté tel quel ou paraphrasé à l'aide d'autres termes. Dans le premier cas, on introduit la citation avec des guillemets. Dans le second, il n'y a pas de guillemets puisqu'ils servent à indiquer les mots prononcés par la personne. Il est donc logique que dans une phrase où les propos ont été paraphrasés, on ne trouve pas de guillemets. Mais dans ce genre de citation paraphrasée, on trouve d'autres indices dans le verbe qui introduit la paraphrase. En ce qui concerne la personne qui est auteur de la citation, il peut s'agir d'une personne physique, clairement identifiable mais aussi d'une compagnie ou d'une organisation qui forme ce qu'on appelle une « personne morale ». La solution d'extraction des citations présentée dans ce mémoire couvrira cette définition de la citation.

Je me suis fortement inspiré de la définition d'une citation du chercheur Ghassan Mourad ou plutôt j'ai trouvé dans sa définition ce que je cherchais. Sa description d'une citation a en effet permis de réduire mes doutes sur la question : qu'est-ce qu'une citation réellement ? Dès le début de ce projet, on m'a fait comprendre que j'étais la seule à pouvoir répondre à cette question. Par conséquent, même dans les situations où j'ignorais si ma phrase était une citation ou non, personne n'avait la réponse et je devais la trouver moi-même.

Ghassan Mourad dit qu'il existe deux types de citations : les citations directes qui utilisent les

guillemets et qui reprennent les propos tels quels et les citations indirectes qui les paraphrasent. Il a écrit une thèse sur l'extraction automatique des citations dans laquelle il s'appuie beaucoup sur les signes typographiques. Signes sur lesquels je m'appuie pour mon module d'extraction des citations mais auxquels j'ajoute également des éléments linguistiques.

La difficulté de l'extraction des citations réside dans la multitude de formes sous laquelle une citation peut se présenter. De même pour l'auteur de la citation. Par exemple, les phrases ci-dessous sont toutes des citations :

"Le PLD devra se renouveler. Si ce n'est pas le cas, un système bipartite ne s'implantera jamais au Japon", a jugé M. Hatoyama.
"Il y a beaucoup de brouhaha autour de Martin et nous avons décidé de ne pas y contribuer en refusant d'alimenter les rumeurs", a déclaré son père, Hans Erik, lui-même un ancien pro, à l'AFP.
"La seule chose qui importe, c'est qu'il puisse se développer en tant que footballeur".
Selon la CGT, la direction refuse de négocier.
Selon l'hebdomadaire Profil, l'Autriche va devoir débourser cette année "13 milliards d'euros pour les retraites, soit plus que ce qu'elle dépense pour les écoles et les universités".

Mais le plus difficile à extraire n'est pas vraiment la citation mais plutôt son auteur.

Un article publié par l'Université de Louvain, en Belgique, prend en compte les citations indirectes ou sans guillemets comme des citations à extraire des articles de presse (Weiser et Watrin, 2012). Ce type de citations sans guillemets peuvent être formalisées par 16 patterns qui sont utilisés pour développer des grammaires d'extraction. Ils soulèvent également la question de la limite de la partie citée dans ce type de citations où il n'y a pas de guillemets pour la délimiter. Ils ont utilisé deux ressources pour extraire ces citations :
Un dictionnaire de verbes qui introduisent la citation et construit à partir des verbes déclencheurs de citations avec guillemets
Une grammaire qui formalise les 16 patterns avec Unitex

D'autres ont utilisés deux ressources de TAL indépendantes (Krestel et al., 2008) :

- Le Reporting Verb Marker qui détecte et taggue les verbes déclencheurs de citations. Il génère une liste de verbes
- Le Reported Speech Finder qui trouve les citations et taggue ses constituants

Ils considèrent que les citations sont de plus en plus importantes pour vérifier la fiabilité d'une information journalistique. Ils définissent eux aussi deux types de citations : directes et indirectes. Deux chercheurs portugais ont créé un logiciel nommé « Verbatim » pour structurer et filtrer la masse d'informations que l'on reçoit (Sarmento et Nunes, 2009). Le logiciel fonctionne pour le portugais uniquement et a pour but d'extraire et classifier les citations et les sujets à partir d'articles de presse. Ils présentent leurs résultats sur une interface web. Ils pensent pouvoir utiliser leur outil comme un « watchdog » afin de comparer les citations d'une même entité et sur le même sujet au fil du temps. Ils admettent qu'il existe une multitude de manière dont la citation peut être exprimée. Selon eux, la citation directe reprend exactement les propos mot pour mot tandis que la citation indirecte est paraphrasée par le journaliste. Ils évitent la résolution des anaphores en ne relevant que les citations dont l'auteur est explicite.

Enfin certains pensent que les citations et ses constituants sont primordiales pour la recherche de plagiat, car une reprise avec citation de sa source aura tendance à être licite tandis qu'une reprise

sans source citée explicitement peut être considérée comme illicite (Poulard et al., 2008). Leur problématique est la suivante : Qui dit quoi sur qui ou quoi et comment ?
Ils réalisent que le repérage des citations et de ses constituants est une tâche complexe dû aux différents types d'informations par lesquels on peut caractériser une citation.
Ils testent deux méthodes pour classifier les discours : une première avec des automates à états finis et une deuxième par apprentissage sur leur corpus annoté manuellement. De même pour la tâche d'identification des constituants d'une citation : une méthode selon l'appartenance aux discours repris (citations directes) ou l'appartenance aux expressions locuteurs.
Ils utilisent pour l'implémentation de leur méthode, un corpus issu de plusieurs journaux différents (Reuters, AFP, Le Soir, Le Figaro, etc…)

3. LA RESOLUTION D'ANAPHORES

Une anaphore est, selon la définition de Wikipédia, un mot ou un syntagme qui sert à reprendre un terme déjà vu plus tôt dans l'énoncé. Elle évite ainsi des répétitions.
Les anaphores peuvent se présenter sous la forme d'un nom, d'un pronom ou d'un adverbe. Pour le projet actuel, on se concentre uniquement sur les anaphores pronominales puisque ce sont les seules qu'on trouve dans les citations de la presse. Un autre type d'anaphores, les anaphores implicites, est également évoqué dans ce mémoire mais n'a pas pu être résolu, faute de temps.
Une anaphore implicite est le nom non officiel donné aux citations entre guillemets qui constitue une phrase entière et dont l'auteur se situe dans une autre phrase.
Exemple :
"C'est une histoire d'amour très pure" entre deux êtres en marge de la société, déclare-t-il aujourd'hui. "Pour moi il n'y avait aucune ambiguïté".

La résolution d'anaphores consiste à résoudre ce à quoi se réfère un pronom ou groupe nominal, dans le cas présent uniquement un pronom.
Dans l'exemple ci-dessous, on comprend dès la première lecture que l'anaphore pronominale « elle » fait référence à Rachel Lambert. Mais si l'anaphore est aussi naturelle pour l'humain, il est plus compliqué de l'expliquer à un ordinateur. Car il faut comme l'humain qu'il prenne en compte le contexte.

Exemple :
"Respecter Vincent c'est reconnaître qu'il n'aurait jamais supporté le moindre acharnement thérapeutique": en accord avec les médecins, Rachel Lambert réclame que son mari maintenu artificiellement en vie dans un état végétatif depuis 2008, puisse "partir dignement". Dans sa voix douce et timide, chaque mot est pesé, souvent ponctué par le silence. La jeune femme de 33 ans, le front barré d'une mèche blonde, protège farouchement sa vie privée pour ne parler que de son mari. "Je suis son épouse, je veux faire connaître l'homme qui était debout avant cet accident qui l'a emporté. Mon seul combat c'est le respect de sa personnalité et de ses volontés", affirme-t-elle.

Pour résoudre les anaphores pronominales, il y a deux indices élémentaires à prendre en compte :

- Le pronom est après son antécédent
- Il doit y avoir correspondance en genre et en nombre

L'algorithme de Hobbs de 1978 permet de résoudre 90% des anaphores pronominales. Il utilise la syntaxe pour trouver le NP candidat correspondant au pronom anaphorique à la 3ème personne. Il effectue une analyse syntaxique de surface : il choisit le dernier NP de la phrase courante et s'il n'y en a pas, il prend comme antécédent, le premier des phrases précédentes. Tout en vérifiant les contraintes de base de genre et de nombre.
En général, les méthodes orientées syntaxe ne suffisent pas à résoudre les anaphores bien qu'elles y ont un rôle important. L'algorithme de Hobbs repose sur différentes contraintes syntaxiques. Bien que son algorithme soit peu coûteux à implémenter et génère d'excellents résultats, il est désormais considéré comme inadéquat.

Lappin & Leass quant à eux, veulent résoudre les pronoms de la 3ème personne et les anaphores lexicales. Ils s'appuient eux aussi sur la syntaxe ainsi que sur des mesures de prépondérance. Leur algorithme contient les éléments suivants :

- Un filtre syntaxique pour exclure la dépendance anaphorique d'un pronom vis-à-vis d'un NP à l'intérieur d'une même phrase
- Un filtre morphologique pour exclure la dépendance anaphorique d'un pronom envers un NP s'il ne correspond pas en genre, en nombre et en personne
- Une procédure pour identifier les pronoms vides de sens
- Un algorithme de liens pour identifier l'antécédent potentiel d'une anaphore lexicale
- Une procédure qui attribue une valeur à différents paramètres de prépondérance
- Une procédure qui identifie les NPs liés anaphoriquement comme une classe équivalente et qui lui attribue une valeur de prépondérance égale à la somme de la prépondérance de ses éléments.
- Et enfin, une procédure de décision pour choisir l'élément préféré parmi une liste de candidats pour un pronom

Enfin Ruslan Mitkov a également travaillé sur la résolution des anaphores. En 1994, il choisit de combiner les méthodes statistiques et linguistiques et en choisissant de se restreindre à un genre particulier plutôt qu'à la langue générale. Son modèle intègre des connaissances sémantique, syntaxique, heuristique, de domaines et de discours.
En 1995, il introduit l'approche sur le raisonnement de l'incertitude. Il se justifie avec les raisons suivantes :

- Même si les informations sur les contraintes et les préférences sont disponibles, il est normal d'assumer le fait que le programme de compréhension du langage naturel de peut pas comprendre la totalité du document en entré
- De plus, les scores des contraintes et des préférences ont été faites par des humains et sont subjectifs

Et en 1997, il propose une méthode qui combine ces deux dernières solutions.

Beaucoup d'autres chercheurs travaillent également sur résolution des anaphores. C'est le cas notamment de Christopher Kennedy et de Branimir Boguraev qui en 1996 proposent une méthode de résolution des anaphores pronominales qui ne nécessite pas de parser. Ils s'appuient sur une version modifiée et étendue de celle développée par Lappin & Leass en 1995. Contrairement à ces

derniers, leur algorithme ne requiert pas d'analyse syntaxique profonde et complète du texte. À la place, en acceptant un compromis au niveau de la qualité de la sortie, ces modifications permettent de résoudre les anaphores en partant d'une sortie tagguée en partie du discours et enrichie des annotations des fonctions grammaticales des éléments lexicaux du texte. Les chercheurs de l'Université d'Auckland ont de leur côté développé une méthode qui utilise une représentation sémantique superficielle des phrases pour résoudre les anaphores pronominales (Ho et al., 2004).

4. OUTILS UNIVERSITAIRES ET INDUSTRIELS EXISTANTS

Outil universitaire

La Plateforme SAPIENS de l'INRIA

L'INRIA (Institut National de Recherche en Informatique et en Automatique) est un institut de recherche français en mathématiques et informatique. C'est un organisme public de recherche dédié aux sciences et technologies du numérique.
L'INRIA a développé sa propre plateforme d'extraction et de visualisation des citations appelée SAPIENS (Station d'Analyse Profonde via Internet pour l'Exploration de NewS). Cette plateforme propose de rechercher les citations d'une personne à travers un nuage d'entités.

Figure 1 - Nuage d'entité de SAPIENS

Il suffit alors de cliquer sur une entité pour obtenir ses propos sous la forme d'un tableau avec dans la colonne de gauche, les dépêches dans lesquelles sont tirées les citations et dans la colonne de droite, la citation.

Figure 2 - Résultats de SAPIENS sur l'entité Éric Woerth

Le tableau est introduit par une brève présentation de l'auteur. En cliquant sur une dépêche, une fenêtre s'ouvre avec la dépêche entière.
SAPIENS gère la résolution des anaphores et des coréférences pour le français. Elle extrait notamment les citations sans guillemets. Les citations sont toutes issues des dépêches de l'AFP, par conséquent SAPIENS ne gère que les citations en français.

Solution industrielle

Trooclick

Trooclick est une entreprise française qui a développé une plateforme en ligne, storyzy, permettant de visualiser toutes les citations d'une personne ou de chercher quelle personnalité a mentionné un terme en particulier. On peut spécifier la période de temps dans laquelle on recherche la citation. Elle ne fonctionne que pour l'anglais et génère uniquement des citations issues de la presse, comme le site de bbc ou ndtv. Elle n'extrait que les citations directes, autrement dit celles entre guillemets mais elle gère la résolution des anaphores pour l'anglais. La plateforme donne un lien direct vers l'article en ligne, origine de la citation.

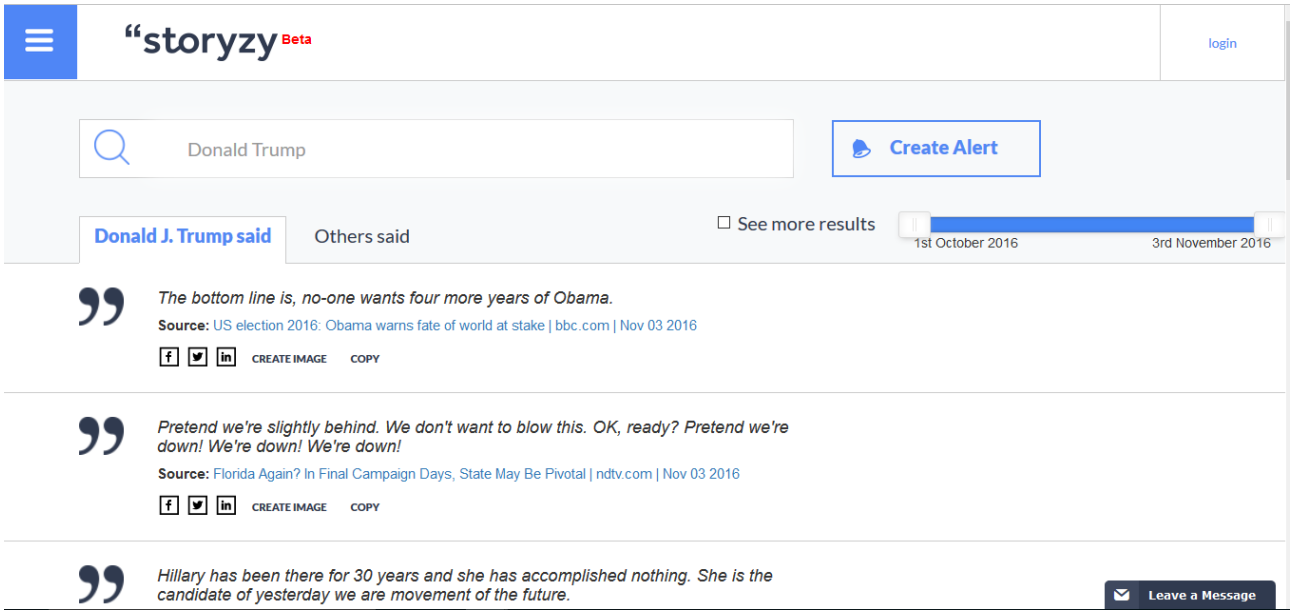


Figure 3 - Résultats de recherche de citations de Donald Trump

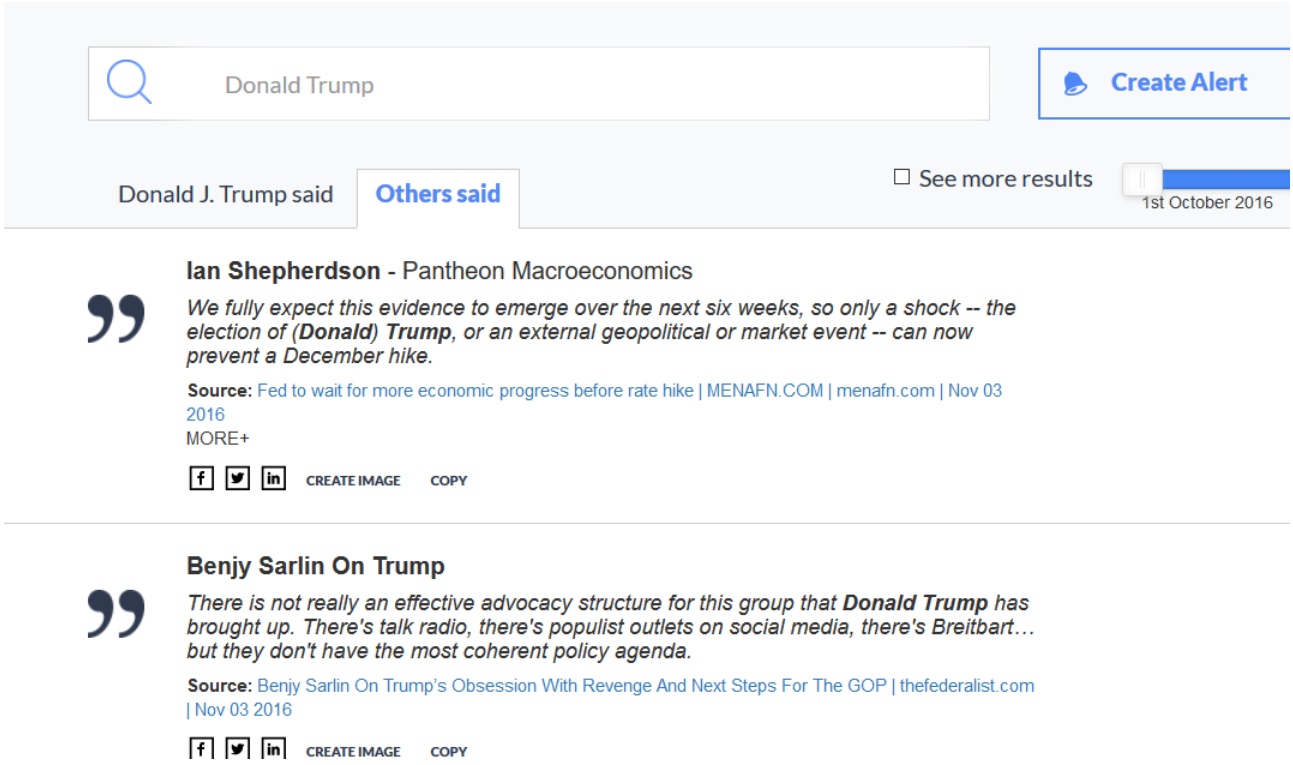


Figure 4 - Résultats de recherche de citations qui contiennent « Donald Trump »

Le Match des mots

Lors des élections présidentielles de 2012, l'AFP et Médialab déploie sur le site de libération.fr un « comparateur de citations » qui proposent aux internautes une plateforme d'extraction et de visualisation de citations politiques issues des dépêches de l'AFP. L'idée était de permettre aux citoyens de comparer les discours des politiciens avant les élections mais aussi de servir d'instrument de travail pour les professionnels des médias et des sciences politiques. L'AFP

travaille sur l'amélioration du comparateur pour l'étendre à d'autres domaines, tel que le sport ou l'économie.
De plus, l'Agence France-Presse utilise actuellement un moteur de recherche des citations, par auteur ou mot-clé.

