MÉMOIRE

présenté devant

Institut National des Langues et Civilisations Orientales (INaLCO)

DÉPARTEMENT INGÉNIERIE LINGUISTIQUE, SPÉCIALITÉ INGÉNIERIE MULTILINGUE

pour obtenir le diplôme de

MASTER

Délivré par

L'Institut National des Langues et Civilisations Orientales

La pertinence des termes filtrants

---

par

ARTHUR BOYER

Sous la direction de :	Daniel Galarreta
Laboratoire d'accueil :	Centre National d'Études Spatiales
Équipe d'accueil :	DSI/IS/GI
Année universitaire :	2013 - 2014

Table des matières

Liste des figures

Remerciements

Je tiens à remercier Thierry Demangeot, responsable de la sous-direction Interface Système (la DSI) ; Aline Jourlin, chef du service Gestion de l'Information (GI) ; Daniel Galarreta, responsable de la cellule Gestion des connaissances et mon maître de stage ; et tout le service GI pour leur chaleureux accueil et leurs conseils bienveillants.

Introduction

Comme un être vivant une entreprise consomme (électricité, eau, matière première, ...), a des besoins (solvabilité, ...), peut croître et même se multipliée... Cette créature-entreprise vit dans un écosystème, un environnement physique, social, économique et juridique dont dépend sa survie. Elle a des droits et des devoirs (personne immatérielle, code du travail) et la nécessite de se nourrir en engrangeant des bénéfices... Cependant son corps n'est pas unique ; comme une fourmilière elle se compose de différentes personnes qui ont chacune une tâche attribuée. Dans son environnement elle est en concurrence avec d'autres entreprises et en partenariat avec d'autres. Enfin, pour survivre à son environnement il lui faut s'adapter.

Notre créature-entreprise vit dans un milieu hostile et changeant, le marché. Avec l'âge elle grandit et acquiert de la sagesse, mais pour atteindre cette sagesse il lui faut apprendre. Tâtonnant, elle découvre son environnement et commet des erreurs. Elle prend des décisions qu'il faut réévaluer ou répéter. Les personnes composant notre créature-entreprise sont chacune dépositaire d'une partie des informations nécessaires au bien-être et au développement de celle-ci. Ces informations jouent également le rôle de mémoire pour notre entreprise. Ces informations ne sont pas qu'un historique des évènements, une chronique des faits et gestes de l'entreprise, elles concernent également leurs « causes », leur « comment ? » et leur « pourquoi ? »

Les personnes porteuses d'informations sont faillibles, elles peuvent oublier les informations ou partir avec. Lorsqu'un membre de l'entreprise part, ses connaissances peuvent être perdues si elles n'ont pas été transmises (des techniques, le classement des répertoires, des contacts, . . .). Ces pertes peuvent ne pas être importantes :

- Les techniques de la personne partante n'étaient plus d'actualités.
- Le savoir-faire de la personne partante est utilisé par différentes personnes restantes.
- Une autre personne qualifiée peut réapprendre ce qui a été perdu.
- . . .

Tout comme ces pertes peuvent parfois être critiques :

- La personne qui est partie était la seule à savoir comment faire fonctionner un logiciel.
- Elle avait développé ses propres outils.
- Elle était la seule à travailler sur son domaine.
- Elle avait sa propre liste de contacts et de clients.
- . . .

Le plus sûr moyen de ne pas perdre de connaissances est de les transmettre. Écrire sur

des documents, enregistrer sur divers supports (numériques ou matériels) les informations essentielles à transmettre et/ou à conserver.

Lorsqu'une entreprise rencontre une situation inattendue elle cherchera à puiser dans ses expériences pour savoir comment y faire face ou à s'inspirer de ses expériences pour trouver une nouvelle solution. Ajouter des connaissances à la mémoire d'entreprise est appelé la capitalisation des connaissances (au sens de rajouter au capital de l'entreprise). La capitalisation des connaissances d'une mémoire d'entreprise permet de transformer les expériences passées en nouveaux moyens de production, de s'inspirer de l'existant pour les projets à venir, elle est l'un des éléments source d'innovation.

Cependant, toutes les entreprises n'ont pas besoin d'une mémoire d'entreprise écrite. La transmission de connaissance peut passer par l'oral. Les sociétés commerciales, l'artisanat, ... en sont des exemples.

Présentation du CNES

Fondé en 1961, le Centre National d'Études Spatiales (CNES) est un établissement public à caractère industriel et commercial. Ses missions vont de l'observation de l'environnement et du climat à l'exploration spatiale en passant par la défense, la sécurité et les télécommunications. Composé en majorité par des ingénieurs, le CNES regroupe plus de 2500 personnes basées sur quatre sites différents (deux à Paris, un à Toulouse et un à Kourou). De par son caractère public le CNES a l'obligation de reverser, selon la loi N°79-18 du 3 janvier 1979 (renforcée en 2008), un exemplaire des documents produits dans le cadre de ses activités aux Archives de France.

La mémoire d'entreprise du CNES

Le CNES travaille sur de la « matière grise » et produit une masse de plusieurs dizaines de milliers de documents par an (62 000 documents traités en 2010). Tous les documents du CNES sont versés aux Archives de France. Ils sont également versés dans les archives du CNES. La valorisation de ce fonds s'effectue au travers du dispositif de la mémoire d'entreprise, duplicat des archives du CNES doté d'un moteur de recherche et de classement automatique. Les documents reversés en mémoire d'entreprise sont aussi bien des publications, des brevets, des rapports sur des projets, des notes techniques, des thèses, ce mémoire ...

Le fonds documentaire du CNES représente plus de 200 000 documents et 300 000 références d'articles. Sont exclus de la mémoire d'entreprise les documents classés "confidentiel" et les documents administratifs.