Figure 5 - Capture d'écran du comparateur de citations de l'AFP en 2012

Ces outils, déjà disponibles sur le marché, donnent des idées d'implémentation pour une future solution d'extraction des citations.

5. POSITION PAR RAPPORT A L'ETAT DE L'ART

Comme la plupart des travaux sur l'extraction des citations, la solution mise en place et présentée dans ce mémoire prend en compte toutes les types de citations, avec ou sans guillemets, quelque que soit l'ordre de l'auteur, de la citation et du déclencheur dans la phrase. Les anaphores sont résolues pour ne manquer aucune citation dont l'auteur serait remplacé par un pronom.
Le corpus sur lequel la solution a été développée est l'AFP, une référence pour la presse française. Les articles vus précédemment qui ont pour but d'extraire les citations, utilisent aussi des corpus issus de la presse. Les tests finaux sont faits sur un autre corpus de l'AFP mais également sur des articles du Monde et de Ouest-France.
De plus, l'algorithme de résolution d'anaphores présenté dans ce mémoire est plus simple et plus rapide à implémenter que les méthodes présentées dans l'état de l'art, tout en retournant des résultats satisfaisants. Il n'utilise pas d'analyse syntaxique, contrairement à la plupart des travaux présentés.
La solution mise en place concerne uniquement les citations dans le domaine de la presse française. Elle est faite à partir de règles (extraction) et d'un script java (résolution).

Nous avons désormais une vue assez large des méthodes et des théories qui existent sur les différents domaines abordés dans ce mémoire. Nous nous en inspirons pour implémenter la

solution d'extraction de citations et la résolution d'anaphores. Mais avant de commencer, nous devons d'abord évaluer la solution actuelle et définir les bases du projet en répondant aux questions posées dans la problématique sur les citations et les anaphores. C'est le sujet du chapitre suivant.

3. Évaluation préliminaire

Nous aborderons dans ce chapitre l'évaluation de la cartouche Living-Quotes et la typologie des citations qui définit précisément une citation et ses différents types. Etant donné que le but de ce projet est d'améliorer la qualité, il est nécessaire de connaître les chiffres de la solution de départ. Les corpus utilisés pour l'évaluation, les critères de validation et les résultats sont également décrits.

1. TYPOLOGIE DES CITATIONS

Pour pouvoir valider la qualité de la cartouche Living-Quotes avec AWB, l'outil de validation de Luxid, il faut d'abord avoir une définition précise de ce qu'est une citation. A la lecture des articles de presse, certaines citations sont évidentes et d'autres non. Les citations les moins évidentes concernent surtout celles dont l'auteur n'est pas directement identifiable et celles qui sont sans guillemets.
Exemple de citation évidente :
"Ces musiques peuvent être extraites du cadre mémoriel : elles peuvent être légères, pimpantes, gaies ...", dit M. du Closel.

Exemples de citations moins évidentes :
M. Aso a annoncé qu'il assumait la responsabilité de cette cuisante défaite et qu'il allait démissionner de la présidence du PLD, probablement après l'élection du nouveau Premier ministre. (Sans guillemets)
Chez les turcophones ouzbeks, kazakhs ou kirghizes, "il y a davantage de différences entre populations d'un même groupe ethnique pour le chromosome Y, alors que l'opposé est observé pour l'ADN mitochondrial", qui se transmet via les femmes (car seul l'ovule l'apporte à l'embryon), d'après l'étude. (Auteur non identifiable)

Pour dissiper les doutes, des recherches sur le sujet ont été faites.

Premièrement qu'en disent les dictionnaires ?
D'après le dictionnaire Larousse, une citation est l' « action de citer, de rapporter les mots ou les phrases de quelqu'un ; paroles, passage empruntés à un auteur ou à quelqu'un qui fait autorité ».
D'après le site linternaute, c'est un « Extrait, passage d'un auteur rapporté mot pour mot. »
D'après le CNRLT, une citation est l' « Action de citer un passage d'auteur, de reproduire exactement ce qu'il a dit ou écrit, oralement ou dans un texte. »

Ensuite quels sont les avis des chercheurs en linguistique et en TAL sur la citation ?

D’après le chercheur libanais Ghassan Mourad, “une citation est tout texte ou fragment textuel rapporté, écrit ou prononcé. Un fragment textuel peut être cité « à la lettre » et entre guillemets : citation directe ; il peut être paraphrasé ou cité à l’aide « d’autres mots » : citation indirecte. Donc toute donnée rapportée, qu’elle soit entre guillemets ou non est une citation.”

Enfin, aux vues de toutes ces définitions, ma vision d’une citation se rapproche de celle de Ghassan Mourad. En effet, je suis d’avis qu’une citation peut être paraphrasée et comme il le dit, elle fait alors partie des citations indirectes.

Après avoir déterminé ce qu’est une citation, on constate qu’elles peuvent se diviser en deux parties distinctes : les citations avec guillemets (citations directes) et les citations sans guillemets (indirectes). Ces dernières n’ont pas de guillemets mais sont tout de même considérées comme des citations car elles peuvent avoir été paraphrasées par le journaliste.

Les citations sans guillemets sont plus risquées à extraire, autrement dit une génération excessive de bruit est possible, car elles sont dépourvues de l’élément majeur d’une citation ordinaire : les guillemets. Elles demandent donc plus de contraintes sur les autres éléments restants, soit l’auteur et le déclencheur.

Les éléments pouvant composer une citation sont :
Obligatoire :

- Un auteur (personne physique, organisation, compagnie, pronoms, coréférences, etc…)
- Un verbe déclencheur (déclarer) ou une préposition (selon, d’après)

Facultatif :

- Des guillemets (« »)
- Deux-points

Les combinaisons correspondent à l’ordre dans lequel sont les éléments dans une phrase. Ces éléments sont :

- Author : l’auteur de la citation
- Trigger : le déclencheur, verbe ou préposition
- Quote : la citation

Par exemple, on peut trouver une citation dans cet ordre : l’auteur, le déclencheur et la citation (Author Trigger Quote = ATQ).
Exemple :
Il s'est ensuite interrogé : "Est-ce que l'aviation, la technique n'a pas évolué d'une façon que l'on n'a pas pris en compte suffisamment ? Cela renvoie à la formation, aux caractéristiques des avions. En tout cas, il faut se poser des questions".

Mais on peut aussi avoir la citation suivie du déclencheur et de l’auteur (Quote Trigger Author = QTA).
Exemple :
"L'Europe recevra tout le gaz que lui fournira la Russie", a affirmé M. Sokolovski.

Ci-dessous la liste des caractéristiques linguistiques, typographiques et positionnels qui ont permis

d’établir cette typologie :

- Avec ou sans guillemets
- Type de déclencheur (verbe ou préposition)
- Phrase entière ou extraits
- Ordre des éléments (ATQ, QTA, QTAQ)
- Deux points (:)
- Nombre d’auteurs

Chaque cas de la typologie est décrit et détaillé ci-dessous avec :

- une représentation formelle en couleurs
- des exemples tirés du corpus de l’AFP
- une description
- les combinaisons possibles qu’ils couvrent
- le nombre de citations qu’il peut contenir.

Pour chaque cas présentés ci-dessous, la citation peut aussi contenir des détails ou « précisions » sur l’auteur (“Jean François Pontal, Directeur Général d’Orange”) ou sur le contexte spatio-temporel dans lequel les propos ont été dits (“dans la matinée”). En général, ces précisions se trouvent entre l’entité nommée et le verbe, c’est pourquoi il est important de les prendre en compte en écrivant les règles d’extraction.

Un tableau récapitulatif est disponible en annexe.

AVEC GUILLEMETS

Pour les citations avec guillemets, tous les types d’auteur sont autorisés.
Il n’y a pas de contrainte sur le verbe déclencheur.

Cas n°1

“passage cité”, verbe précisions entité nommée

Exemples :

- "Ces musiques peuvent être extraites du cadre mémoriel : elles peuvent être légères, pimpantes, gaies ...", dit M. du Closel.
- Le Premier ministre Ehud Olmert, dont les propos ont été rapportés par un haut responsable israélien, a affirmé que "l'opération n'a pas été lancée à Gaza pour qu'elle se termine avec le même nombre de tirs de roquettes qu'au début de l'offensive".

Note : La spécificité de ce type de citation est qu’elle contient une seule citation, qui en général est

une phrase complète qui est compréhensible sans avoir besoin de la remettre dans le contexte. La citation commence ou finit la phrase.

Deux combinaisons sont possibles pour ce cas :
- Quote Trigger Author (QTA)
- Author Trigger Quote (ATQ)

Nombre de citations possible pour ce cas : 1

Cas n°2

entité nommée précisions verbe **:** “passage cité”

Exemples :
- Jean François Pontal, Directeur Général d’Orange a déclaré : "Les activités d’Orange continuent de gagner en puissance”
- Face aux questionnements, le PS martèle dans un communiqué sa position : "Nous prônons la création d'une contribution énergie-climat dont le produit serait intégralement consacré à des compensations sociales liées à la lutte contre le changement climatique".

Note : La spécificité de cette citation est qu’elle est introduite par deux points (« : »).

Une seule combinaison possible pour ce cas : Author Trigger Quote (ATQ).

Nombre de citations possible pour ce cas : 1

Cas n°3

phrase “passage cité” phrase, verbe précisions entité nommée

Exemples :
- Une partie de ces équipements ont été fournis par la Russie, dont les bombardements ont commencé à "diminuer la capacité de combat (...) des groupes terroristes", a souligné le général Ayoub.
- Manuel Valls a défendu, mercredi 15 juin sur France Inter, l’action du renseignement et de l’antiterrorisme en France après l’assassinat d’un policier et de sa compagne lundi soir, affirmant qu’il n’y avait eu ni « négligence » ni « manque de discernement » dans le suivi du tueur de Magnanville

Note : Contrairement au cas n°1, ce type de citations contient des extraits de phrases citées, qui ne

sont pas compréhensibles sans la phrase complète. Il est supposé que le journaliste paraphrase les propos tout en citant entre guillemets certaines parties.

Deux combinaisons possibles pour ce cas :

- Quote Trigger Author (QTA)
- Author Trigger Quote (ATQ)

Nombre de citations possible pour ce cas : entre 1 et 5

Cas n°4

"passage cité", selon/d'après/pour précisions entité nommée

Exemples :

- Nombre de partitions écrites par ces auteurs mis à l'index ont été détruites, ou égarées, et après-guerre "il ne restait que très peu de traces de ces œuvres", selon M. du Closel.
- Même constat chez Oleg Orlov, président de l'organisation de défense des droits de l'homme Mémorial: "Ce sont les extrémistes qui ont pris le dessus dans la lutte clandestine. La place des séparatistes (...) a été prise par ceux qui veulent réaliser leur utopie d'un Etat islamiste dans le Caucase".
- D'après la CGT, "aucune sanction n'est prévue s'il n'y a pas d'accord, mais il existe un risque que le patronat perde ce cadeau fiscal".

Note : Citations avec une préposition comme déclencheur. Les prépositions qui déclenchent ce type de citations sont :

- Selon
- D'après
- Pour
- Chez

Attention ! La préposition « pour » génère plus de bruit que les autres, à utiliser avec précaution.

Deux combinaisons possibles pour ce cas :

- Quote Trigger Author (QTA)
- Trigger Author Quote (TAQ)

Nombre de citations possible pour ce cas : entre 1 et 3

Cas n°5

“passage cité, verbe/préposition précisions entité nommée. passage cité”

Exemples :

- « D'année en année, les quotas ne sont pas remplis et ce nouveau quota n'a donc pas lieu d'être aussi élevé, a déclaré Truls Gulowsen, le président de Greenpeace Norvège. Mais ce n'est pas un gros problème : c'est une activité qui est en voie d'extinction dans la mesure où les consommateurs préfèrent une pizza à la viande de baleine ».
- "Or, ajoute-t-il, les lettres recommandées qui contiennent des cartes bleues, des objets de valeur, sont régulièrement la cible des braqueurs".

Note : La spécificité de cette citation est qu’elle inclut le déclencheur et l’auteur entre les guillemets.

Une seule combinaison possible pour ce cas : Quote Trigger Author Quote (QTAQ).

Nombre de citations possible pour ce cas : 1

Cas n°6

« passage cité », verbe/préposition entité nommée, verbe/préposition « passage cité »

Exemples :

- Selon Yann Marec du Midi Libre, les socialistes "ont désormais une patronne", mais il avertit contre les "assauts répétés de ses alliés potentiels prompts à bombarder le navire amiral de la gauche".
- "Discutons, mettons-nous autour d'une table, trouvons une possibilité qui soit socialement soutenable", a-t-il ajouté, soulignant que "la lutte contre la dégradation climatique ne peut pas attendre que la gauche arrive au pouvoir".
- "Il faut de l'altruisme, il faut de la générosité", il est "temps de dialoguer et de parler", a-t-elle poursuivi, ajoutant : "il y a des membres du PS et des écolos qui veulent que ça bouge, ils ont raison, je suis avec eux".

Note : La spécificité de cette citation est d’avoir 2 déclencheurs et donc au moins 2 citations qui se situent autour de l’auteur.

Quatre combinaisons possibles pour ce cas :

- Quote Trigger Author Trigger Quote (QTATQ)
- Author Trigger Quote Trigger Quote (ATQTQ)
- Trigger Author Quote Trigger Quote (ATQTQ)
- Trigger Quote Author Trigger Quote (TQATQ)

Nombre de citations possible pour ce cas : entre 2 et 4

Cas n°7

entité nommée précisions verbe/préposition « passage cité », entité nommée précisions verbe/préposition « passage cité »

Exemples :

- Alors que le journal dit maintenant s'attendre à l' « offensive de Berlin », un porte-parole du ministère allemand des Finances a seulement commenté que la décision était une « initiative bienvenue ».
- Foudres de Daniel Cohn-Bendit qui a jugé la finaliste à la présidentielle "à l'Est", alors que la numéro un des Verts, Cécile Duflot, invitée de La Rochelle, fustigeait des "démagogues déconnectés de la réalité" et défendait la fiscalité écologique, "première pierre d'une nouvelle redistribution".

Note : La spécificité de cette citation est que, dans la même phrase, il y a : 2 auteurs distincts, 2 déclencheurs et donc au moins 2 citations.

Deux combinaisons possibles pour ce cas :
- Author Trigger Quote Author Trigger Quote (ATQATQ)
- Trigger Author Quote Author Trigger Quote (TAQATQ)

Nombre de citations possible pour ce cas : entre 2 et 3

SANS GUILLEMETS

Les cas sans guillemets sont à traiter avec plus de précaution.

Pour les citations sans guillemets, seuls les types d'auteur listés ci-dessous sont autorisés :
- Une personne physique
- Une compagnie
- Une organisation

Cas n°8

entité nommée verbe que/de précisions passage rapporté

Exemples :

- Vivendi Universal a annoncé ce jour que son offre d'achat contre espèces de toutes les

actions ordinaires en circulation de Houghton Mifflin avait expiré à minuit, heure de New York, le 6 juillet 2001.
- Weidong Yin a déclaré qu'il espérait produire deux millions de doses par mois.

Note : La spécificité de ce cas est que l'on extrait uniquement les citations dont le verbe déclencheur suggère que l'auteur a dit ces propos mais d'une autre manière et qu'ils ont été paraphrasés par le journaliste.

Les verbes déclencheurs sont : déclarer et annoncer. De plus, le verbe doit absolument introduire une proposition subordonnée (que/de) et jamais un groupe nominal.

Une seule combinaison possible pour ce cas : Author Trigger Quote (ATQ)

Cas n°9

passage rapporté, verbe précisions entité nommée

Exemples :

- Le processus nécessaire pour envoyer une photo prend moins d'une minute, et le coût est de 20-25 cents par déclic, a déclaré M. Kahn
- La Fédération internationale de l'automobile (FIA) enquête sur l'accident de l'ex-pilote Renault Nelson Piquet Jr. au Grand Prix de Singapour 2008, qui aurait été prémédité afin de faciliter la tâche de Fernando Alonso, a affirmé dimanche la télévision brésilienne Rede Globo.

Note : La spécificité de ce cas est que l'on extrait uniquement les citations dont le verbe déclencheur suggère une déclaration de la part de l'auteur. Les verbes déclencheurs sont :

- Déclarer
- Annoncer
- Affirmer
- Indiquer
- Préciser
- Assurer
- Dire
- Expliquer
- Souligner
- Ajouter

Une seule combinaison possible pour ce cas : Quote Trigger Author (QTA)

Cas n°10

selon/d'après/pour précisions entité nommée, passage rapporté

Exemples :

- Selon Bernard Conio, Directeur des Ventes et du Marketing, NEC France devient ainsi l'un des leaders sur le marché français des systèmes de transmission hertzienne.
- Il s'agit parfois de personnages en surpoids ridiculisés (deux cas), d'hommes au physique ingrat (petits, gros, chauves, boutonneux...) mais attirant de jolies filles grâce à leur argent (cinq cas) ou d'allusions douteuses à la sexualité défaillante de personnages âgés (deux cas), selon l'ARPP.

Note : Comme pour le cas n°4, le déclencheur ici est une préposition. Les prépositions qui déclenchent ce type de citations sont :

- Selon
- D'après
- Pour
- Chez

Attention ! Les phrases qui contiennent « selon des témoins », « selon les services d'urgence » ou « selon la police » ne doivent pas être extraits car l'auteur n'est pas identifiable, elles ne font pas référence à quelqu'un en particulier.

Deux combinaisons possibles pour ce cas :

- Trigger Author Quote (TAQ)
- Quote Trigger Author (QTA)

Au total, une dizaine de citations a pu être identifiée, chacune étant unique.

La typologie des citations s'est faite à partir de la lecture des articles de presse utilisés lors de la validation de la cartouche Living-Quotes. Elle s'est enrichie au fur et à mesure des lectures de dépêches et a été revue plusieurs fois pour donner ce qui est présentée ci-dessus. Toutes les citations rencontrées et validées comme correctes ou partiellement correctes sont enregistrées dans un fichier regroupant environ 1000 citations (CitationsAFP.txt). De même, les citations ayant été validées comme incorrectes, appelées les « fausses citations », sont regroupées dans un autre fichier (FaussesCitationsAFP.txt).