L'approche d'archivage papier a été abandonnée en 2002 au profit du versement de lots d'archives, par projet, sur CD-Rom, jusqu'en 2008. Depuis, les archives sont en cours de dématérialisation. La mémoire d'entreprise du CNES est organisée selon près de 200 projets et 80 métiers et spécialités du CNES. A l'aide d'une base de connaissances les documents sont indexés grâce à un programme de reconnaissance de termes.

Objectifs

Pour pouvoir faciliter le travail de recherche documentaire des agents du CNES et valoriser les investissements et coûts d'archivage, le service DSI/IS/GI, [1] où j'ai effectué mon stage, a pour mission d'organiser et de capitaliser les ressources informationnelles du CNES.

Le service Gestion de l'Information (GI) est responsable de la tenue du centre documentaire et d'informations du CNES, des commandes de documents (magazine, livre, brevet, papier) pour les agents du CNES et de plusieurs corps de métier, la veille informationnelle, la traduction et la gestion de connaissances.

Le service Gestion des Connaissances (GC) dirigé par Daniel Galarreta à pour mission la gestion de la Mémoire d'Entreprise (ME) : la récupération des documents à reverser dans la ME, l'organisation de la ME et la mise en valeur de la ME. Le service GC doit également gérer la sauvegarde des connaissances des experts quittant le CNES. Pour permettre la capitalisation du patrimoine du CNES, le service de GC possède un moteur de recherche sémantique permettant de faire des recherches sur des documents indexés en mémoire d'entreprise. Le moteur de recherche sémantique du CNES permet de contextualiser les requêtes pour augmenter la pertinence des documents rapportés. Ce moteur de recherche sera présenté en détail durant le chapitre 1.

L'indexation des documents se fait à l'aide de la base de connaissances du CNES qui est constituée à partir de l'ensemble des ontologies des domaines/métiers. Une ontologie est un classement de termes d'un vocabulaire de façon hiérarchique et transversale dont des relations sont formalisées, qui sera présenté en détail au chapitre 1.3.

Les ontologies sont enrichies au fur et à mesure de séances de travail menées entre une terminologue et un expert. Les ontologies sont construites sur deux objectifs différents :

- Mettre à jour les relations sémantiques de termes composant le domaine/métier
- Permettre l'indexation des documents du domaine/métier.

Les relations sémantiques utilisées par le moteur de recherche permettent aux agents du CNES un gain de temps en rapportant un ensemble de documents plus réduit et de meilleure qualité. Elles aident les utilisateurs à effectuer leurs recherches, en proposant des termes sémantiquement proches.

L'indexation des documents par projets et par métiers permet de pouvoir gérer les évolutions des savoir-faire, leurs obsolescences et leurs caractères parfois critique.

Par exemple, lorsque la station spatiale MIR a été envoyée dans l'espace en 1986, les Russes n'avaient pas envisagé qu'elle devrait un jour redescendre sur terre. Lorsqu'elle a commencé à vieillir, il a fallu retrouver, en urgence, les documents concernant sa construction. Ces documents ont servi à calculer sa résistance à la rentrée dans l'atmosphère et sa trajectoire. Mir est sortie de son orbite, volontairement, le 23 mars 2001.

1. Direction du Système d'Information (DSI), dans la sous-Direction Interfaces et Système (IS), le service Gestion de l'Information et de la connaissance (GI).

Problématique

Mon stage au service de Gestion des Connaissance (GC) était situé sur deux axes : l'enrichissement des ontologies et l'amélioration de l'archivage automatique. Ces deux axes ont comme objectif commun de permettre une meilleure recherche et de faciliter la transmission des connaissances.

Lorsque l'on ajoute un document en mémoire d'entreprise, il se range (théoriquement) automatiquement dans une catégorie correspondant à son contenu. L'archivage automatique (filtrage de documents) est réalisé grâce à un parsage du document et à une reconnaissance des termes.

Cette reconnaissance de termes est possible moyennant l'existence préalable de listes déterminées de termes.

- Chaque liste doit appartenir à une catégorie du système de classement de l'archive.
- Les termes composant les listes doivent être caractéristiques de la catégorie de classement de l'archive.
- Les termes composant les listes ne doivent pas être retrouvables dans d'autres catégories du classement de l'archive.

L'objet de ce mémoire va porter sur la pertinence des formes textuelles utilisées pour le filtrage et l'indexation des documents. Autrement dit : une forme textuelle appartenant à une liste permet-elle la discrimination de la catégorie du classement pour laquelle elle est prévue ?

Nous verrons tout au long de ce mémoire le système de classification de l'archivage (chapitre 2), la composition des listes de vocabulaire permettant l'indexation et le classement des documents (chapitre 3), l'utilisation de ces listes via le moteur de recherche ATIC (chapitre 1) et enfin nous chercherons à répondre à la problématique en présentant les travaux effectués lors de mon stage au CNES (chapitre 4).

1 Les moteurs de recherche

1.1 Que cherche-t-on et pourquoi ?

Pour faire en sorte de répondre plus rapidement à des demandes ou simplement pour retrouver des informations permettant de ne pas avoir à refaire ce qui a déjà été fait, le CNES a cherché à partir de 2004 à capitaliser ses connaissances grâce à un moteur de recherche sur sa mémoire d'entreprise.

1.2 Définition

Un moteur de recherche est un outil qui parcourt et indexe automatiquement, dans une base de données textuelles, le contenu des documents qu'il visite. Cette indexation permet aux utilisateurs de retrouver un document à partir d'une requête formulée dans un langage spécifique ou par des mots-clés. Les moteurs de recherche ont quatre principales fonctions :

Retrouver de l'information : Cas d'un expert connaissant les références d'un document qu'il souhaite consulter. Également le cas d'un expert ayant une information qu'il cherche à confirmer.

Consolider une connaissance : Cas d'un expert ayant une connaissance vague d'un sujet donné, qu'il cherche à approfondir. C'est également le cas des personnes cherchant à s'instruire sur une connaissance ou à gagner une nouvelle compétence.