Au contraire, qu'est ce qui n'est pas considéré comme une citation ?
Tout d'abord, ce qu'on appelle ici les fausses citations. Des mots entre guillemets mais qui n'ont rien d'une citation, excepté les guillemets.

Exemple :
Frank Evans, un ancien boucher connu en Espagne, comme "l'Anglais", a tué deux taureaux d'environ 500 kilos dans la petite place de Benalmadena, près de Malaga en Andalousie (sud).

Ensuite toutes les citations sans guillemets qui vont au-delà de la définition d'une citation. Ces phrases qui ont un auteur, un verbe déclencheur de type annoncer, déclarer mais qui n'introduisent pas de proposition subordonnée complétive conjonctive. En résumé, toutes les phrases qui après le verbe déclencheur ne contiennent pas de conjonction de coordination « que » ne sont pas des citations paraphrasées.

Exemples :

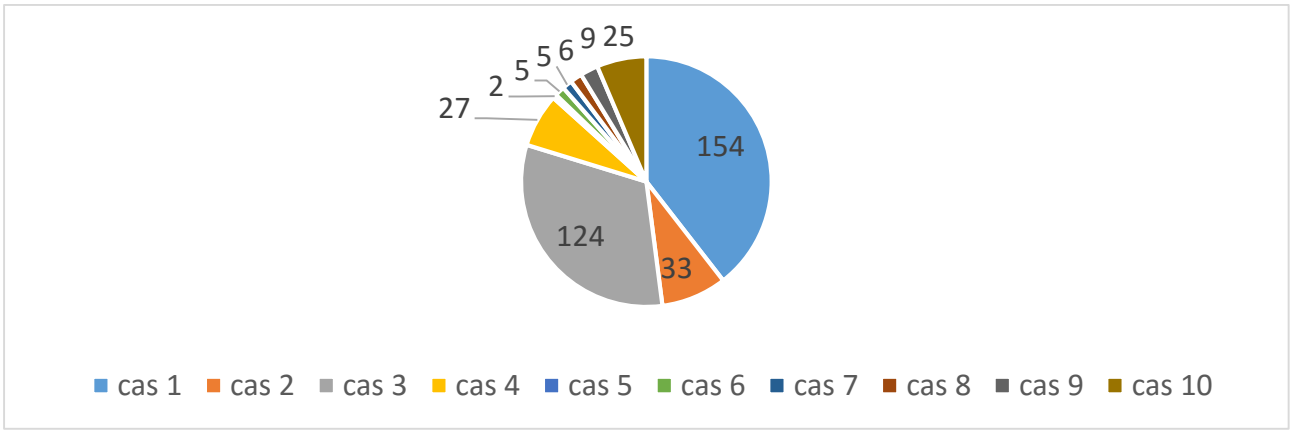
- Mais selon le sélectionneur Razvan Lucescu, il devrait être apte pour le match contre la France, samedi au Stade de France. C'est considéré comme une citation
- France Télécom a annoncé vendredi le rachat d'Unanimis, régie britannique spécialisée dans la publicité de marques sur internet, pour un montant qui n'a pas été précisé. Ceci n'est pas une citation.

Dans ce cas, la phrase ne peut pas être considérée comme une citation car elle ne laisse pas penser que la personne concernée ait pu dire quelque chose d'approchant, même paraphrasé. En d'autres termes, le doute est présent quand on essaye de savoir si l'auteur peut avoir dit cela ou non. Quand il y a ce doute, la phrase n'est pas considérée comme une citation.

Parmi les cas présentés ci-dessus, on peut retrouver les caractéristiques d'un cas dans un autre cas, comme les deux points (cas n°2) ou le fait qu'il peut y avoir plusieurs citations dans la même phrase (cas n°3). Mais les détails sur le nombre de citations possibles et les combinaisons possibles rendent chaque cas de citations unique.
Exemple cas 4 : selon/d'après/pour M. Smith « … », « … »
Exemple cas 6 : M. Smith a dit : « … » et ajouta que « … »

Et s'il reste une incertitude sur le cas à attribuer à une citation, c'est le plus élevé qui est choisi.

Répartition des citations :



La répartition s'est calculée à partir d'un échantillon de 99 documents comprenant en tout 390 citations.

2. PRESENTATION DE LA CARTOUCHE LIVING-QUOTES

Nous présentons dans cette partie, la cartouche Living-Quotes, solution développée par Temis pour extraire les citations.

Tout d'abord, une présentation de ce que la cartouche Living-Quotes fait :

- Elle exploite les annotations de la TM360, autrement dit les entités nommées
- Elle contient des règles qui reprennent la sortie de la TM360, soit des relations et des entités
- A partir des données de la TM360 et des règles, elle créé alors une relation « Citation » avec un attribut « Who » pour l'auteur
- Elle surligne la phrase entière qui contient la citation sans distinguer la citation à proprement dite.

Ce dernier point ne permet pas d'identifier immédiatement les propos cités. On suppose que la solution a été développée de cette façon pour minimiser les traitements et uniformiser les extractions et ainsi éviter de créer une sortie d'extraction propre aux citations avec guillemets et une autre pour celles sans guillemets. Cependant nous ne suivrons pas cette volonté, au contraire nous ferons en sorte que les éléments auteur et citation soient extraits de façon distincte.

Ensuite, ce que la cartouche Living-Quotes extrait :

- Les citations avec guillemets s'il y a un verbe déclencheur

Exemple :
A ce propos, le secrétaire américain à la Défense Ashton Carter a déclaré s'attendre à ce que la Russie "commence à subir des pertes" humaines "dans les prochains jours".

- Les citations sans guillemets s'il y a le déclencheur « selon »

Exemple :
Malgré la richesse du sous-sol en Guinée-parmi les premiers producteurs mondiaux de bauxite et riche en minerai de fer, or, diamant et pétrole-plus de la moitié de la population vit avec moins d'un euro par jour, selon l' ONU .

- Les citations en anglais, en français et en allemand

Enfin ce que la cartouche Living-Quotes n'extrait pas, autrement dit, les manques :

- Les anaphores

Lors de la comparaison avec l'autre outil d'extraction des citations, la moitié des citations

manquées par Temis était dû à une anaphore pronominale au niveau de l'auteur de la citation.

- Les citations incomplètes

Exemple :
"Alcatel se réjouit de cette étape supplémentaire dans sa coopération avec Thomson multimédia, qui établit un lien étroit entre deux des activités de cœur de métier de nos compagnies. Je n'ai pas de doute que cette opération va permettre à l'équipe modems d'accélérer son développement et à nos deux groupes d'apporter au marché les solutions de bout en bout qu'il attend pour évoluer vers la vidéo interactive " a déclaré Serge Tchuruk, Président-Directeur Général d' Alcatel

- Les citations sans guillemets et sans selon

Les seules citations sans guillemets que la cartouche extrait sont celles qui contiennent la préposition « selon ». Or ce ne sont pas les seules.

Pour finir cette présentation sur la cartouche Living-Quotes, voici une liste des problèmes identifiés lors de la validation :

- Verbes déclencheurs

En validant les résultats de la cartouche Living-Quotes, on cherche à savoir pourquoi certaines citations ne sont pas ou mal extraites. Pour cela, on modifie la phrase et on l'annote de nouveau afin d'identifier ce qui bloque. Et parfois il s'agissait du verbe déclencheur qui ne faisait pas partie de la liste. On identifiait ce problème en remplaçant le verbe en question par le verbe déclarer.

- La distance entre le verbe et l'auteur

On a constaté que certaines citations n'étaient pas extraites à cause de la distance entre ces deux éléments. En effet, comme présenté dans la typologie des citations, il peut y avoir des précisions, des compléments de lieux ou de temps entre l'auteur et le verbe déclencheur et les règles développées dans Living-Quotes ne prennent pas en compte ces situations. On a pu identifier ce problème en supprimant justement ces « précisions ».

- Les fausses citations

Certaines phrases ont été identifiées et extraites en tant que citations mais selon la typologie des citations, elles n'en sont pas.
Exemples :
Vivendi Universal annonce aujourd'hui son accord avec la société Houghton Mifflin, basée à Boston et l'un des leaders américains de l'édition scolaire, pour acquérir l'ensemble des actions formant le capital de Houghton Mifflin, à travers une offre d'achat à un prix de 60 dollars par action.
Alcatel annonce la fin des discussions en cours avec Lucent Technologies

- L'auteur de la citation

Il y a eu quelques erreurs au niveau du nom de la compagnie mais cela est dû à la TM360. Dans l'exemple ci-dessous, seul NTT a été extrait comme auteur.
Exemple :
Le test du service de téléphonie mobile de troisième génération, lancé cette semaine au Japon, a été atteint par un défaut de fonctionnement du service e-mail, qui s'est prolongé pour 18 heures, a déclaré l'opérateur NTT DoCoMo vendredi.

A la suite de cette analyse de la cartouche, les points sur lesquels il faut être attentif lors de l'implémentation des règles d'extraction sont les suivants :

- Les anaphores
- L'auteur
- La distance entre les éléments
- Les verbes déclencheurs
- La sortie finale

De manière générale, nous nous concentrerons sur l'amélioration des résultats rendus par la cartouche Living-Quotes en extrayant de la manière la plus rigoureuse possible, les citations décrites dans la typologie.

3. CORPUS

Nous verrons dans cette partie les corpus utilisés pour constituer la typologie.

Dans un premier temps, on a étudié les citations à travers les articles de GsmBox et ensuite avec les dépêches de l'AFP. Mais les quelques cas de citations identifiés dans le premier corpus ont été retrouvés dans le second.

En tout, 630 documents validés dont 81 articles de GsmBox et 550 dépêches de l'AFP.

Le corpus utilisé est un corpus de presse composé à partir des articles de l'AFP, GsmBox, Ouest-France et Le Monde, soit téléchargé depuis les sites de partage de fichiers de l'entreprise ou soit récupéré à partir des ressources utilisées pendant le master, c'est le cas pour le corpus de Le Monde 2014.

Le corpus utilisé pour définir la typologie sert par la suite de corpus de développement pour les règles d'extraction.

Tout le travail effectué s'est majoritairement appuyé sur des données textuelles de l'AFP, on peut donc supposer que les citations relevées sont très spécifiques aux dépêches de l'AFP et moins applicables sur un autre corpus de presse. L'AFP représente pour la presse française une référence (sûre et objective). En effet plusieurs journaux comme Libération reprennent les articles de l'AFP pour rédiger les leurs. Cependant l'idéal aurait été d'avoir un très large éventail de corpus de presse pour être sûr de couvrir toutes les citations existantes. Mais l'on notera que l'analyse de corpus s'est d'abord faite sur un corpus de GsmBox, un journal spécialisé dans les nouvelles technologies et que la première version de la typologie s'est construite dessus. Et lors de la seconde analyse sur l'AFP, la plupart des cas de citations précédemment observés ont été retrouvés en même temps que l'apparition de nouveaux cas. On peut donc dire que la typologie recouvre un panel assez large des citations.

4. ANNOTATION ET VALIDATION DANS AWB

On présente ci-dessous les critères qui nous ont permis d'évaluer les extractions de la cartouche Living-Quotes :

Validation	Situations
Correct	• Auteur et citation correctes
Partiellement correct	• Auteur correct mais citation incomplète • Citation complète mais auteur incorrect • Citation incomplète et auteur incorrect
Incorrect	• Fausses citations

Tableau 1 - Critères de validations des citations

Justifications :

- Partiellement incorrect : il m'a été conseillé de valider comme partiellement correcte les citations dont l'auteur est bien extrait mais dont la citation est incomplète car une partie de la citation a tout de même été relevée. De même pour les citations complètes mais dont l'auteur n'a pas été bien extrait, soit parce qu'un autre terme a été extrait, soit parce que l'auteur n'est pas complet car au moins un des deux éléments est identifié. Enfin les citations dont l'auteur remplit un des cas précédents et dont la citation est incomplète pour les mêmes raisons dites précédemment.
- Correct : Seules sont correctes les citations dont l'auteur est bon et la citation complète
- Incorrectes : Les citations qui sont notées en tant qu'incorrectes sont les cas de fausses citations.

Annotation manuelle :
Toutes les citations manquantes, c'est-à-dire qui n'ont pas été extraites par la cartouche Living-Quotes ont été annotées manuellement. En lisant les dépêches, toutes les citations correspondantes aux cas de la typologie ont été ajoutées et annotées en tant que citation, sans préciser l'attribut auteur « who ».

5. RESULTATS ET DISCUSSIONS

Suite à la validation et l'annotation manuelle des 659 dépêches de l'AFP 2009 contenant en tout 2037 citations, les résultats sont les suivants :

	Précision	Rappel	F-mesure
Strict	79%	15%	25%
Tolérant	86 %	16%	27%

Tableau 2 - Résultats de la cartouche Living-Quotes

La précision représente le pourcentage de bonnes extractions sur l'ensemble des extractions et le rappel, le pourcentage de bonnes extractions sur les extractions qui étaient attendues. La F-mesure donne la performance du système en utilisant les deux mesures, grâce à la formule :

$$F = 2.PR / P+R$$

On constate que les erreurs d'extraction sont peu fréquentes (86%) mais le nombre de citations manquées est très élevé (16%). L'accent est donc à mettre sur l'augmentation du rappel. Tout en gardant une précision équivalente.

Les résultats ont été présentés à l'agent commercial concerné et voici les informations qui en sont ressorties à la suite de la présentation :

- Améliorer l'extraction des citations en s'appuyant sur la typologie pour la future implémentation
- Résoudre les anaphores
- Extraire les citations sur plusieurs phrases
- Viser les chiffres suivants pour la nouvelle solution à mettre en place :
  - Précision : entre 80 et 90%
  - Rappel : environ 75%
- Sortie souhaitée :
  - la phrase complète
  - nom de l'auteur
  - « la citation »

Ce chapitre a permis de définir ce qu'est une citation et comment elle se présente. On a vu qu'il existait une dizaine de cas différents. On a aussi désormais des mesures de référence issues de l'ancienne solution d'extraction des citations et des points d'attention pour diriger/mener dans le bon sens l'implémentation de la nouvelle.

Comme expliqué dans la présentation du contexte dans le chapitre 1, le stage est partagé entre les deux technologies, Luxid et Cogito Studio et la phase suivante qui concerne l'implémentation de la nouvelle solution se fait avec Cogito Studio et non Luxid. La raison principale de ce changement d'outil est la situation de transition dans laquelle se trouve Expert System France. La plupart des projets en cours continuent à être élaborés avec les outils Luxid de Temis mais pour une meilleure fusion avec l'entreprise Expert System, Luxid sera associé à Cogito Studio dans un futur proche afin de tirer profit du meilleur des deux et les projets à venir seront développés dans le résultat de cette fusion. Le projet d'extraction des citations étant un nouveau projet partant de zéro, il a alors été choisi pour être l'un des premiers à être développé avec Cogito Studio afin d'acquérir des connaissances sur l'outil, d'avoir un cas d'usage concret pour aider au développement de la

transition du côté de la R&D et comparer les pour et les contre des deux outils.

4. Implémentation de la solution

Après avoir établi une typologie complète des citations et avoir identifié les principaux problèmes de la cartouche Living-Quotes, tous les éléments sont réunis pour développer une nouvelle solution d'extraction des citations dans Cogito Studio.
Dans cette partie, nous allons décrire la solution implémentée, les ressources utilisées, les règles développées, la méthodologie appliquée, les limites et l'évaluation.

1. COUVERTURE ET EXPLICATION DES REGLES

Un projet dédié appelé « ExtractionCitations » a été créé dans Cogito Studio pour écrire les règles d'extraction. Présenté plus haut, l'outil Cogito Studio repose sur l'analyse de son réseau sémantique Sensigrafo et de son désambiguïseur pour développer les règles. Ces règles sont écrites en langage E, un langage propriétaire d'Expert System spécifique à l'extraction d'informations.
L'environnement Cogito Studio est semblable au Skill Cartridge Builder (SCB) de Temis. Ils permettent tous les deux d'écrire les règles d'extractions. Cependant chacun des deux outils accepte une entrée différente : le SCB prend en entrée des documents au format lux, tandis que Cogito Studio, lui, traite les documents convertis au format Cogito Studio. Le langage propriétaire de Temis pour écrire les règles, le langage TSL, est composé de balises comme le XML entre lesquelles l'on développe les règles. Alors que le langage E de Cogito Studio est fait d'accolades comme le JSON.
Les règles écrites dans le SCB peuvent se faire à partir d'expressions régulières, de listes, de lexiques externes et d'éléments morphologiques. Les règles qui utilisent des éléments morphologiques s'appuient sur les analyses de Xelda, l'outil de Luxid qui réalise une tokénisation et une analyse morphologique du texte entré. A ce niveau-là, Cogito Studio permet aussi d'écrire des règles faites d'expressions régulières (PATTERN), de listes (LIST), d'éléments morphologiques (TYPE), de rôles syntaxiques (ROLE) et de mots-clés (KEYWORD). Mais il permet en plus d'exploiter les ressources sémantiques du Sensigrafo, avec la possibilité de choisir un concept précis (SYNCON) ou bien de prendre tous sa descendance (ANCESTOR).

Les règles qui sont développées s'appuient entièrement sur la typologie des citations.

La sortie idéale contient un auteur, une citation et la phrase dans sa totalité.

Organisation du projet « ExtractionCitations » dans Cogito Studio
Tout d'abord, une présentation de la manière dont le projet ExtractionCitations s'organise. Il est composé de plusieurs fichiers, tous présentés ci-dessous, qui une fois compilés ensemble forment

le module d'extraction, à condition qu'ils ne génèrent aucune erreur. Tous les fichiers, excepté les listes, ont l'extension .efe qui est l'extension pour les fichiers de règles d'extraction.

- Le fichier main :

Le fichier source principal du projet. Il permet de spécifier quels sont les fichiers à prendre en compte lors de la compilation. Tous les fichiers doivent être précédés du mot-clé IMPORT. Si un fichier est absent ou en commentaire dans le fichier main, il ne sera pas compilé.