Découvrir une information : Cas d'une veille sur un sujet donné. Partir du connu ou d'une notion pour aller vers l'inconnu en avançant à tâtons. C'est le cas le plus difficile car une personne ne peut pas rechercher une connaissance dont il n'a a priori pas entendu parler.

Diffuser : Cas des informations à partager, des cours à transmettre, des connaissances à conserver ...

1.2.1 Évaluer un moteur de recherche : rappel et précision

Les performances d'un moteur de recherche sont basées sur le rappel et la précision de leurs recherches.

Le rappel : Nombre de documents pertinents retrouvés au regard du nombre de docu-

ments pertinents que possède le fonds documentaire.

La précision : Nombre de documents pertinents retrouvés, rapportés au nombre total de documents proposés par le moteur de recherche pour une requête donnée.

On peut formaliser ces définitions de rappel et précision par les formules suivantes :

$$\text{Le rappel} : R = \frac{VP}{VP+FN}$$
$$\text{La précision} : P = \frac{VP}{VP+FP}$$

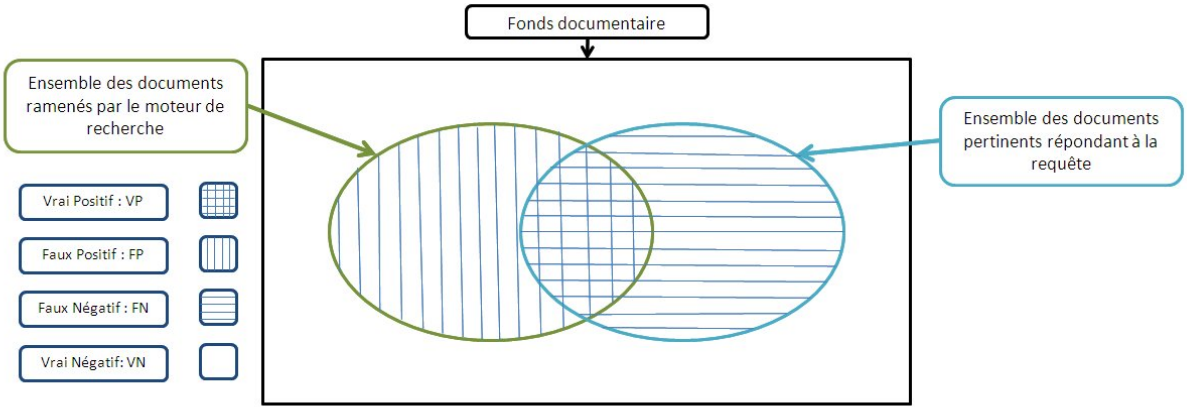

Figure 1.1 — Schéma synthétique : rappel et précision

1.2.2 Méthodes de recherche

La méthode de recherche plein-texte, la plus courante, consiste à retrouver les documents contenant les mots-clés de la requête. Une recherche plein-texte permet (en général) un excellent rappel, mais donne une très mauvaise précision.
Lorsqu'une recherche sera effectuée en plein-texte, tous les documents contenant au moins une occurrence d'un des mots clés de la requête seront rapportés. Les documents rapportés seront très nombreux et contiendront les documents recherchés, mais aussi beaucoup de documents ne correspondant pas à la recherche (bruit).

Un moteur de recherche sémantique est un moteur de recherche qui interprète une requête. Les moteurs de recherche dits sémantiques actuels, utilisent les relations sémantiques des mots-clés de la requête pour affiner les résultats. Ce qui permet une meilleure précision. Par exemple le moteur de recherche google renvoie une fiche bibliographique lorsque l'on fait une recherche sur une personne avant de donner des liens vers des sites répondant à la requête.

1.3 Ontologie

Comme vu précédemment un moteur de recherche indexe le fonds documentaire pour pouvoir renvoyer rapidement un document en réponse à une requête.

Plus haut nous avons défini une base de connaissances, par ce qui la composait. Nous allons ici faire l'exercice inverse et expliquer ce qui la compose, pour présenter le besoin d'une base de connaissances.

1.3.1 Du vocabulaire à la base de connaissance

1.3.1.1 D'une langue de spécialité à un vocabulaire

Les métiers du CNES se caractérisent par des savoirs utilisés pour la réalisation d'objectifs différents. Pour pouvoir utiliser ses connaissances, chaque métier utilise un vocabulaire spécifique et de manière plus générale, une langue de spécialité, pour caractériser les objets qu'il élabore. **Un vocabulaire** est un ensemble de termes définis par une communauté afin de labelliser des objets ou concepts. Avoir un vocabulaire contrôlé permet de partager les informations entre les membres du groupe. Le même vocabulaire utilisé par une communauté différente pourra avoir un sens différent ; autrement dit la signification du vocabulaire est dépendante du groupe qui l'utilise.

1.3.1.2 Du vocabulaire à une taxinomie

Si l'on ne définit pas les relations entre les termes du vocabulaire, le vocabulaire n'est qu'une liste de termes. En classant les termes de façon hiérarchique (selon les hyperonymes et les hyponymes), les relations entre les termes permettent de rajouter du sens à ces derniers. D'un vocabulaire contrôlé, nous passons à un vocabulaire organisé.

Exemple : un "chat" appartient à la catégories "félins" qui est une des catégories de "mammifères" et "chat siamois" est une sous-catégorie de "chat".

Cette organisation s'appelle **une taxinomie**.

1.3.1.3 D'une hiérarchisation à une transversalité

Si l'on veut encore ajouter du sens à l'organisation du vocabulaire en taxinomie, on peut également donner des informations sur les sujets connexes en donnant des termes relatifs.

Exemple : à la catégorie "chat" nous pouvons associer "omnivore", "animal domestique" ... Car ce sont des traits qui catégorisent l'ensemble "chat".

Cet élargissement du champ de connaissance forme un ensemble structuré appelé **un thésaurus** (les traits supplémentaires correspondent à ce qui est habituellement désigné comme « voir aussi »).

1.3.1.4 L'ontologie, une formalisation des liens

Lorsque tous les liens entre les termes sont eux-mêmes spécifiés, cette nouvelle organisation est appelée **une ontologie**.

"An ontology is an explicit specification of a conceptualization. The term is borrowed from philosophy, where an Ontology is a systematic account of Existence. For AI systems, what "exists" is that

which can be represented. When the knowledge of a domain is represented in a declarative formalism, the set of objects that can be represented is called the universe of discourse. This set of objects, and the describable relationships among them, are reflected in the representational vocabulary with which a knowledge-based program represents knowlegde. Thus, in the context of AI, we can describe the ontology of a program by defining a set of representational terms. In such an ontology, definitions associate the names of entites in the universe of discourse (e.g., classes, relations, functions, or other objects) with human-readable text describing what the names mean, and formal axioms that constain the interpretation and well-formed use of these terms. Formally, an ontology is the statement of a logical theory." [Gruber *et al.*, 1993]

En pratique lorsque l'on dit qu'un terme est en relation avec un autre terme, la relation entre les deux termes prend un sens précis, que l'on peut représenter par une flèche entre les deux termes. On nomme cette flèche en y attachant le type de relation qu'elle désigne ("sorte de", "partie de", "idée associée", "sert à", "conséquent de" ...). Une ontologie correspond donc à un vocabulaire, contrôlé et organisé, et à la formalisation explicite des relations créées entre les différents termes du vocabulaire. Une ontologie permet donc une organisation des connaissances à l'intérieur d'un domaine donné. Les connaissances correspondent aux liens entre les éléments du vocabulaire.

Par exemple, un chien de chasse peut se définir par les relations suivantes : *Sorte de :* canin ; *Taille :* de 120 à 150 cm ; *Chant :* aboiement ; *Sert à :* chasser ; *Régime :* omnivore ; *Voir aussi :* animal domestique ;

1.3.1.5 Une base de connaissance

Dans un secteur multidisciplinaire comme celui des activités spatiales *un ensemble d'ontologies* constitue **la base de connaissance** de l'ensemble de ces activités. C'est en tout cas ainsi que ces différentes ontologies sont comprises.

1.4 Exemple du CNES

1.4.1 Une recherche guidée

Grâce à la base de connaissances du CNES constituée d'ontologies, le moteur de recherche de la plateforme ATIC utilisé sur la mémoire d'entreprise peut en plus d'une recherche plein-texte ressortir les relations sémantiques des requêtes (voir chapitre 1.5).

Les ontologies de la base de connaissances regroupent les concepts ainsi que toutes leurs instances (synonymes, variantes orthographiques, acronymes) en français et en anglais et l'ensemble des relations sémantiques (« sorte de », « idée associée à », « partie de », « étape de », ...).

Ainsi lorsqu'un utilisateur lance une recherche sur « trains », le moteur de recherche de la plateforme ATIC permet un affinement et un élargissement de la recherche par les concepts

hyponymes [1] (train de marchandise, train de voyageurs) et hyperonymes [2] (véhicules), pour permettre à l'utilisateur de naviguer plus facilement.

De plus, grâce à l'indexation des documents en métier et projet, la possibilité de rechercher selon le domaine et donc de contextualiser sa recherche existe également (voir figure n°1.2)

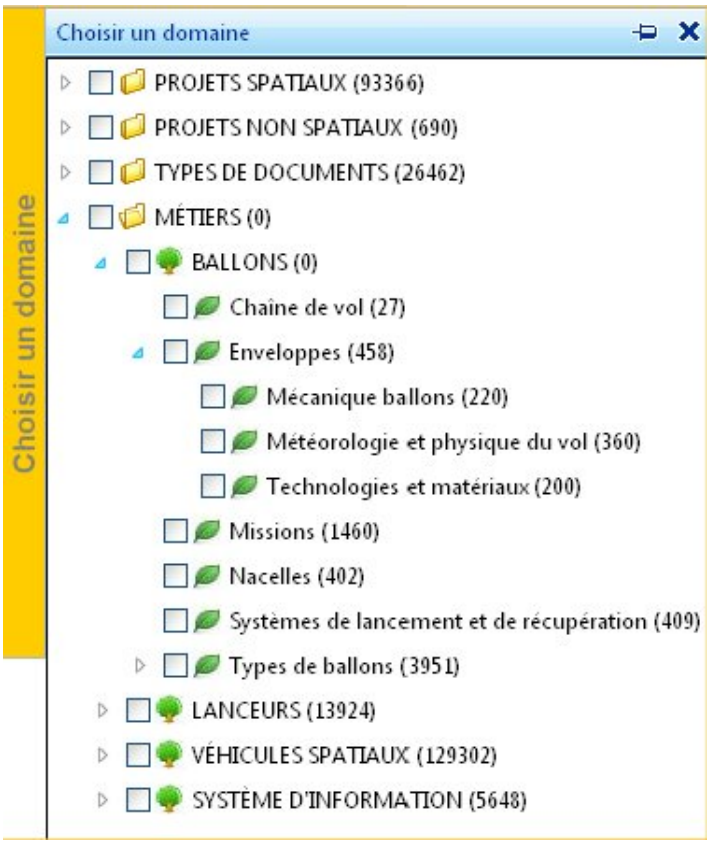

Figure 1.2 — Choix du domaine de recherche dans le moteur de recherche de la plateforme ATIC

1.4.2 Exemple de recherche

À une recherche sur « véhicule », le moteur de recherche de la plateforme ATIC retrouvera tous les documents contenant le mot véhicule, mais également ceux contenant ses hyponymes (voitures, voiture de sport, voitures familiales, camions, fourgons, camionnette, trains, train de marchandises, train de voyageurs, ...), leurs synonymes (voitures, bagnoles, automobiles) et également les documents contenant les équivalents de ces mots en anglais (voir figure n° 1.3).