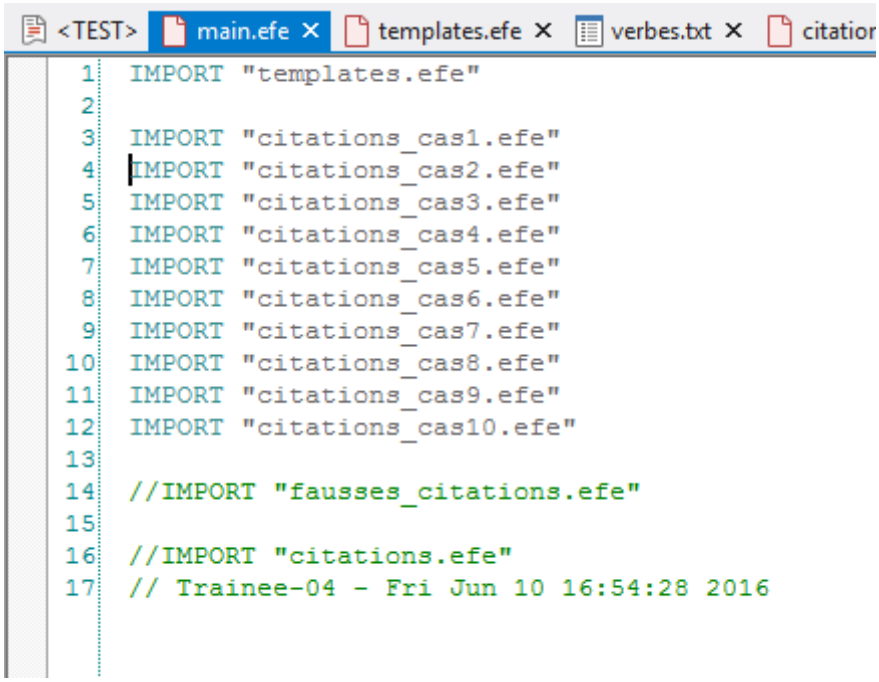


Figure 6 - Fichier main du projet ExtractionCitations

- Le fichier templates :

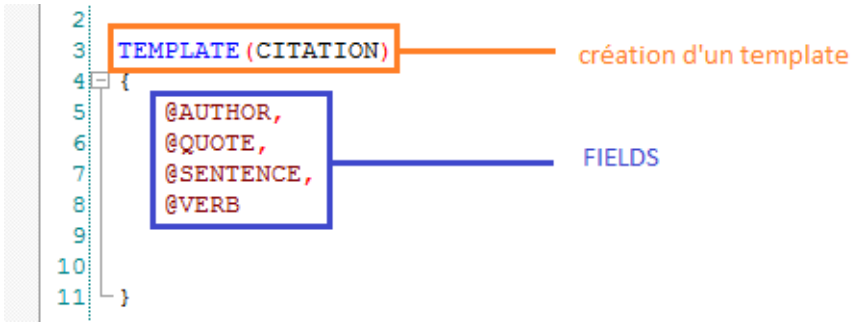


Figure 7 - Schéma d'un template

Un template est l'équivalent d'un tableau de base de données. Il constitue la première étape lors d'un projet d'extraction. Il définit les structures de données à remplir appelées les « champs ». En d'autres termes, il permet de lier une certaine séquence de valeurs pour représenter quelque

chose. Par exemple, on peut regrouper sous un template nommé « données personnelles », des champs comme « nom », « adresse » et « téléphone ». Ces trois champs représentent ensemble des informations personnelles.
On peut extraire plusieurs templates dans un même projet. Mais ici le but est d'extraire les citations donc un seul template nommé « CITATION » a été créé. En ce qui concerne les champs, ils représentent tout ce que l'on veut extraire dans une citation, soit l'auteur (@AUTHOR), la citation en elle-même (@QUOTE) et la phrase entière (@SENTENCE). On a aussi un champ pour le verbe déclencheur (@VERB) qui n'est pas exigé pour la sortie finale mais qui est un élément capital pour extraire une citation.
On peut voir sur la capture d'écran qu'il y a plusieurs champs pour l'auteur (@AUTHOR2), la citation (@QUOTE5) et le verbe (@VERB2). C'est dû à la typologie des citations qui dit que selon les cas définis il peut y avoir plusieurs citations et un ou deux auteurs et verbes déclencheurs. Les champs définis dans le template ne sont pas forcément tous présents dans les règles.

Ce fichier templates.efe permet donc de définir le template souhaité (CITATION) et les champs qui le composent (@AUTHOR, @QUOTE, …).
Il n'est pas nécessaire de créer un fichier à part pour définir un template, il peut tout à fait être placé dans les fichiers où sont écrites les règles. Mais pour ce projet, il a été déclaré séparément pour plus de lisibilité et une meilleure organisation.

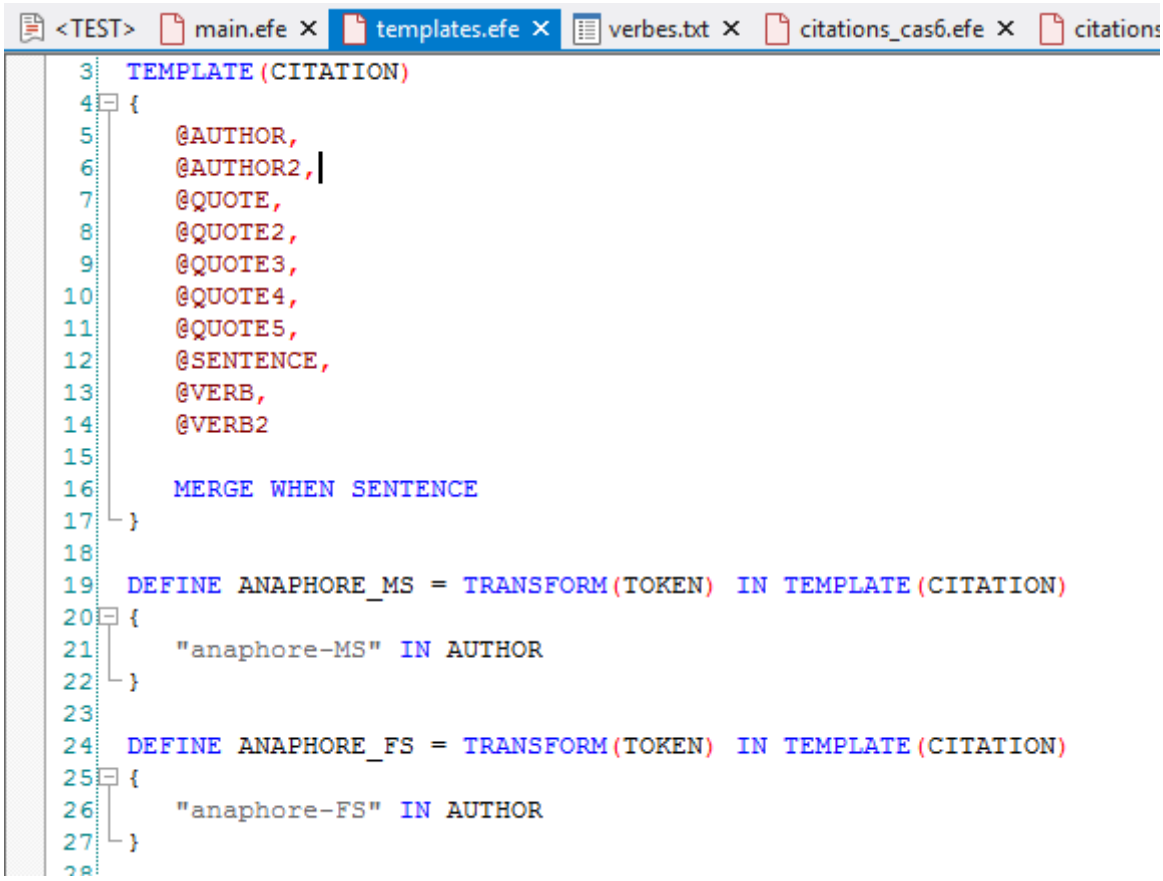


Figure 8 - template CITATION du projet

- Les fichiers de règles :

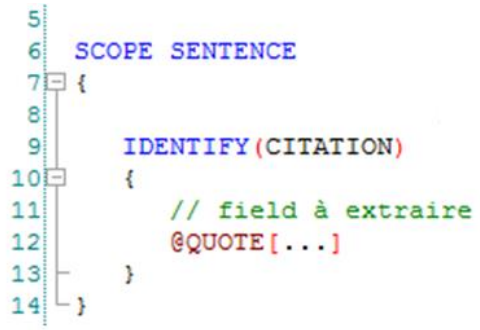


Figure 9 - Schéma d'une règle d'extraction

Ce sont les fichiers dans lesquels les règles sont implémentées. Pour chaque cas de citations, un fichier de règles est créé pour plus de lisibilité car :

- les différentes combinaisons possibles entre les trois éléments sont nombreuses
- les règles n'ont pas de priorité
- et elles doivent être mutuellement exclusives

Chaque fichier de règles contient entre 22 et 88 règles.
Ces règles écrites permettent de spécifier la nature du ou des champs à reconnaître dans un contexte délimité, ici la phrase.
Les éléments qui peuvent composer ces règles sont présentés dans le tableau ci-dessous :

Attribut	Valeur	Description
KEYWORD	String	Identifies a token from the exact sequence of characters to be found in a text
LEMMA	String	Identifies a token from the base form of a word contained in Sensigrafo
SYNCON	Syncon ID	Identifies a concept from its syncon ID
ANCESTOR	Syncon ID	Identifies a chain of concepts deriving from a syncon ID
LIST	List	Identifies a token by specifying the numerical ID of a syncon contained in Sensigrafo and considering the syncon itself as a bare container of lemmas
TYPE	Type	Identifies a token considering its word class
PATTERN	Regular expression	Identifies a token using a regular expression
ROLE	Role	Identifies a token considering its role in the clause
POSITION	Position	Identifies a token considering its position in the text
RELEVANT	List	Identifies a token by verifying if it is a relevant element in the text

Tableau 3 - Attributs pour les règles d'extraction

Des opérateurs permettent de combiner ces différents attributs. Il en existe trois types :

- Opérateurs de position

Ils permettent de spécifier l'ordre d'apparition des éléments dans une règle.

- Opérateurs logiques

Ils exigent que les deux éléments concernés par l'opérateur soient liés syntaxiquement.

- Opérateurs booléens

Ils prennent en compte la présence ou non de plusieurs tokens dans le texte.

Pour les besoins du projet, les éléments suivants sont utilisés :

- Expression régulière pour l'identification de la citation proprement dite, soit la section de texte entre guillemets
- Liste de lemmes correspondant à des verbes introducteurs de discours
- Les parties du discours pour l'identification des auteurs de citation
- Les mots-clés pour une liste courte de mots à trouver tel quel
- Les rôles pour contraindre l'extraction des auteurs à ceux qui sont sujets
- Les ancestor pour exclure quelques concepts lors de l'extraction des auteurs
- Les opérateurs de position pour couvrir toutes les combinaisons possibles

Figure 10 - Une règle pour les citations de cas 1

- Les listes de verbes :

Les listes permettent d'externaliser un lexique quand il est large. Elles sont au format txt et doivent contenir un mot par ligne.

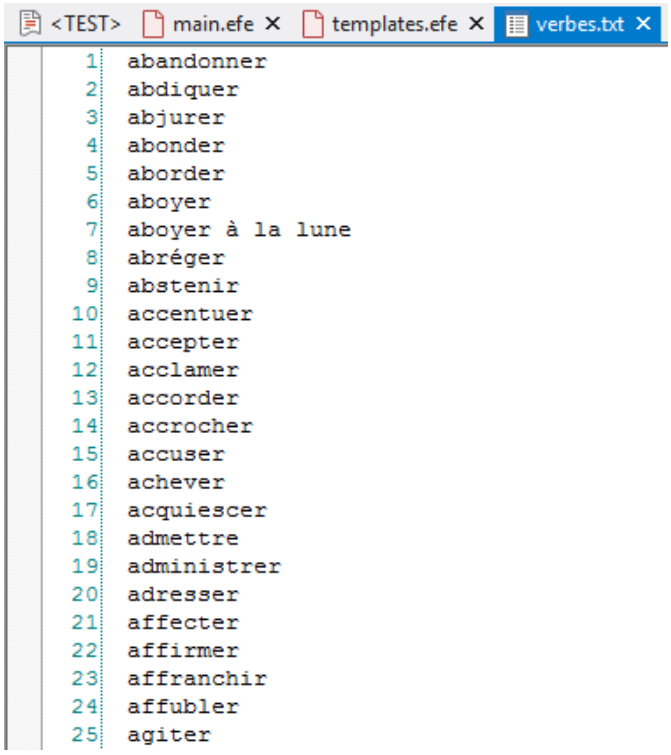


Figure 11 - Liste des verbes déclencheurs

Pour résumer, le projet se compose de :

- un fichier main
- un fichier templates
- 10 fichiers de règles

On présente ensuite la méthodologie suivie pour écrire les règles d'extraction.

Ce que l'on souhaite reconnaître : la citation et son auteur.
On a appliqué la méthodologie présentée ci-dessous de la même manière pour chaque cas de citations.
Cette méthodologie est constituée des étapes suivantes :

1. Création d'un fichier d'échantillon de citations appartenant au même cas
2. Modélisation des phrases
3. Écriture des règles. Test de développement sur le fichier échantillon. Modification des règles et tests jusqu'à des résultats optimaux
4. Tests supplémentaires sur des dépêches entières de l'AFP prises au hasard dans le corpus complet de dépêches converties au format Cogito Studio. Relève des erreurs à corriger. Modification des règles si ces erreurs sont récurrentes

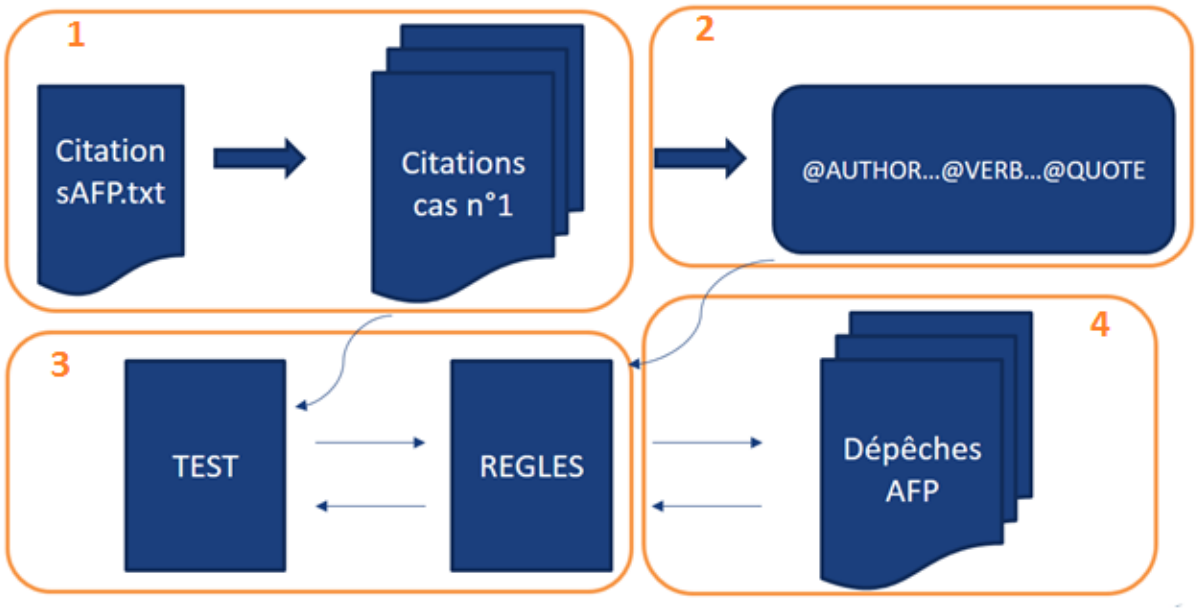


Figure 12 - Schéma de la méthodologie suivie pour écrire les règles

Détails des étapes :

1. Mise en œuvre

Les citations rencontrées dans les 630 documents ont été relevées et rassemblées dans un fichier (CitationsAFP.txt). Ce fichier contient environ 1000 citations. Il a permis de constituer 10 fichiers, qui correspondent aux 10 cas de citations et qui regroupent chacun un échantillon allant de 10 à 30 citations par cas (TestUnitaireCas1.txt). Ils ont été utilisés pour développer et tester les règles. Pour s'assurer d'aucun effet de bord à l'implémentation d'une nouvelle règle, les autres échantillons de citations sont laissés dans le panneau de test et les règles des cas déjà implémentés sont également compilées.

2. Modélisation des phrases

Pour chaque cas, l'échantillon de citations correspondant a été analysé et modélisé pour définir au mieux les règles d'extraction. A partir de chaque échantillon, qui contient entre 10 et 30 exemples variés pour chaque type de citations, on modélise chacune des phrases pour avoir un aperçu de l'agencement de chaque élément et avoir une idée des distances entre les éléments afin d'extraire un maximum de citations.

Pour illustrer cette étape, voici trois exemples de citation avec leurs modélisations en dessous :

Dans un discours télévisé, le chef de gouvernement du Hamas Ismaïl Haniyeh a également assuré que "le peuple palestinien vaincra les chars" israéliens en cas d'incursion terrestre.

@AUTHOR <0 :2> @VERB <0 :1> @QUOTE

"Il faut de l'altruisme, il faut de la générosité", il est "temps de dialoguer et de parler", a-t-elle poursuivi, ajoutant : "il y a des membres du PS et des écolos qui veulent que ça bouge, ils ont raison, je suis avec eux".

@QUOTE <0 :3> @QUOTE2 @AUTHOR @VERB @VERB2 @QUOTE3

Lundi, un éditorial publié par le quotidien chinois Global Times, tout en prenant soin de ne pas nommer la Birmanie, s'inquiétait des "gouvernements impopulaires" dans les pays frontaliers de la Chine.

@AUTHOR <0 :13> @VERB <0 :1> @QUOTE

2. Ecriture des règles

On décrit dans cette partie les règles implémentées dans Cogito Studio.

Pour les citations avec guillemets, les trois éléments, auteur, citation et déclencheur, doivent être présents, tandis que pour les citations sans guillemets, il y a seulement le déclencheur et l'auteur.
Lors de l'écriture des règles d'extraction, on privilégie le bruit au silence, autrement dit, on préfère avoir plus de mauvaises extractions que de manquer des extractions.
L'écriture des règles se fait à force de tests sur l'échantillon et de modifications des règles.

A la fin, on obtient des règles qui peuvent s'appuient sur les éléments suivants :

CITATIONS

principes

Les citations avec guillemets sont extraites grâce à une expression régulière qui accepte trois types de guillemets différents :

- "..."
- +...+
- « ... »

Pour les citations sans guillemets, on extrait la phrase complète avec une expression régulière.

limites

Les règles d'extraction des citations avec guillemets extraient aussi des guillemets qui ne sont pas délimiteurs d'une citation et cela crée du bruit.

AUTEURS

principes

Pour extraire les auteurs des citations, les règles dépendent de l'analyse sémantique qui a été faite en amont car elles s'appuient essentiellement sur les parties du discours.

Pour les citations avec guillemets, plusieurs types d'auteur sont à prendre en compte. Les règles implémentées dans Cogito Studio permettent d'extraire toutes les citations dont l'auteur est :

- Un nom propre, excepté les lieux et les journaux[1]
- Un nom propre précédé d'un titre (M., Mme)
- Un nom composé
- Un pronom personnel (il, elle)
- La combinaison d'un adjectif et d'un nom commun
- Un nom commun comme communiqué, source ou étude

Pour les citations sans guillemets, les auteurs potentiels sont plus limités car il y a plus de risques de générer du bruit.