En contextualisant les recherches, les moteurs de recherche sémantique évitent le bruit engendré. En effet, un signe peut avoir plusieurs référents si le contexte n'est pas connu.

1. Hyponyme : Mot dont le sens est plus spécifique que celui d'un autre.
Les co-hyponymes peuvent être incompatibles (soit l'hyponyme A, soit l'hyponyme B) ou non.
2. Hyperonyme : Mot dont le sens inclut celui d'un autre mot.

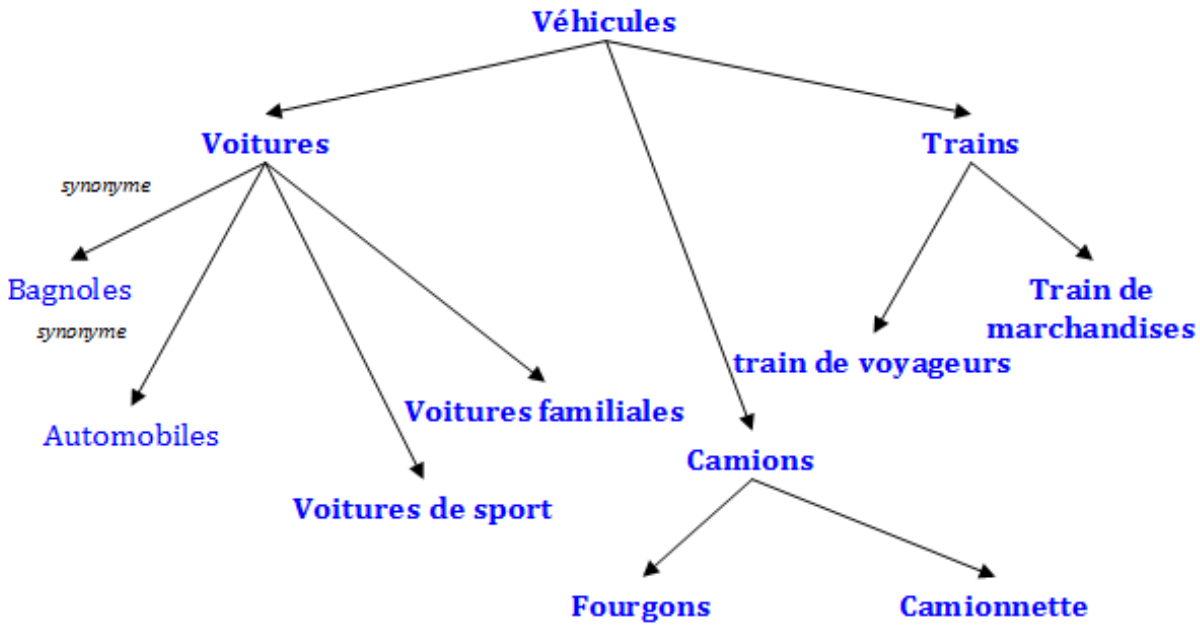

Figure 1.3 — Exemple de réseau sémantique

Exemple : jaguar : la voiture, l'animal.

Une requête comme « voiture de sport » sortira, avec un moteur de recherche en texte plein, tous les documents ayant « voiture », « sport » et « voiture de sport » ; le moteur de recherche sémantique ne rapportera que les documents appartenant à l'ensemble « véhicules ». Les moteurs de recherche sémantique combinent les avantages des recherches en texte plein et des ontologies. Le rappel et la précision des moteurs de recherche sémantique sont très bons.

1.5 ATIC

ATIC, ATelier d'Ingénierie des Connaissances, est la plateforme utilisée au CNES permettant la création, l'enrichissement et l'exploitation de la base de connaissance. Développé par la société Arisem, ATIC a été adopté au CNES en 2004. Arisem est une filiale du groupe Thales [3].

Atic est également la plateforme permettant le rangement automatique de la mémoire d'entreprise. Ce point sera développé ultérieurement (voir chapitre 4.2).

3. Page officielle d'Arisem :
`https://www.thalesgroup.com/fr/content/arisem-vous-connaissez`

1.5.1 Composants d'ATIC

ATIC utilise le moteur de recherche Lucene. Lucene est un logiciel libre, c'est un projet open source de la fondation Apache. Lucene indexe automatiquement les documents en utilisant une table d'indexation.

Pour organiser la base de connaissances, ATIC a utilisé ITM (Inteligence Topic Map) comme environnement pour classer ses ontologies. ITM, utilise le langage SKOS [4] (Simple Knowledge Organization System) pour exprimer les relations entre les concepts. Cependant, c'est pour s'affranchir du format propriétaire que le CNES a développé sa propre base de connaissance grâce à Protégé [5].

Les ontologies sont actuellement écrites en OWL [6] (Web Ontology Language) puis converties en RDF [7] (Ressource Description Framwork) pour être exploitables.

La plateforme ATIC utilise également des composants d'analyse linguistique qui fournissent des méta-données à Lucene. Les méta-données sont trouvées par l'utilisation des ontologies (hyponymes, hyperonymes, synonymes ...), d'un traducteur (équivalence en anglais).