En effet dans une citation sans guillemets, il n'y a plus que 2 éléments : l'auteur et le déclencheur. La règle étant moins précise, il y a donc plus de risques d'extraire du bruit. C'est pourquoi, il est nécessaire de resserrer les deux autres éléments qui constituent la citation.
Les auteurs potentiels s'arrêtent aux entités nommées de type :

- Personne physique
- Compagnie
- Organisation

Si une phrase contient un déclencheur mais que l'auteur n'est pas une personne identifiable et distincte, la citation n'est pas extraite car elle ne pourra pas être rattachée à quelqu'un en particulier.

limites

Avec toutes les règles développées pour couvrir les différents types d'auteur, il y a parfois plus d'un auteur pour la même citation.

[1] Les éléments surlignés en jaune ont été rajoutés plus tard, après l'évaluation intermédiaire des règles

DECLENCHEURS

principes

Le déclencheur correspond à un verbe ou une préposition qui introduit la citation. Avec l'auteur, ils constituent les deux éléments nécessaires à l'extraction des citations.

Pour les citations avec guillemets, les verbes déclencheurs sont regroupés dans une liste et extraits sous la forme de lemmes. On peut donc extraire le verbe sous toutes ses formes. Ils sont essentiellement issus du concept « verbes de communication linguistique » du Sensigrafo et ont été relevés manuellement en parcourant la descendance complète du concept. Ce concept « verbes de communication linguistique » a été choisi car il est le concept qui contient le plus de verbes déclencheurs recherchés. Les règles développées prennent compte la présence ou non de l'auxiliaire « avoir » ou « être » devant le verbe déclencheur, ce qui double le nombre de règles.

Pour les citations sans guillemets, la liste de verbes est plus restreinte car une citation sans guillemets est plus risquée à extraire qu'une citation avec guillemets. La liste des verbes choisis est dans la typologie des citations.

Les prépositions sont :

- Selon
- Pour
- D'après
- Chez

Si une phrase contient une citation et un auteur potentiel mais aucun déclencheur, la citation n'est pas extraite. Le déclencheur est le lien entre l'auteur et la citation.

limites

La préposition « pour » génère beaucoup de bruit car elle est très utilisée en dehors des citations.

Pour lier ces trois éléments dans une règle d'extraction, on utilise les opérateurs de position évoqués plus haut. Ils autorisent la présence d'un ou plusieurs tokens entre les éléments. Voici un tableau qui résume la fonction de ces opérateurs :

Operator	Name	Description
>>	Strict sequence	Elements must be in the sequence specified in the rule and no token is allowed between them
>	Loose sequence	Elements must be in the sequence specified in the rule and only elements with low semantic value are allowed (adjectives, adverbs, conjunctions, articles) between them

<>	Flexible sequence	Elements must be in the sequence specified in the rule and any token is accepted between them
<	Loose sequence with right reference	Equivalent to > except in the presence of a negated attribute
<<	Loose sequence with right reference	Equivalent to >> except in the presence of a negated attribute

Tableau 4 - Opérateurs de position

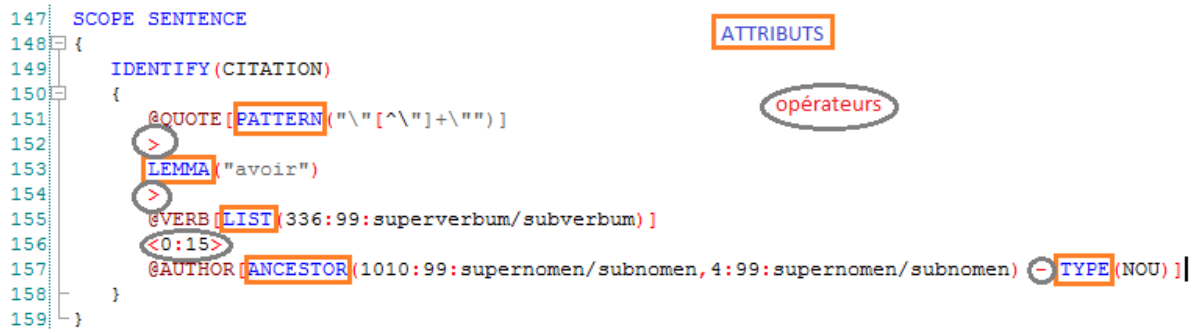


Figure 13- Exemple de règle avec l'utilisation des opérateurs de position

4. Tests supplémentaires

Une fois que les règles extraient parfaitement l'échantillon de citations, elles sont testés sur des dépêches entières prises au hasard dans le corpus de l'AFP 2015. Les dépêches ont été converties au format Cogito Studio. En plus de tester les règles implémentées sur d'autres citations que celles de l'échantillon, ces tests permettent également de voir si les règles extraient bien des citations et les bonnes dans un document entier.

En testant ainsi les règles sur plusieurs dépêches AFP, on note les erreurs et si elles sont récurrentes on modifie les règles.

Des règles mutuellement exclusives, une contrainte importante
Chaque cas prend uniquement en compte ce qu'il couvre et exclut le reste des possibilités. Cela est nécessaire sinon les cas ne sont pas exclusifs et les extractions peuvent contenir des doublons. Pour tous les cas, les développements ont été faits de telle sorte qu'il n'y ait pas d'extraction d'une même citation par deux ou plusieurs règles différentes. Pour ce faire, le cas n°1 et n°2, par exemple, excluent qu'il y ait plus d'une citation. Cette obligation pour les règles d'être mutuellement exclusives est dues aux contraintes de Cogito Studio présentées par la suite.
Par exemple, on souhaite extraire la citation suivante :

"Je suis sûre qu'elles s'accommoderont de la nouvelle situation politique", a cependant prédit Mme Hama.

Cependant :

- Si j'ai une règle qui extrait les citations dont l'auteur est un nom propre humain (NPH) , l'auteur de cette citation sera « Hama »
- Si j'ai écrit une autre règle qui extrait les citations dont l'auteur est « Mme,M. » suivi d'un nom propre, cette citation aura aussi pour auteur « Mme Hama »
- Enfin si j'ai une dernière règle qui extrait les citations dont l'auteur est un nom, alors cette citation aura un auteur équivalent à « Mme »

Au final, trois auteurs sont extraits pour la même citation alors qu'une seule extraction est choisie pour la sortie finale et le hasard ne prend pas forcément la bonne.

En écrivant les règles, il faut donc s'assurer qu'elles sont exclusives les unes par rapport aux autres.

2. RESSOURCES UTILISEES

Les ressources utilisées pour extraire les citations sont :

- L'outil Cogito Studio
- Le corpus de l'AFP de l'année 2009
- Le corpus de l'AFP de l'année 2015

Pour écrire les règles d'extraction
Dans Cogito Studio, on a utilisé :

- le Sensigrafo, réseau sémantique multilingue
- des fichiers sources pour l'extraction (extension .efe) : citations_cas1.efe, citation_cas2.efe, etc…
- des listes (verbes.txt, author_nou.txt)
- le Wiki d'Expert System, documentation en ligne

Pour les tests de développements
Le corpus de l'AFP de l'année 2009 est le corpus utilisé pour valider la cartouche Living-Quotes. Il contient 659 dépêches et 2037 citations. Mais il a également servi de corpus de développement pour les règles d'extraction puisqu'à partir des citations rencontrées lors de cette validation, on a chargé un fichier appelé CitationsAFP.txt des 1000 premières citations, qui lui-même a servi à créer les fichiers d'échantillon de citations pour chaque cas. Ainsi on a 10 fichiers, un pour chaque cas de la typologie des citations, qui contiennent entre 10 et 30 citations sur lesquels on s'appuie pour développer et tester les règles d'extraction.

Pour faire les tests supplémentaires
Le corpus de l'AFP contient toutes les dépêches de l'année 2015, soit en tout 273 315 dépêches. Toutes les dépêches ont été convertis au format lux et Cogito Studio pour pouvoir les utiliser dans les deux outils. Ce corpus est beaucoup trop important. Par conséquent, on l'a découpé en un

sous-corpus composé de seulement 500 dépêches pour faciliter l'accès aux dépêches et ainsi aux tests. C'est à partir de ce sous-corpus que l'on prend au hasard les dépêches qui servent à faire les tests supplémentaires expliqués précédemment.

3. AVANTAGES ET LIMITES DE L'OUTIL ET DE LA SOLUTION IMPLEMENTEE

On liste ici les différentes contraintes de Cogito Studio.

REPETITION ET EXCLUSION

Les trois points suivants contraignent l'écriture des règles avec Cogito Studio. En conséquence, les règles doivent être le plus précises possibles et mutuellement exclusives afin que chacune soit unique et qu'il n'ait pas de doubles extractions sur une même phrase.

modularité

L'outil Cogito Studio ne permet pas de réutiliser une règle déjà implémentée. Si l'on souhaite reprendre une règle existante et la compléter, il faut la dupliquer.

optionalité

Dans Cogito Studio, il est n'est pas possible de mettre en option les différents éléments d'une règle. S'il y a plusieurs possibilités dans une même règle, il faut réécrire cette règle autant de fois qu'il y a de possibilités.
Dans l'exemple ci-dessous, les deux règles sont identiques excepté la partie concernant la présence du lemme « être » ou « avoir » : la première exige un des deux auxiliaires tandis que l'autre impose qu'il n'y en ait pas (! = négation).

```
// rule 1
//"L'Europe recevra tout le gaz que lui fournira la Russie", a affirmé M. Sokolovski.
SCOPE SENTENCE
{
   IDENTIFY(CITATION)
   {
      !PATTERN("\"[^\"]+\"" , "(\«[^\»]+\»)" , "\+[^\+]+\+")
      <0:15>
      @QUOTE[PATTERN("\"[^\"]+\"" , "(\«[^\»]+\»)" , "\+[^\+]+\+")]
      >
      LEMMA("avoir","être")
      >
      @VERB[LEMMA(LOAD "verbes.txt")]
      <0:15>
      !KEYWORD("M.","Mme","Me","Pr","Dr")
      <0:15>
      @AUTHOR[TYPE(NPH) - ANCESTOR(123109,119,687)]
      <0:15>
      !PATTERN("\"[^\"]+\"" , "(\«[^\»]+\»)" , "\+[^\+]+\+")
   }
}

// rule 2
//"Martine Aubry a repris la barre, voire fixé un cap", se félicite Paul Quinio dans Libération.
SCOPE SENTENCE
{
   IDENTIFY(CITATION)
   {
      !PATTERN("\"[^\"]+\"" , "(\«[^\»]+\»)" , "\+[^\+]+\+")
      <0:15>
      @QUOTE[PATTERN("\"[^\"]+\"" , "(\«[^\»]+\»)" , "\+[^\+]+\+")]
      >
      !LEMMA("avoir","être")
      >
      @VERB[LEMMA(LOAD "verbes.txt")]
      <0:15>
      !KEYWORD("M.","Mme","Me","Pr","Dr")
      <0:15>
      @AUTHOR[TYPE(NPH) - ANCESTOR(123109,119,687)]
      <0:15>
      !PATTERN("\"[^\"]+\"" , "(\«[^\»]+\»)" , "\+[^\+]+\+")
   }
}
```

Figure 14 - Deux règles presque identiques

priorité des règles

Toutes les règles écrites dans un projet Cogito Studio sont sur un seul et même niveau, il n'y a pas de priorité entre elles. Si deux règles extraient une information différente sur une même phrase, là où il est attendu qu'une seule bonne extraction, il sera impossible de dire laquelle on priorise. Le système aura alors plusieurs choix et en prendra un au hasard. Il est donc très important d'écrire des règles mutuellement exclusives et très précises pour qu'aucune ne se croisent.

ANALYSES ET DESAMBIGUISATION

Les points évoqués ci-dessous concernent, les analyses du désambigüiseur sémantique de Cogito. Elles sont énumérées dans leur ordre d'application. Pour traiter le français, elles peuvent constituer des obstacles au développement des règles dans Cogito Studio quand le taux d'erreur est trop important. Si une erreur est générée lors d'une des analyses, elle se répercute et est aggravée dans

les analyses suivantes.
On présente ci-dessous des schémas pour comprendre où et comment intervient le désambiguïseur de Cogito.

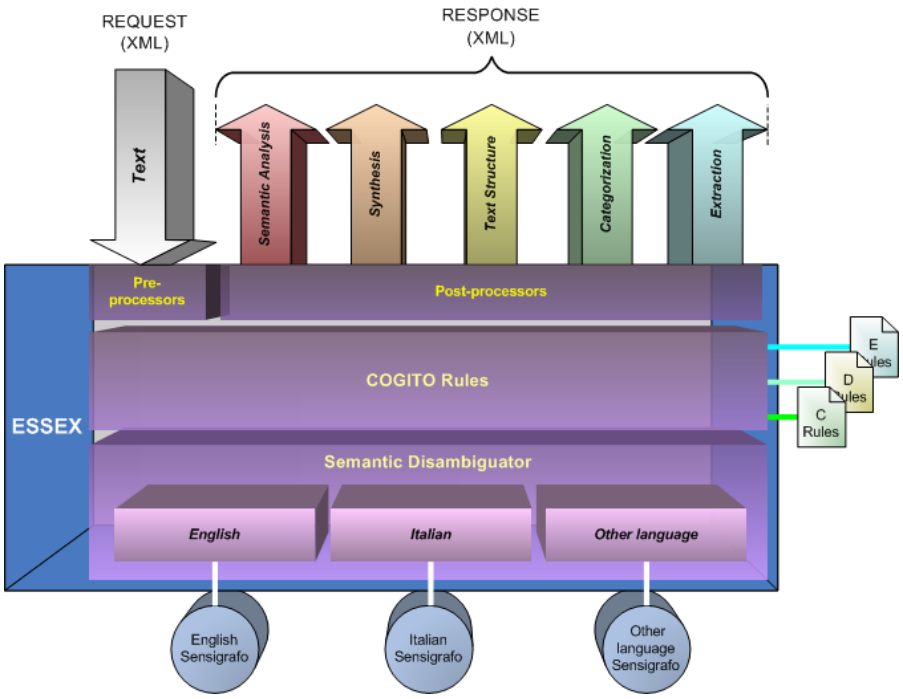


Figure 15 -Fonctionnement de Cogito Studio et ESSEX

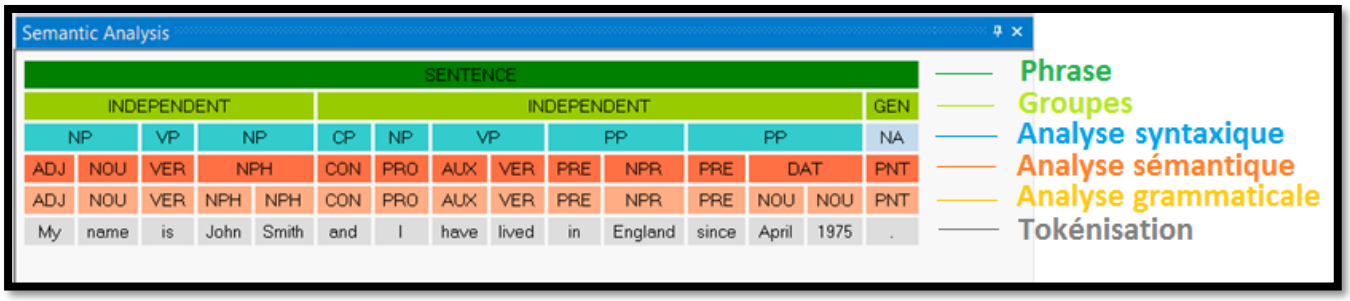


Figure 16 - Analyse sémantique complète du désambiguïseur de Cogito Studio

analyse lexicale

Cette première étape permet de découper le texte en tokens. Pour cela, le texte est d'abord nettoyé et normalisé. La sortie qui en résulte sera utilisée dans la phase suivante.
En général la tokénisation est correcte.

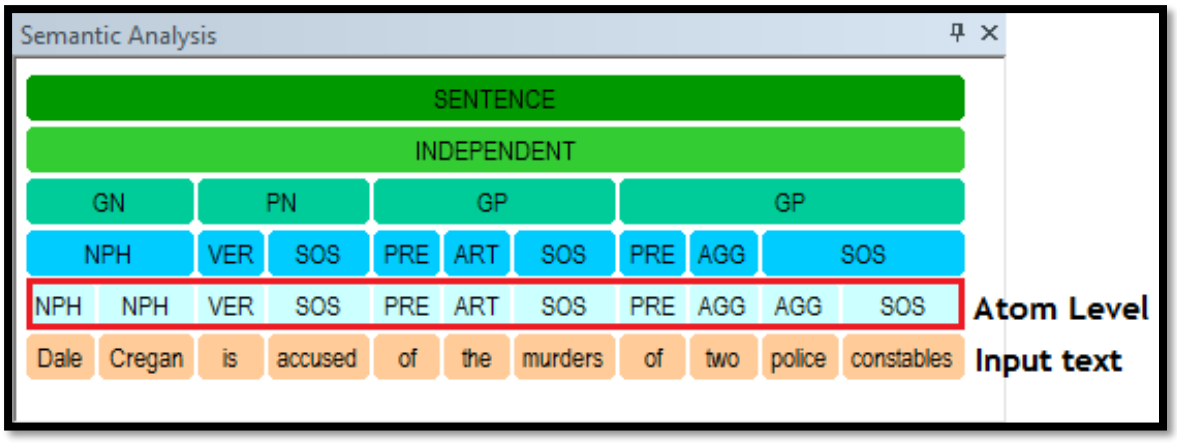


Figure 17 - Analyse lexicale

analyse grammaticale

Lors de cette étape, on attribue une partie du discours à chaque token. Cette phase identifie les tokens connus et inconnus du réseau sémantique. Elle regroupe les tokens s'il y a des collocations, des mots-composés, des expressions idiomatiques, etc... Enfin elle associe à chaque mot ou groupe de mots connus leur lemme. Le résultat de cette analyse est un ensemble d'éléments grammaticalement et lexicalement identifiés et classifiés.

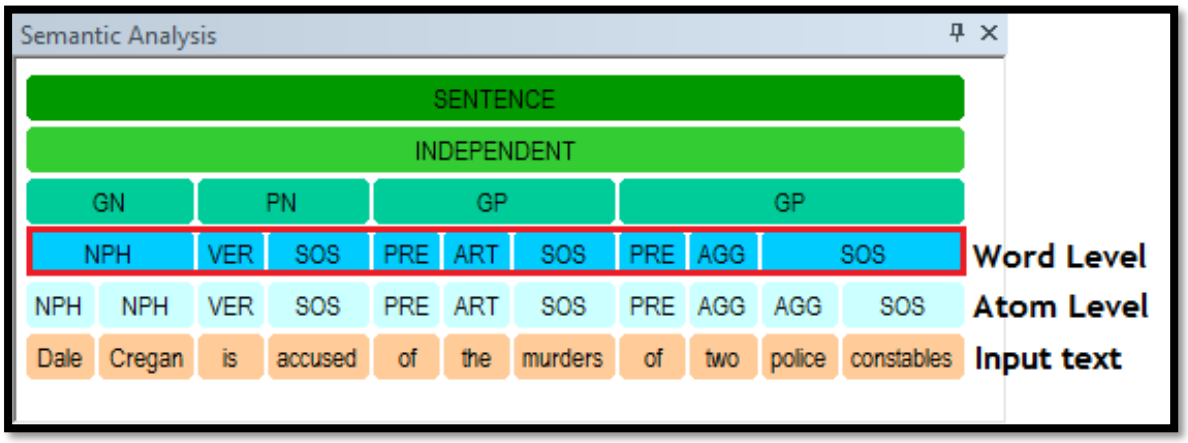


Figure 18 - Analyse grammaticale

Cette phase peut parfois générer des erreurs. Le plus souvent c'est une autre catégorie grammaticale qui a été associée au mot.

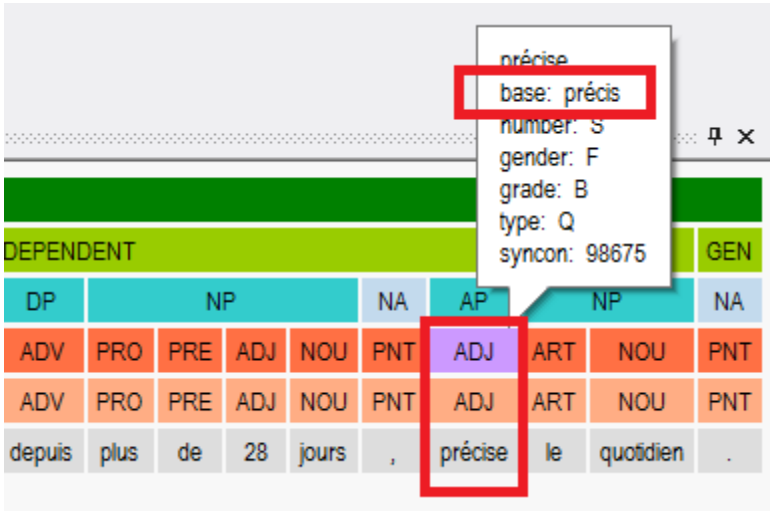


Figure 19 - Exemple d'erreur de désambiguïsation

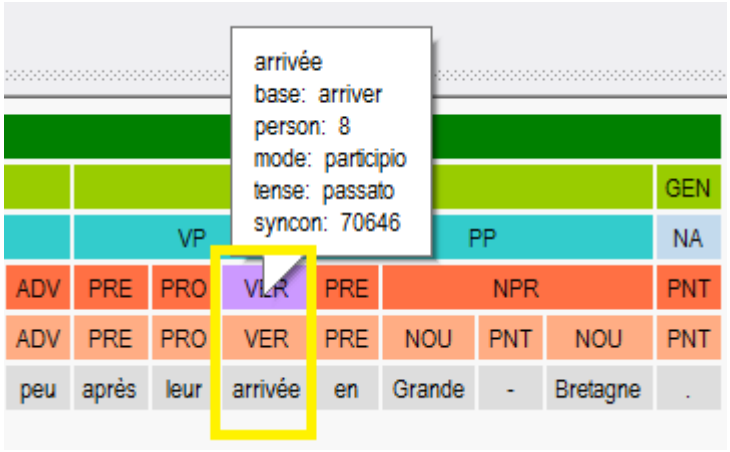


Figure 20 - Exemple d'erreur de désambiguïsation 2

analyse syntaxique

L'analyse syntaxique regroupe les mots par groupes (nominal, verbal), puis par propositions et finalement par phrases. Elle attribue un rôle à chaque groupe (sujet, objet) et indique les relations entre les verbes et leurs sujet, les verbes et leurs objets, etc…

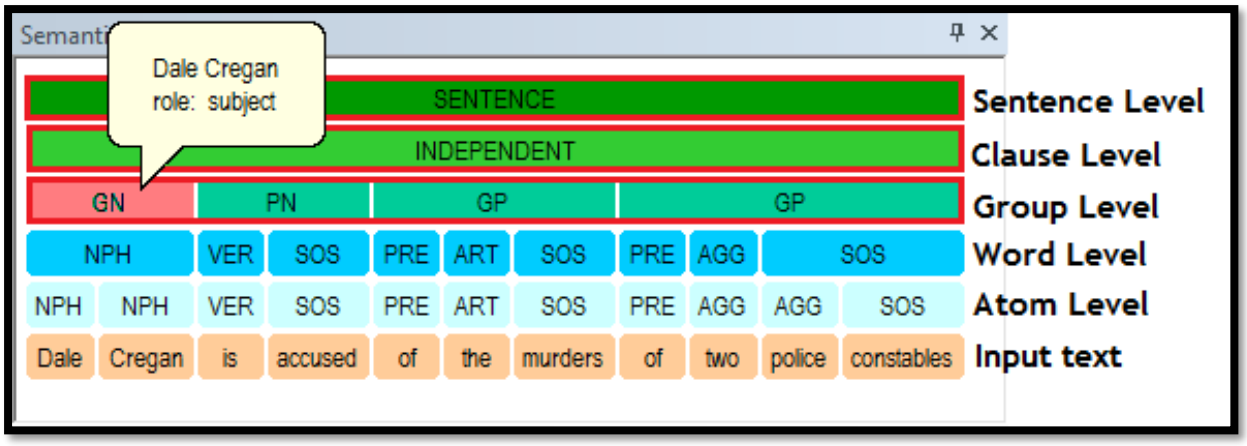


Figure 21 - Analyse syntaxique

Les rôles ne sont pas toujours bien détectés :

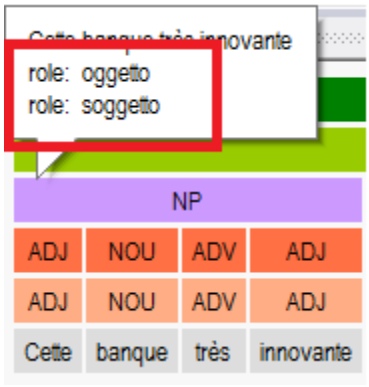


Figure 22 - Exemple d'erreur d'attribution des rôles

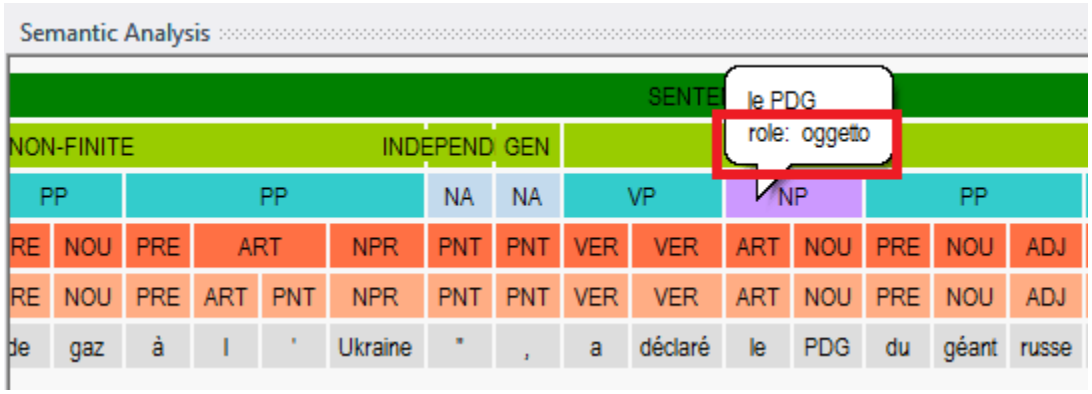


Figure 23 - Exemple d'erreur d'attribution des rôles 2

analyse sémantique

Sensigrafo est le réseau sémantique multilingue de Cogito Studio. C'est une vaste base dictionnairique directement lié et accessible depuis l'outil. Il est composé de milliers de concepts liés entre eux par des relations d'hyperonymie et d'hyponymie. Le taux d'erreur pour attribuer à un mot le bon concept est de 50%, autrement dit, un mot sur deux est mal désambiguïsé.
Pour la dernière phase, chaque token reconnu lors de l'analyse grammaticale est associé à un concept du Sensigrafo. Chaque token est associé à plusieurs concepts. Mais après la prise en compte de plusieurs critères, un seul reste lié au token. Ces critères de sélection sont :

- la fréquence d'utilisation
- le domaine
- les attributs
- la cohérence sémantique

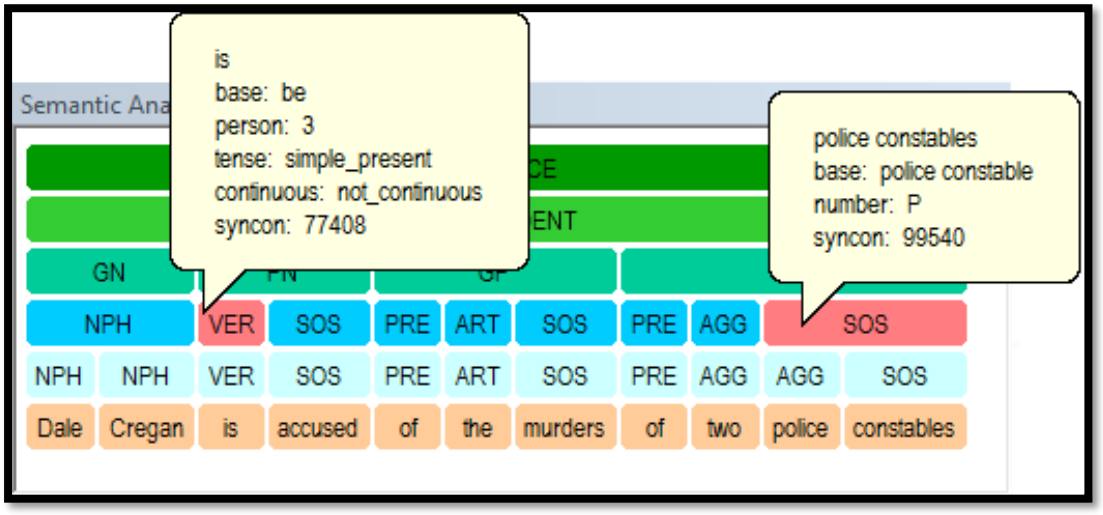


Figure 24 - Analyse sémantique

Cependant, la qualité de l'analyse sémantique pour le français est assez basse et un mot sur deux se voit attribué le mauvais concept.

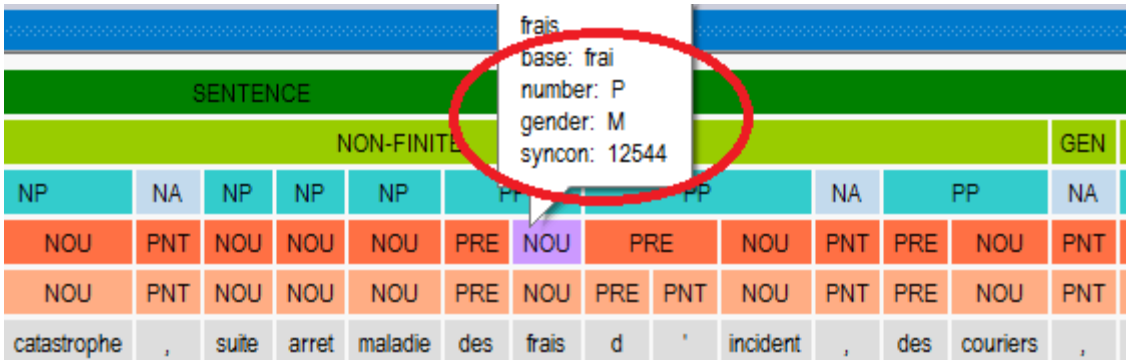


Figure 25 - Exemple d'erreur d'attribution de syncon

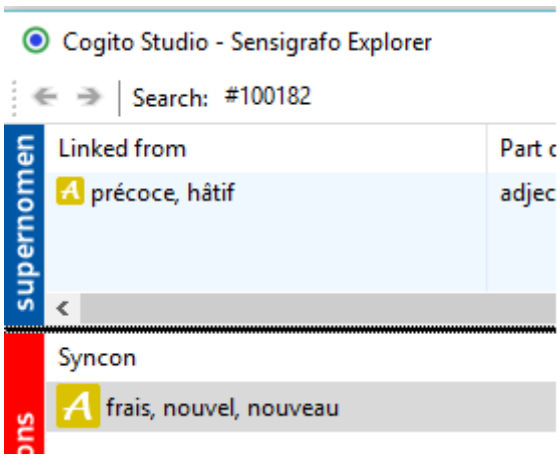


Figure 26 - Définition du syncon choisi à tort

Les difficultés rencontrées

Hormis les différents obstacles générés par les contraintes de Cogito Studio, les principales difficultés rencontrées concernent l'auteur de la citation. Le nombre d'auteurs possibles d'une citation augmentait au fur et à mesure des analyses et des tests effectués (la ministre danoise, un importateur vietnamien, la jeune fille, le Premier ministre).

4. ÉVALUATION DES REGLES IMPLEMENTEES

Cogito Studio possède un outil d'évaluation, le Test Bench, mais actuellement il n'est disponible que pour évaluer les projets de catégorisation. Donc le seul moyen d'évaluer la qualité est l'AWB de Temis. Pour évaluer les règles Cogito Studio dans AWB, il a fallu développer une passerelle ESSEXBridge qui permet d'exploiter, de visualiser et de valider les résultats de Cogito Studio depuis l'environnement Luxid et donc AWB. C'est l'équipe de la R&D qui a développé cette passerelle, notamment en s'aidant des deux seuls projets d'extraction développés dans Cogito Studio et disponibles à Expert System France : le projet d'extraction des citations et celui d'Opinion Mining. Cette passerelle transforme les sorties Cogito Studio en format lux, ce qui rend les sorties de Cogito Studio exploitables dans les outils de Luxid.

Fonctionnement de la ESSEXBridge :

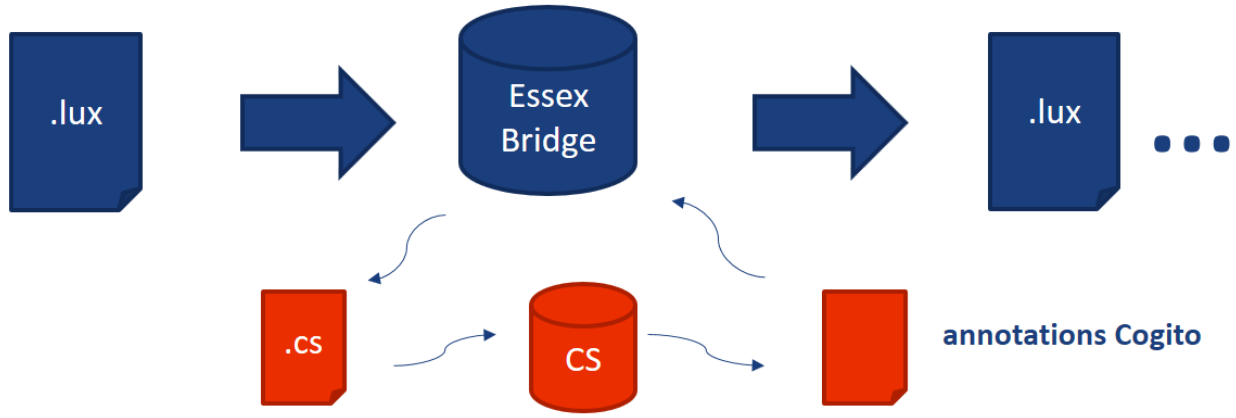


Figure 27 - Fonctionnement de la cartouche ESSEXBridge

La ESSEXBridge est intégrée dans le plan d'annotation du projet et contient le LPK (Linguistic PacKage), le fichier compilé et compressé du module d'extraction des citations de Cogito Studio.

L'évaluation des règles implémentées s'est faite sur un échantillon de 99 dépêches de l'AFP 2015 contenant 390 citations. C'est le même corpus qui a servi à calculer la répartition des types de citations dans la partie typologie. Le corpus utilisé ici, n'est pas exactement le même qui a servi à évaluer la cartouche Living-Quotes. Ce dernier était un corpus de l'AFP datant de 2009 tandis que celui-ci date de 2015. De plus, il n'a que 99 documents alors que l'autre en contenait 659.

Le mode de validation reste le même que celui pour évaluer la cartouche Living-Quotes pour pouvoir comparer les résultats des deux solutions.

Voici les résultats du module ExtractionCitations de Cogito Studio :

	Précision	Rappel	F-mesure
Strict	41%	17%	24%
Tolérant	84 %	35%	50%

Tableau 5 - Résultats intermédiaires du projet ExtractionCitations

Pour rappel, la cartouche Living-Quotes avait une précision de 86%, un rappel de 16% et une f-mesure de 27%.
On constate une nette amélioration du rappel comme prévu tout en ayant gardé une bonne précision. Cependant le rappel reste peu élevé. En effet, 255 citations n'ont pas été extraites. Pour comprendre pourquoi, une analyse a été faite sur toutes les citations validées en partiellement correcte et en incorrecte ainsi que celles manquées. Elle recense toutes les erreurs et les manques du module, leurs causes et sur quels éléments se trouve le problème (auteur, déclencheur, citation) en incluant les anaphores et les coréférences.
Cette analyse nécessite une vérification dans Cogito Studio pour chaque citation.