1.5.2 Exemple

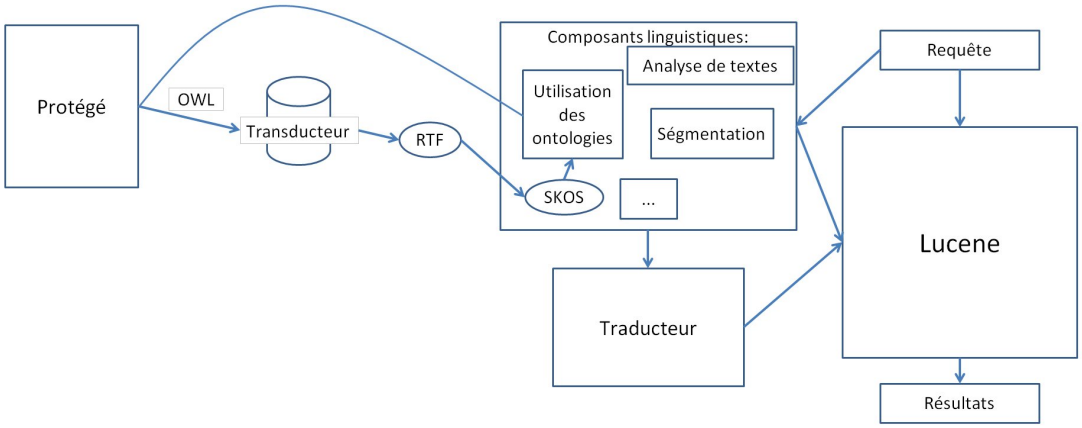

Figure 1.4 — Schéma de fonctionnement de la plateforme ATIC

4. SKOS est un langage formel utilisé permettant une représentation de vocabulaire contrôlé et structuré. SKOS est une recommandation officielle du W3C.
Pour plus de renseignement :
`http://www.w3.org/2004/02/skos/`

5. Protégé est un éditeur d'ontologie en Open-source. Il a été développé par l'Université de Standford par le groupe de travail de Mark Musen.

6. OWL est un langage de représentation de connaissance conçu comme une extension de RDF.
OWL est une recommandation du W3C pour le web sémantique.
Pour plus de renseignement : `http://www.w3.org/TR/owl-guide/`

7. RDF est un modèle de graphe formel. Sa syntaxe utilise des triplets pour définir la **relation** entre un **sujet** et un **objet**.
RDF est interprétable par SKOS et par OWL

ATIC utilise les éléments de la requête pour faire une première recherche plein texte avec Lucene. Il va ensuite analyser la requête pour récupérer d'autres données à l'aide d'ontologies [8]. Ces données vont être recherchées en plein-texte par Lucene. Les équivalents en anglais de ces données (et de la requête) seront également transmis à Lucene. Les résultats seront croisés pour obtenir le plus de précision possible.

Par exemple lorsque l'on fait une recherche sur "conduire une Jaguar" ATIC va rechercher avec Lucene dans la base de donnée les documents contenant les mots "conduire", "Jaguar" et la séquence "conduire une Jaguar". Ensuite la base de connaissance permettra de faire ressortir une série de termes en rapport à la conduite et aux Jaguars. Ces termes ainsi que ceux de la requête seront également lemmatisés puis traduits en anglais. L'ensemble des termes obtenus sera utilisé pour une nouvelle recherche.

Les résultats de la première recherche et de la seconde seront croisés pour éliminer les résultats les moins pertinents et mettre en valeur ceux répondant le plus aux deux recherches. l'ensemble "véhicules" qui est l'ensemble commun à "conduire" et "Jaguar" (contrairement à l'ensemble des animaux) sera surement le plus représenté.

8. Pour plus d'information, en français, sur les ontologies et l'éditeur d'ontologie Protégé : voir [Noy et McGuinness, 2000]

2 Classification

2.1 Historique

"On classe comme on peut, mais on classe." [Lévi-Strauss, 1962]

2.1.1 Les premières classifications

Les classifications représentent l'organisation d'un domaine. Depuis longtemps, l'Homme s'efforce de nommer, classifier, organiser ce qui l'entoure. Très tôt, dès Théophraste (-372 -287), les premières classifications de type utilitaire sont apparues. Jusqu'aux ébauches d'une méthode de classification scientifique apportée par le naturaliste Carl Von Linné, les systèmes de classements consistaient à rassembler dans un catalogue raisonné les entités dont on avait besoin. La classification linnéenne, ou classification classique, est fondée sur l'analyse comparée des caractères morphologiques des espèces. Ce genre de classification a été utilisée jusqu'à la seconde moitié du XXe siècle où la classification phylogénétique est apparue. La classification phylogénétique s'inspire du mouvement cladistique. C'est-à-dire elle classe, ici les êtres vivants, selon des liens de parenté dans un cadre évolutionniste. Cette approche met en avant un nœud souche, dans une arborescence, qui est l'ancêtre ou l'ensemble commun.

2.1.2 Les classements automatiques

Avec l'avènement de l'informatique, des méthodes de classement automatisé sont apparues également. Ces classifications existent en deux types différents : le **classement supervisé** et la **classification non-supervisée**. Ces deux types de classement correspondent à deux approches différentes. La classification non-supervisée regroupe les données à classer de façon statistique. Les catégories ainsi créées nécessitent d'être nommées. Cette approche a l'avantage de ne pas être influencée par des conceptions extérieures au milieu où le classement a eu lieu (voir figure 2.1). La classification supervisée quant à elle se réfère à des catégories déjà existantes. Les données vont se classer, par apprentissage sur un jeu de données déjà classé, dans les catégories existantes.(voir figure 2.2)