En conclusion de cette analyse :

Onze raisons dont les deux plus importantes se situent sur l'auteur :
- l'extraction du nom de l'auteur : 25%

- les anaphores (pronominales et implicites) : respectivement 23% et 10%

Les deux autres raisons suivantes concernent les citations :
- les citations sans guillemets : 13%
- le nombre de citations dans une phrase : 13%

Les autres raisons moins importantes concernent :
- les fausses citations : 8%
- les multi-phrases : 6%
- les coréférences : 5%

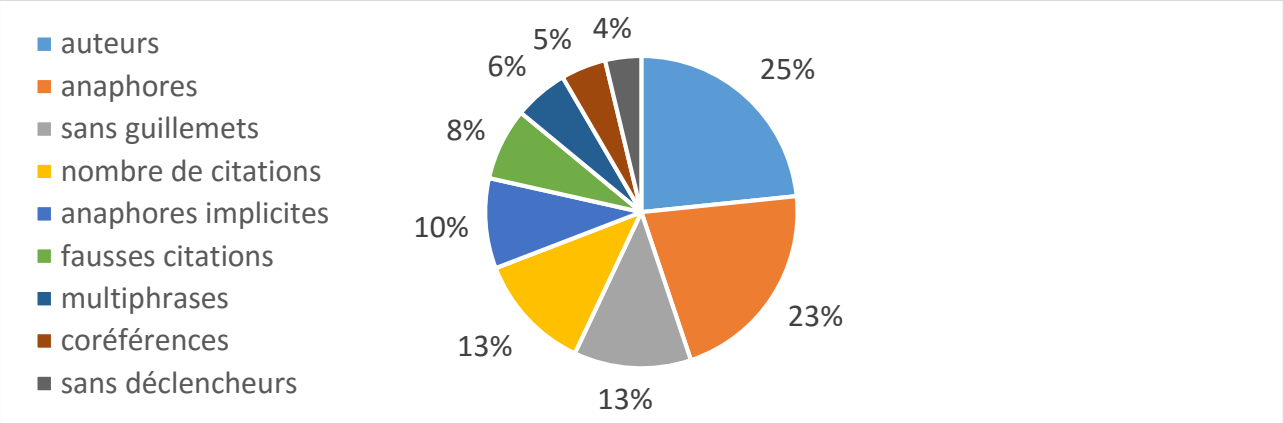


Figure 28 - Répartition des causes d'erreurs et de manques

Face à cette conclusion, deux axes pour améliorer le rappel sont mis en priorité :
- l'amélioration des règles pour une meilleure extraction des auteurs
- l'écriture d'un script pour la résolution d'anaphores

Amélioration des règles d'extraction

Suite aux résultats des règles et à l'évaluation, on a constaté que l'extraction de l'auteur était une des principales causes des citations manquées. On décide donc de réécrire et d'améliorer les règles de façon à extraire plus de citations et plus d'auteurs en suivant la même méthodologie. Pour cela, on recense tous les types d'auteurs de citations possibles rencontrés depuis le début dans les dépêches et on extrait en plus les auteurs de type noms communs comme « étude », « communiqué », on exclut tous les noms de journaux et les noms de lieux des auteurs de type nom propre et on extrait les pronoms personnels pour la future résolution d'anaphores.
Les règles sont aussi réécrites de façon beaucoup plus organisée pour une meilleure visibilité en prenant en compte :

- toutes les combinaisons : QTA (Quote Trigger Author), ATQ (Author Trigger Quote), QTAQ (Quote Trigger Author Quote), etc…
- le nombre de citations présentes dans la phrase (de 1 à 5)
- le nombre de types d'auteurs possibles (11)

On se sert davantage de la négation pour que les règles soient le plus exclusives possibles entre elles et qu'elles ne génèrent aucun doublon. On a recours à la transformation pour normaliser les pronoms en "anaphore" afin d'identifier rapidement les anaphores à résoudre lors de l'étape de

résolution des anaphores.

Le module d'extraction des citations de Cogito Studio a été développé dans le but d'extraire plus de citations que la cartouche Living-Quotes et les résultats démontrent que les règles ont répondu jusqu'ici aux attentes d'amélioration. Cependant les résultats ne sont pas encore optimaux car il manque la résolution d'anaphores, prochaine et dernière étape de ce projet.

5. Résolution d'anaphores

L'extraction des citations étant terminée, nous allons maintenant aborder la résolution des anaphores. Nous présenterons dans ce chapitre le travail effectué pour résoudre les anaphores, fortement présentes dans les citations. Ce chapitre présente la résolution des anaphores pronominales et implicites mais pour des raisons de temps, seule la résolution des anaphores pronominales a été implémentée.

1. PRESENTATION DE DIFFERENTES APPROCHES POUR RESOUDRE LES ANAPHORES

Cogito Studio gère la résolution d'anaphores mais elle n'est pas exploitable pour le français car les résolutions sont rares et le plus souvent mauvaises. Temis, de son côté, a aussi développé une méthode pour résoudre les anaphores mais seulement pour l'anglais et les résultats sont peu concluants.
Nous avons pris connaissance dans la partie de l'état de l'art de quelques méthodes et théories applicables pour la résolution des anaphores qui ont toutes l'air de très bien fonctionné. Cependant les ressources disponibles pour le projet ne nous permettent pas d'implémenter une solution similaire, en tout cas pas avec l'intervalle de temps qui nous restait. Mais les contraintes et les algorithmes appris grâce à ces articles, ont inspiré une méthode plus simple et unique qui permet quand même de résoudre des anaphores pronominales. Ce sont notamment les contraintes de genre et de nombre qui sont les contraintes de base d'une résolution d'anaphores pronominales, auxquelles on ajoute les contraintes syntaxiques (l'auteur candidat doit être sujet de sa phrase) et de position dans le texte (l'antécédent candidat ne soit pas être dans une citation)

2. APPROCHE CHOISIE ET JUSTIFICATIONS

Parmi toutes les méthodes présentées, nous avons choisi d'implémenter notre propre méthode de résolution d'anaphores. Ce choix est en majorité dû à des contraintes de temps, étant donné qu'il ne restait qu'un mois pour implémenter une solution mais aussi par rapport aux ressources qui étaient disponibles pour le projet. Cependant, il est possible que la solution soit améliorée par la

suite.

En ce qui concerne les ressources disponibles, Expert System France ne possède pas de solution qui permette d'effectuer une analyse syntaxique, il faut donc résoudre les anaphores sans ces informations.

On a réalisé une analyse manuelle sur un échantillon d'anaphores (cas_anaphores_a_resoudre.txt) pour vérifier si les algorithmes choisis étaient viables et s'ils valaient la peine d'être implémentés. Il y a donc deux pistes à étudier : celle pour la résolution des anaphores pronominales et celle des anaphores implicites.
Les algorithmes choisis sont présentés ci-dessous sous la forme de schémas.

- Piste 1 (résolution d'anaphores pronominales) : pour chaque pronom qui est auteur d'une citation, noter son genre et son nombre, remonter dans le texte pour trouver le premier nom propre qui correspond aux contraintes de type (personne, compagnie), genre, nombre, qui doit être sujet de la phrase et ne pas se trouver dans une citation

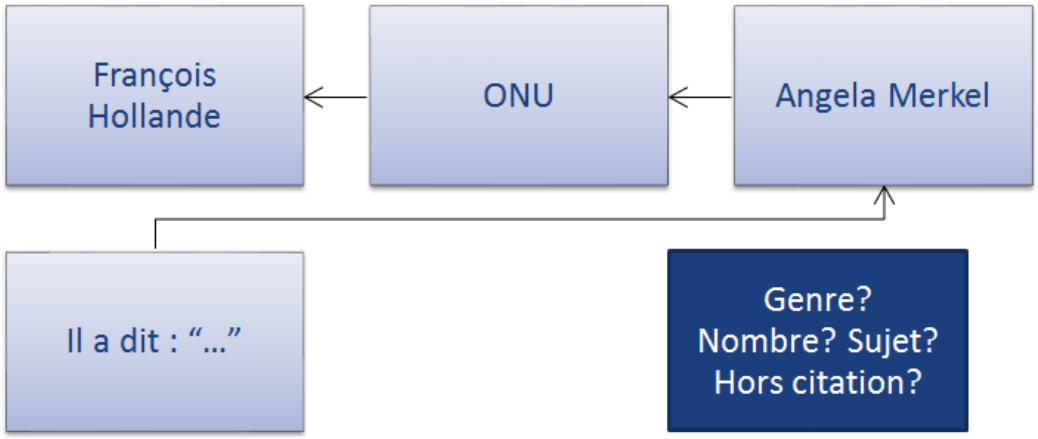


Figure 29 - Schéma de l'algorithme de résolution d'anaphores pronominales

Les résultats sur 25 dépêches :

- Nombre d'anaphores pronominales à résoudre : 48
- Nombre de cas qui fonctionnent (avec les contraintes de sujet, genre et nombre) : 42
- Nombre de cas qui fonctionnent (avec uniquement les contraintes de genre et nombre) : 23

- Piste 2 (résolution d'anaphores implicites) : prendre l'auteur de la citation précédente

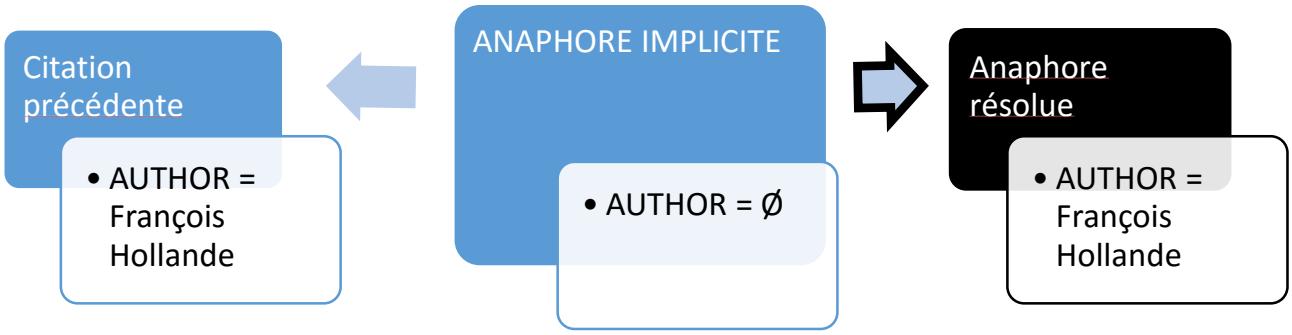


Figure 30 - Schéma de l'algorithme de résolution des anaphores implicites

Les résultats sur 25 dépêches :

- Nombre d'anaphores implicites à résoudre : 42
- Nombre de cas qui fonctionnent (si on a résolu l'anaphore pronominale avant) : 33

Suite à cette analyse, on constate pour la résolution d'anaphores pronominales, que s'il manque les informations syntaxiques, la moitié seulement des anaphores sera résolue en suivant cet algorithme, contre presque 90% si on avait ces informations. En ce qui concerne la résolution d'anaphores implicites, l'algorithme permettrait d'en résoudre environ 80%.

3. RESSOURCES UTILISEES

Pour résoudre les anaphores, on s'est servi des entités nommées extraites par la TM360, en choisissant uniquement les entités nommées de type personne, compagnie et organisation, les seules pouvant être l'auteur d'une citation.

On a également exploité les extractions de Cogito Studio pour retrouver les anaphores à résoudre. Comme dit plus tôt, les anaphores pronominales ont été extraites en tant qu'auteur d'une citation avec les règles d'extraction du projet Cogito Studio et elles ont été normalisées de la façon suivante afin de les identifier plus facilement :

- anaphore-MS : pronom masculin singulier
- anaphore-MP : pronom masculin pluriel
- anaphore-FS : pronom féminin singulier
- anaphore-FP : pronom féminin pluriel

Ainsi le script de résolution utilisera ces deux informations et remplacera par la bonne entité nommée l'anaphore pronominale rencontrée.

4. IMPLEMENTATION

Pour résoudre les anaphores pronominales, un script java est développé.

Mais avant de l'écrire, il faut connaître les données sur lesquelles on travaille. Dans le cas présent, les données à modifier avec le script java sont au format lux. Ces données contiennent les extractions de citations de Cogito studio, avec les anaphores à résoudre et les entités nommées extraites par la cartouche TM360. Voici un exemple de document au format lux généré par la chaîne de traitement :

```
1   <lux >
2       <doc  id="AFP_NEWS_2953" >
3           <content >
4               <text  mimetype="text/plain" > Le prince jordanien Ali bin Al Hussein, qui a annoncé mardi sa candidature à la présidence de la
                Fifa, a cherché à développer le football en Asie avec une attention particulière portée aux jeunes et aux femmes. Agé de 39 ans,
                ce demi-frère du roi Abdallah II est connu dans les milieux sportifs comme un homme modeste, calme et très humain. Vice-président
                de la Fifa pour l'Asie et membre du Comité exécutif de la Confédération asiatique de football (CAF), il dirige également depuis
                1999 la Fédération jordanienne de football et occupe d'autres postes de responsabilité dans ce domaine. En annonçant sa
                candidature, le prince Ali a expliqué vouloir redorer le blason de la Fifa, éclaboussée par plusieurs affaires de corruption. Le
                football mondial "mérite une gouvernance de classe mondiale" et la Fifa doit être "une organisation de service et un modèle
                d'éthique", a-t-il dit en éreintant implicitement la gestion controversée du Suisse Joseph Blatter, candidat à 78 ans à un
                cinquième mandat. En 2012, le prince a créé le projet de développement du football asiatique (AFDP, à but non lucratif) qui a
                pour mission de développer le football à travers l'Asie, en particulier auprès des jeunes, et de valoriser la place des femmes.
                Il se concentre aussi sur la responsabilité sociale du sport. L'AFDP a notamment mené, avec succès, la campagne pour lever
                l'interdiction faite aux femmes voilées de jouer. A son crédit également, la proposition -acceptée- d'augmenter le nombre de pays
                participant à la Ligue des champions d'Asie. - Prince, général et sportif - Le prince Ali se présente comme "un fervent partisan
                du football féminin". "Je suis déterminé à aborder toutes les questions pertinentes afin de veiller à ce que toutes les filles et
                les femmes puissent jouer ce beau jeu à travers le continent (asiatique)", déclarait-il en 2011. Né le 23 décembre 1975 d'un
                troisième mariage du feu roi Hussein avec la reine Alia, une Jordanienne d'origine palestinienne tuée dans un accident
                d'hélicoptère en 1977, le prince Ali a fait ses études aux Etats-Unis, où il a obtenu en 1993 un diplôme de la Salisbury School,
                au Connecticut. Comme la plupart des membres de la famille royale de Jordanie, il a ensuite rejoint l'Académie militaire royale
                de Sandhurst en Grande-Bretagne, dont il est sorti en 1994. Il a servi comme chef de la sécurité spéciale du roi de 1999 à 2008,
                et a rang de général dans l'armée jordanienne. Il est aussi amateur de lutte gréco-romaine. Sa soeur, la princesse Haya, est
                mariée au souverain de Dubaï et vice-président des Emirats arabes unis, Mohammed ben Rached al-Maktoum. Marié depuis 2004 à
                l'ex-journaliste algérienne Rym Brahimi, il a une fille et un garçon. La princesse Rym, fille de Lakhdar Brahimi, conseiller du
                secrétaire général de l'ONU, a travaillé à CNN, pour qui elle a surtout couvert l'Irak. bur-kt/tp/cbo/es </text>
5           </content>
6           <metadata >
23          <boundaries  name="sentence"  set=" converter " >
46          <knowledge  name="EXTRACTION" >
99          <knowledge  name="Knowledge" >
193         <status  valid="true" >
196     </doc>
197 </lux>
```

Figure 31 – Fichier au format lux

Le fichier contient une première balise « content » qui regroupe le texte du document. Ensuite, il y a la balise « metadata » sous laquelle sont toutes les métadonnées liées au document. Sous la balise « boundaries » sont listés le début et la fin de chaque phrase du texte. Enfin, il y a deux balises « knowledge » qui contiennent les informations qui vont être utilisées pour le script de résolution d'anaphores. La première « EXTRACTION » contient les extractions des citations avec Cogito Studio :

Figure 32 - Fichier au format lux : résultats des extractions de Cogito Studio (knowledge Extraction)

Et la deuxième « Knowledge » contient les extractions d'entités nommées de la TM360 :

Figure 33 - Fichier au format lux : résultats des extractions de la TM360 (knowledge Knowledge)

Sous chaque balise « knowledge », on a trois autres balises :

- La balise « types » qui comprend un résumé du type des éléments rencontrés dans le texte
- La balise "descriptors" qui donne un résumé des différentes annotations du texte
- La balise "annotations" qui recense toutes les annotations effectuées sur le texte

Avec ce fichier lux généré à la sortie de la chaîne, on possède toutes les informations nécessaires pour résoudre l'anaphore.

Dans le script java, on suit les étapes de l'algorithme suivant :

Le principe général est qu'à chaque fois qu'on rencontre une anaphore à résoudre, on remonte dans le texte pour retrouver une entité nommée qui peut correspondre.

- On parcourt le fichier lux
- On enregistre dans un set toutes les auteurs des citations du texte
- On enregistre dans un autre set toutes les annotations de type Entity/Person de la TM360
- Pour chaque annotation de type = "/Template/CITATION", on descend dans sa balise <a> dont l'attribut name="AUTHOR"
- Si refName = "anaphore-MS", alors on enregistre le début de la citation et le début de l'auteur ; si refName ≠ "anaphore-MS" alors on passe à l'annotation suivante

  - 1ère option : vérifier si l'auteur de la citation précédente correspond
    - On parcourt de la fin au début le set qui contient les auteurs de citations
    - On prend le premier auteur qui se présente
    - On descend dans le knowlegde des annotations de la TM360
    - Si dans la balise <a> dont le name="Gender" , la value ="Male", alors on remplace anaphore-MS par le nom de l'auteur; si value ≠ "Male" alors on passe à la citation suivante

  - 2ème option : on vérifie les entités Personnes
    - On parcourt de la fin au début le set qui contient les annotations de la TM360
    - On prend la première entité personne qui se présente
    - On descend dans le knowlegde des annotations de la TM360
    - Si dans la balise <a> dont le name="Gender", la value ="Male", alors on remplace anaphore-MS par le nom de l'entité Personne ; si value ≠ "Male" alors on passe à l'entité personne suivante

Le script n'est pas terminé, il lui manque une 3ème et 4ème option pour parcourir également les annotations de types Compagny et Organisation. Mais pour cela, on aurait besoin des informations de genre et de nombre de la compagnie ou de l'organisation potentiellement auteur de la citation.

Tutorial_22_LuxLoopOnFilteredAnnotations.java | AnnotationToCategory.java | MyFirstScriptTester.java | input.lux | Anaphores.java

```
85                // get the SortedSet of the quotations annotated
86                SortedSet<LuxAnnotation> concepts =
87                        docHelper.getSortedAnnotations(
88                                LuxAnnotationPredicates.filterAnnotationsByType("/Template/CITATION"));
89
90                for (LuxAnnotation a : concepts) {
91                    System.out.println("citations : " + a);
92
93                    // get the <a> subtag corresponding to name="AUTHOR"
94
95                    List<LuxAttribute> attr =
96                            a.getAttributes();
97                    //System.out.println("attributs : " + attr);
98
99                    for (LuxAttribute c : attr) {
100                       System.out.println("attribut : " + c);
101
102                       // search for the one with AUTHOR
103
104                       String attrName = "AUTHOR";
105                       //System.out.println("getName : " + c.getName());
106
107                       if (c.getName().equals(attrName)) {
108                           //System.out.print("auteur trouvé!\n");
109
110                           String anaphora = "anaphore-MS";
111
112                           if (c.getRefName().equals(anaphora)) {
113                               System.out.println("anaphore trouvée : " + anaphora);
114
115                               // get the beginning of the quotation
116                               int debut_citation_s = a.getStart();
117                               System.out.println("debut citation : " + debut_citation_s);
```

Figure 34 - Début du script de résolution des anaphores

Le script développé pour résoudre les anaphores est intégré à la chaîne de traitement du projet, après l'extraction des citations par Cogito Studio. Le processus complet du projet se présente ainsi :

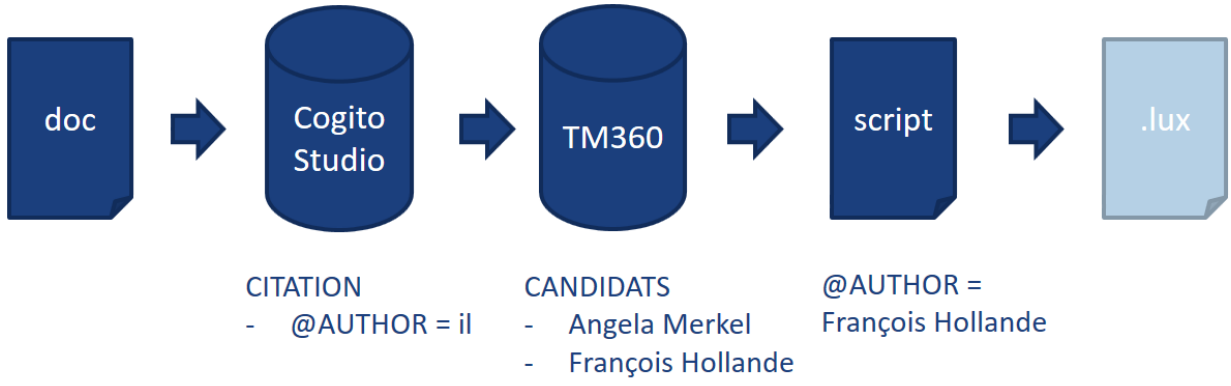


Figure 35 - Schéma de la chaîne de traitement complète

Le script n'a pas pu être terminé, mais il permet néanmoins d'avoir une première idée pour résoudre les anaphores pronominales et il peut servir de base pour une future reprise.