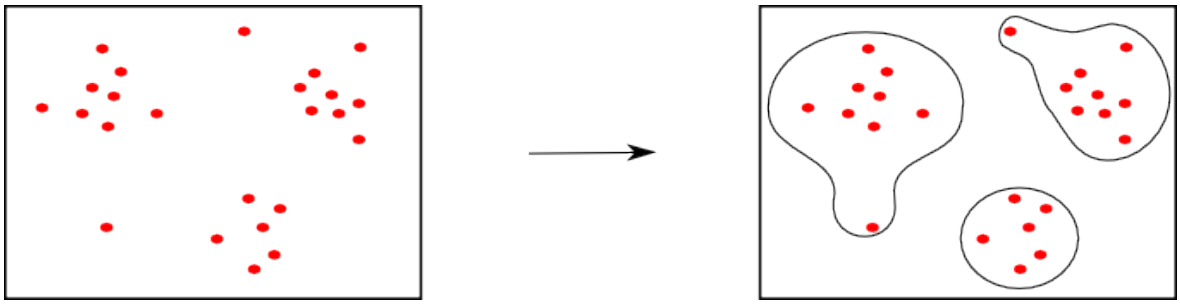

Figure 2.1 — Exemple d'une classification non-supervisée

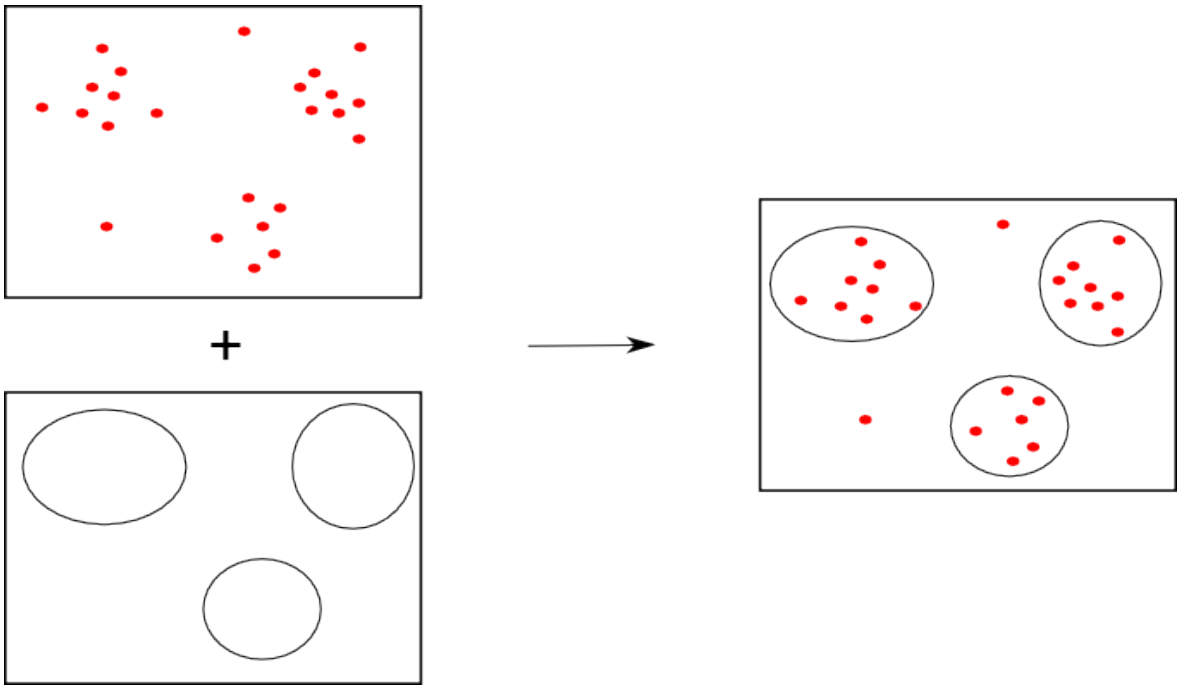

Figure 2.2 — Exemple d'un classement supervisé

2.2 Comment définir des catégories de classement pertinentes

2.2.1 Partir de l'usage

Si l'on poursuit l'hypothèse de Sapir-Whorf ([Sapir, 1921], [Whorf *et al.*, 1956], [Hoijer, 1954]), étant donné que le langage est dépendant de la culture de celui qui s'exprime, et que le classement, l'organisation des connaissances d'un domaine est dépendant du langage pour lequel il est prévu, on peut en conclure que tout classement est dépendant de la culture de la communauté classifiante. Autrement dit, l'organisation de l'information est dépendante de l'usage.

"Human beings do not live in the objective world alone, nor alone in the world of social activity as ordinarily understood, but are very much at the mercy of the particular language which has become the medium of expression for their society. It is quite an illusion to imagine that one adjusts to reality essentially without the use of language and that language is merely an incidental means of solving

specific problems of communication or reflection. The fact of the matter is that the 'real world' is to a large extent unconsciously built upon the language habits of the group. No two languages are ever sufficiently similar to be considered as representing the same social reality. The worlds in which different societies live are distinct worlds, not merely the same world with different labels attached... We see and hear and otherwise experience very largely as we do because the language habits of our community predispose certain choices of interpretation." [Sapir, 1921]

En **partant de l'usage prévu**, la conception d'une organisation de l'information pour une communauté donnée devient donc **pertinente**.

2.2.2 Conséquences pratiques

Pour classer le fonds documentaire de la mémoire d'entreprise, une organisation a été proposée. Cette organisation est orientée vers les agents du CNES et prend en compte le fait qu'ils en seront les utilisateurs. Le classement des compétences du CNES est réparti selon l'organigramme des métiers du CNES (transposition aux métiers de l'organigramme de l'entreprise) et des projets réalisés. Cette organisation permet au personnel du CNES de pouvoir rechercher facilement des informations ayant trait à leurs domaines de compétences et de pouvoir retrouver rapidement toutes informations critiques (voir introduction) concernant un projet.

Cette organisation, cependant, ne permet pas d'automatisation du classement du fonds documentaire du CNES. Il y a deux raisons principales faisant qu'un classement par apprentissage non-supervisé n'aie pas été retenu comme solution pour le CNES.

La première est que le fonds documentaire recouvre des disciplines différentes (aéronautique, cryogénie, ballons, radio-fréquence, astronomie, génie mécanique, traitement d'information ...) et qu'au sein de ces disciplines, la répartition en nombre de documents peut être inégale (il y a par exemple beaucoup plus de documents sur le domaine des antennes que sur le domaine de la biologie). En conséquence, certaines disciplines seraient noyées dans la masse des autres documents si une répartition statistique avait lieu.