5. EVALUATION FINALE DE LA CHAINE

On présente ci-dessous les derniers résultats de la solution sur un corpus de 100 documents de l'AFP après l'amélioration des règles et la résolution d'anaphores :

	Précision	Rappel	F-mesure
Strict	50%	30%	37%
Tolérant	85%	60%	70%

Tableau 6 - Résultats finaux du projet

Pour rappel, les résultats de la chaîne sans la résolution d'anaphores étaient de 84% de précision, 35% de rappel et 50% pour la f-mesure. On constate donc une amélioration du rappel comme souhaité, ainsi qu'une augmentation de la précision grâce à l'amélioration des règles d'extraction. L'objectif d'élever le rappel à 70% n'a cependant pas été atteint. Une des raisons de ce rappel encore moyen, se trouve peut-être dans la présence de multi-phrases et des anaphores implicites qui n'ont pas encore été résolues.

L'ajout d'une résolution des anaphores pronominales pour les citations a permis d'améliorer la qualité de la chaîne de traitement. La solution présentée au début avait un rappel très bas de 16%. Les objectifs fixés au début du projet ont été remplis mais il reste certains points sur lesquels travailler pour encore améliorer la solution. Ces points sont présentés dans le chapitre suivant qui constitue la dernière partie de ce mémoire.

6. Perspectives d'améliorations et pistes

Nous avons extrait les citations, résolue les anaphores afin d'offrir une solution optimale qui extrait le plus de citations possible. Pour conclure ce mémoire, nous présenterons ici les différentes pistes d'améliorations possibles pour une future reprise de la solution.

1. GESTION DES MULTI-PHRASES

Dans le projet actuel, les citations qui sont composées de plusieurs phrases ne sont pas extraites. Pour les prendre en compte, il faut gérer les SEGMENTS, une technique de l'outil Cogito Studio qui demande du temps mais qui permet de gérer les citations à un autre niveau que la phrase.

2. RESOLUTION DES COREFERENCES

Les coréférences ne sont pas gérées dans le projet actuel. Elles nécessitent une résolution à part.

3. RESOLUTION DES ANAPHORES IMPLICITES

Les anaphores implicites doivent d'abord être extraites avec Cogito avant d'être résolues. Il faut pour cela analyser l'anaphore implicite : comment se présente-t-elle ? Est-ce qu'elle s'étend du début à la fin de la phrase ?
Ensuite il sera possible d'écrire un script java de post-traitement pour les résoudre.
L'idée d'algorithme qui résout l'anaphore implicite en lui attribuant l'auteur de la citation précédente permettrait, hypothétiquement, de les résoudre à 80%.

4. UTILISATION DES OPERATEURS LOGIQUES

Les opérateurs logiques permettent d'extraire les éléments d'une phrase selon leur relation syntaxique.

Opérateur	Variant	Description
&VS	&SV	Relation entre un verbe et son sujet
&VO	&OV	Relation entre un verbe et son objet direct
&SO	&OS	Relation entre un sujet et son objet direct
&SS		Relation entre deux sujets dépendants du même verbe
&OO		Relation entre deux objets directs dépendants du même verbe

Tableau 7 - Opérateurs logiques

Si l'analyse syntaxique était meilleure dans Cogito Studio, ces opérateurs seraient très utiles pour l'extraction des citations car :

- On n'aurait plus besoin de prendre en compte toutes les distances possibles entre un verbe déclencheur et son sujet, autrement dit l'auteur de la citation
- L'auteur, quel qu'il soit, serait toujours extrait et on n'aurait pas besoin de spécifier tous les types d'auteur possibles
- S'il y a plusieurs auteurs associés à un verbe déclencheur, l'opérateur permettrait de les extraire

Mais pour le moment l'analyse syntaxique n'est pas assez bonne pour les utiliser.

5. TRI ET CLASSEMENT DES VERBES DECLENCHEURS

La principale liste de verbes contenue dans le fichier « verbes.txt » est issue d'une première liste de verbes de discours (150), auxquels on a ajouté l'ensemble des verbes de communication linguistiques tirés du Sensigrafo (450). Ces verbes ont été relevés manuellement tels quels et peuvent être triés si du bruit est généré à cause d'eux.
Il est aussi possible de les classer selon leur catégorie sémantique.

6. FAUSSES CITATIONS

Une fausse citation peut correspondre à deux situations :

- Une phrase qui contient des guillemets pour citer un titre de livre, de film, pour introduire une procédure judiciaire, ou encore pour utiliser une expression

Exemple :

Le récent roman "Au Zénith" de la dissidente Duong Thu Huong, qui raconte un amour caché du "Président", sacrifié par le régime, est ultra-tabou à Hanoï.

- Une phrase sans guillemets mais qui ne contient pas d'auteur identifiable et distinct de type Personne, Compagnie ou Organisation.

Exemple :
Dans la soirée, l'aviation israélienne a repris ses bombardements, visant pour la troisième fois depuis le début de l'opération des tunnels de contrebande de Rafah, dans le sud, selon des témoins.

Une solution pourrait être implémentée pour être sûre de ne pas extraire ces fausses citations. Pour une analyse plus profonde de ce problème, un fichier « FaussesCitationsAFP.txt » a été créé et regroupe les fausses citations du corpus de l'AFP.

7. CITATIONS SANS DECLENCHEURS

Elles sont peu fréquentes, mais certaines citations, sans parler des citations à anaphores implicites, ne sont introduites par un aucun déclencheur. Peut-être peut-on les considérer comme des citations avec anaphores implicites ? Un fichier nommé « citations_sans_declencheurs.txt » en regroupe une vingtaine pour une future analyse.

Exemples :

- Sur "l'avenir de l'Europe" et pour faire avancer "la convergence franco-allemande" M. Hollande rencontrera la chancelière Merkel dimanche à Strasbourg.
- La France doit continuer à "soutenir l'opposition" syrienne.
- La nouvelle base avancée française au nord du Niger servira à frapper les jihadistes "à chaque fois qu'ils sortiront des lieux où ils se sont cachés".
- Hollande veut une taxe sur les transactions financières en Europe "pour 2016, au plus tard en 2017", "la plus large possible" et qui puisse servir à financer la lutte contre le réchauffement climatique.

- Pendant un an, il a "vécu" dans les meubles de son prédécesseur avant de refaire son bureau, avec notamment un meuble qui "rappelle" l'usine de Moulinex.

Pour analyser ce cas, un fichier a été créé et regroupe ces citations :
citations_sans_declencheurs.txt

8. CANDIDATS AUTEURS

Le but est de générer une meilleure liste de candidats auteurs avec plus d'informations, dans l'idéal avec des informations syntaxiques et sémantiques.

Actuellement, les candidats auteurs pour la résolution d'anaphores sont extraits à partir de la TM360 (question de temps et de qualité) mais il manque une information importante pour augmenter, voire doubler, le taux de résolution des anaphores : l'analyse syntaxique. Or aucune solution Luxid ne permet une telle analyse. Une solution serait donc d'extraire avec des règles dans Cogito Studio tous les noms propres qui sont analysés en tant que sujet. Une autre solution serait d'utiliser un outil externe à l'entreprise qui analyse la syntaxe et qui puisse être intégré dans le projet à travers un script en java. On pourrait alors écrire un script qui donne en sortie l'analyse syntaxique d'un document et dont on pourrait exploiter les résultats.

Conclusion

Contrairement à ce qu'on pourrait penser, la citation n'est pas un simple discours rapporté mis entre guillemets et pour l'extraire il ne s'agit pas seulement de relever dans un texte tout ce qui est entre guillemets. En effet, il y a beaucoup de cas, d'auteurs, de déclencheurs différents qui combiner tous ensemble donnent des possibilités de citations croissantes.
Malgré les difficultés rencontrées, nous avons réussi à extraire les citations de la typologie. Les citations avec guillemets et sans guillemets sont extraites. Il reste quelques cas qui n'ont pas été pris en compte lors de l'implémentation et qui requièrent des analyses et des implémentations futures. Comme c'est le cas des citations sur plusieurs phrases (multi-phrases) ou des anaphores implicites.
On a vu que l'extraction des citations pouvait être utilisée à des fins diverses. Par exemple, pour la recherche de plagiat (Poulard et al., 2008), pour surveiller les citations d'une même personne sur un sujet précis au fil du temps (Sarmento et Nunes, 2009), pour générer un résumé selon une opinion (Stoyanov et Cardie, 2006), pour l'indexation des documents par les citations (Ritchie et al., 2006), ou encore pour savoir ce que pense X à propos de Y à travers les questions-réponses (Somasundaran et al., 2007). On peut constater que la principale utilisation des citations concerne la politique (van Atteveldt, 2014). Pour permettre, par exemple, aux citoyens de comparer les discours des politiciens avant les élections (le Match des Mots, 2012) ou servir d'instrument de travail pour les professionnels des médias et des sciences politiques. Les citations peuvent aussi être utilisées pour l'Opinion Mining (Balahur et al., 2009).

Bibliographie

Articles scientifiques

Setzer, A. et Gaizauskas, R. (2000). Annotating Events and Temporal Information in Newswire Texts. In *Proceedings of LREC-2000*, pp. 1287–1294

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L. , Strassel, S. et Weischedel, R. (2004). The automatic content extraction (ACE) program-tasks, data, and evaluation. *In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC-2004)*

Riloff, E. (1999). Information extraction as a stepping stone toward story understanding. *In Understanding Language Understanding: Computational Models of Reading*, pp. 435–460

Balahur, A., Steinberger, R., van der Goot, E., Pouliquen, B. et Kabadjov, M. (2009). Opinion Mining on Newspaper Quotations. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Volume 03*, pp. 523–526

Krestel, R., Bergler, S. et Witte, R. (2008). Minding the Source: Automatic Tagging of Reported Speech in Newspaper Articles. In *Proceedings of the Sixth International Language Resources and Evaluation Conference (LREC 2008)*

van Atteveldt, W. (2014). Quotes as Data Extracting Political Statements from Dutch Newspapers by applying Transformation Rules to Syntax Graphs.

Sarmento, L. et Nunes, S. (2009). FEUP - Automatic Extraction of Quotes and Topics from News Feeds. In *DSIE'09 - 4th Doctoral Symposium on Informatics Engineering,* pp. 1-12

Poulard, F., Waszak, T., Hernandez, N. et Bellot, P. (2008). Repérage de citations, classification des styles de discours rapporté et identification des constituants citationnels en écrits journalistiques. In *Traitement Automatique des Langues Naturelles*, pp.450-459

Ho, H., Min, K. et Yeap, W. (2004). Pronominal Anaphora Resolution Using a Shallow Meaning Representation of Sentences. In *PRICAI 2004: Trends in Artificial Intelligence*, pp. 862–871

Kennedy, C. et Boguraev, B. (1996). Anaphora for Everyone: Pronominal Anaphora Resolution Without a Parser. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 1*, pp. 113–118

Watrin, P. et Weiser, S. (2012). Extraction of Unmarked Quotations in Newspapers A Study Based on Direct Speech Extraction Systems. In *LREC - International Conference on Language Resources and Evaluation*

Mourad, G. (2000). Présentation de connaissances linguistiques pour le repérage et l'extraction de citations. In *Conférence RECITAL 2000, Lausanne 16*

De La Clergerie, E., Sagot, B., Stern, R., Denis, P., Recourcé, G., et Mignot, V. (2009). Extracting and Visualizing Quotations from News Wires. In *LTC 2009 - 4th Language and Technology Conference*, pp.522-532

Sagot, B., Danlos, L. et Stern, R. (2010). A Lexicon of French Quotation Verbs for Automatic Quotation Extraction. In *7th international conference on Language Resources and Evaluation - LREC 2010*

Thèses

Ghassan Mourad, Analyse informatique des signes typographiques pour la segmentation de textes et l'extraction automatique des citations

Présentations

Information Extraction and Named Entity Recognition. Christopher Manning, Université de Stanford.

Quelques expériences autour du flux de dépêches AFP (15 Mars 2016). Éric de la Clergerie, INRIA. Journée “Information, Médias et Informatique”

Livres

Recent Advances in Natural Language Processing IV: Selected papers from RANLP 2005, Nicolas Nicolov, Kalina Bontcheva, Galia Angelova and Ruslan Mitkov

Liens consultés

https://en.wikipedia.org/wiki/Information_extraction

http://archinfo41.hypotheses.org/571

http://www.christian-faure.net/2007/05/30/introduction-au-text-mining/

https://fr.wikipedia.org/wiki/Deixis

http://la-conjugaison.nouvelobs.com/regles/grammaire/la-proposition-subordonnee-completive-109.php

www.linternaute.com/dictionnaire/fr/definition/citation/

http://www.cnrtl.fr/definition/citation

http://www.larousse.fr/dictionnaires/francais/citation/16228

https://fr.wikipedia.org/wiki/Anaphore_(grammaire)

Wiki d’Expert System :

http://wiki.expertsystem.com/wiki/bin/view/Main/

Outils d'extraction des citations :

http://storyzy.com/

http://passage.inria.fr/SAPIENS/

Glossaire

- Luxid : technologie de Temis, comprenant notamment les outils SCB, AWB et Xelda
- Cogito Studio : outil d'Expert System pour développer des règles d'extraction, de catégorisation ou de domaines. Il intègre des liens directs vers le Sensigrafo, le Test Bench et la documentation Wiki.
- AWB : outil Luxid pour la validation des résultats d'une cartouche de connaissance sur un corpus choisi
- SCB : outil Luxid pour écrire, éditer et développer des règles d'extraction et les exporter en sous forme de cartouche de connaissance.
- Cartouche de connaissance : module Luxid qui contient un ensemble de fichiers de règles permettant de catégoriser des documents ou d'extraire des informations.
- Wiki : documentation en ligne sur la technologie d'Expert System
- Sensigrafo : réseau sémantique multilingue d'Expert System lié à Cogito Studio. Il contient des syncons qui correspondent à des concepts du monde. Pour le Sensigrafo français, le dictionnaire qui a servit à le remplir est Le Robert.
- Langages C/D/E : langages propriétaires d'Expert System. Le langage C pour écrire les règles de catégorisation, le D pour les domaines et le langage E pour l'extraction.
- LPK : Linguistic PacKage Repository. Fichier compilé à partir du projet Cogito Studio et réutilisable
- Syncon : noeud qui représente un concept dans le Sensigrafo, le réseau sémantique multilingue de Cogito Studio. Chaque syncon a un identifiant unique qui le lie à une définition précise du mot
- ESSEXBridge : cartouche Luxid permettant de lier Cogito Studio à Luxid
- ESSEX : serveur d'Expert System qui englobe Cogito Studio, le désambiguiseur et Sensigrafo
- TM360 : cartouche de connaissances pour extraire les entités nommées

- LivingQuotes : cartouche de connaissances pour extraire les citations
- Xelda : outil Luxid de tokénisation et d'analyse morphologique
- Template : tableau de données à remplir lors d'un projet d'extraction dans Cogito Studio
- Main: fichier principal d'un projet Cogito Studio qui permet de préciser quels sont les fichiers à compiler
- Champs : structures de données du template à remplir lors de l'extraction dans Cogito Studio
- Test Bench: outil de validation de la qualité des règles de Cogito Studio, valable uniquement pour les projets de catégorisation

Listes des tableaux

Table des figures

Annexe

Tableau récapitulatif de la typologie des citations :

Cas	Modèle	Auteur	Déclencheur	Combinaisons	Nombre de citations
		AVEC GUILLEMETS			
1	"passage cité", verbe précisions entité nommée	Tous types	Tous types de verbes	QTA ATQ	1
2	entité nommée précisions verbe : "passage cité"	Tous types	Tous types de verbes	ATQ	1
3	phrase "passage cité" phrase, verbe précisions entité nommée	Tous types	Tous types de verbes	QTA ATQ	De 1 à 5
4	"passage cité", selon/d'après/pour précisions entité nommée	Tous types	Prépositions	TAQ QTA	De 1 à 3
5	"passage cité, verbe/préposition précisions entité nommée. passage cité"	Tous types	Tous types de verbes Prépositions	QTAQ	1

6	« passage cité », verbe/préposition entité nommée, verbe/préposition « passage cité »	Tous types	Tous types de verbes Prépositions	QTATQ ATQTQ TAQTQ TQATQ	De 2 à 4
7	entité nommée précisions verbe/préposition « passage cité », entité nommée précisions verbe/préposition « passage cité »	Tous types	Tous types de verbes Prépositions	ATQATQ TAQATQ	De 2 à 3
SANS GUILLEMETS					
8	entité nommée verbe que/de précisions passage rapporté	Entité nommée : Personne Compagnie Organisation	Verbes : - Déclarer - Annoncer	ATQ	Ø
9	passage rapporté, verbe précisions entité nommée	Entité nommée : Personne Compagnie Organisation	Verbes : - Déclarer - Annoncer - Affirmer - Indiquer - Préciser - Assurer - Dire - Expliquer - Souligner - Ajouter	QTA	Ø
10	selon/d'après/pour précisions entité nommée, passage rapporté	Entité nommée : Personne Compagnie Organisation	Prépositions	TAQ QTA	Ø

Tableau de l'attribut TYPE :

Type (Word Class)	**Description**
ADJ	Adjective
ART	Article
AUX	Auxiliary verb
ADV	Adverb
CON	Conjunction
NOU	Noun
NPH	Human proper noun
NPR	Proper noun
PNT	Punctuation mark

PRE	Preposition
PRO	Pronoun
VER	Verb