La seconde raison est que les documents qui y sont produits appartiennent à de nombreux genres différents (didactique, brevet, notes, retour d'expérience, publication, livre, encyclopédie, présentation PowerPoint, ...), un apprentissage non-supervisé ne distinguerait plus les différences entre disciplines, mais entre genre textuel.

Une troisième raison serait que pour pouvoir construire notre système de base de connaissance, nous avons recours à des experts (voir chapitre 4). Cependant, aucun domaine ne peut être maîtrisé par un seul expert. Il y a au sein d'un même domaine plusieurs disciplines ayant chacune plusieurs métiers.

C'est pourquoi nous avons adopté une organisation selon l'organigramme des métiers du CNES.

En revanche, à l'intérieur de ces métiers, nous découpons le métier en spécialités sur une base statistique textuelle (voir chapitre 2.3). Ce découpage en :

$$\text{Domaine} \supset \text{Disciplines et technologies} \supset \text{Métiers} \supset \text{Sous-domaine}$$

Exemple :

Ballon ⊃ Nacelle pointée ⊃ Missionnisation ⊃ Architecture mécanique

représente une arborescence profonde du fonds documentaire et permet aux utilisateurs de rechercher précisément les informations dont ils ont besoin. Ces spécificités permettent de répartir les documents existants.

2.3 Remarque sur l'approche statistique textuelle mise en œuvre

L'approche statistique textuelle mise en œuvre dans le cadre du CNES repose sur la méthodologie développée par Max Reinert [1] à l'Institut Statistique de l'Université de Paris. Cette méthode s'est concrétisée sous la forme de l'outil Alceste, qui est maintenant la propriété de la société Image [2]. Alceste est l'acronyme d'Analyse des Lexèmes Cooccurrents dans les Enoncés Simples d'un Texte. Cet outil base ses principes sur l'analyse des données et la structure sémantique du texte, grâce à la répartition du vocabulaire dans le corpus. Cet outil est principalement utilisé comme aide à l'analyse et l'interprétation.

Alceste, à partir d'un corpus, effectue une première analyse détaillée de son vocabulaire, et constitue le dictionnaire des mots ainsi que de leur racine, avec leurs fréquences. Ensuite, par fractionnements successifs, il découpe le texte en segments homogènes contenant un nombre suffisant de mots, et procède alors à une classification de ces segments en repérant les oppositions les plus fortes. Cette méthode permet d'extraire des classes de sens, constituées par les mots et les phrases les plus significatifs. Les classes obtenues représentent les idées et les thèmes dominants du corpus. L'ensemble des résultats triés selon leur pertinence, accompagnés de nombreuses représentations graphiques et de différents rapports d'analyse, permet à l'utilisateur une interprétation aisée et efficace. (source `http://www.image-zafar.com/fr/logiciel-alceste`)

2.3.1 Utilisation pratique d'Alceste

Pour pouvoir utiliser l'outil Alceste, il faut regrouper tout le corpus en un fichier. Chaque document du corpus peut être clusterisé s'il apparaît précédé d'au moins un titre commençant par une étoile (exemple : ∗module_Radar ∗chapitre_8). Cette séparation est appelée par Alceste "**Unité de Contexte Initial**" (UCI).

Cela implique qu'il faut nettoyer le corpus avant de pouvoir le soumettre à une analyse Alceste. En effet certains documents utilisent l'étoile comme notation pour le signe de la multiplication, faire une liste à puces ou en tant qu'astérisque. Nous avons nettoyé manuellement le corpus puis, grâce à un script, nous concaténons les fichiers contenus dans un dossier, en un fichier ayant les annotations adéquates.

Alceste découpe alors les UCI en liste de termes et en segments de quelques lignes, d'une longueur fixée par un nombre de mots pleins (10 à 20). Cette segmentation essaie de respecter, si possible, les coupures proposées par la ponctuation. Ces segments sont appelés par

1. `http://www.printemps.uvsq.fr/reinert-max-143304.kjsp`
2. `http://www.image-zafar.com/`

Alceste "**Unités de Contexte Élémentaires**" (UCE). La ponctuation étant importante, tout corpus contenant des présentations Powerpoint, qui sont très souvent composées de morceaux de phrase sans ponctuation, aura des problèmes de segmentation.

2.3.2 Base statistique d'Alceste

Alceste utilise l'algorithme du Chi2(X^2) prouvé par [Kendall, 1967], [Benzécri, 1973] et [Reinert, 1986] [3]. Pour une matrice booléenne, ayant en ordonnée les termes et en abscisse les UCE, une classe est définie comme un ensemble de lignes. Une ligne ne peut appartenir qu'à une seule classe. Le système Alceste réalise ces rapprochements par des statistiques de classification hiérarchique descendante. Ces statistiques sont indépendantes du sens et cherchent à déterminer l'organisation et la structuration des éléments qui constituent le corpus.

Pour rassembler les UCE en classe, Alceste recherche les dichotomies entre les UCE (répartition des termes entre UCE). Une fois la dichotomie maximale trouvée, il sépare le texte en deux classes. Il recommence le processus en partant de la classe la plus grande, jusqu'à atteindre un nombre de classes déterminé.

À chacune de ces classes est attachée une liste de termes.

2.4 Résumé

Le classement du fond documentaire du CNES se fait par apprentissage supervisé via la plateforme ATIC (rappel chapitre 1.5). Les documents de la mémoire d'entreprise sont organisés selon les métiers et missions du CNES. Ces domaines sont eux même subdivisés en sous-domaines qui peuvent eux même être subdivisés en parties plus petites, etc...

Pour découper les domaines nous avons recours à des interviews avec un expert qui nous donne son point de vue sur le découpage taxinomique de son domaine. Lors de cette interview nous récupérons également un corpus identifié par l'expert avec l'aide du terminologue(voir chapitre 3). À l'aide du logiciel Alceste nous découpons statistiquement le corpus pour en déduire des classes qui serviront à un découpage taxinomique du domaine.

Lors d'un second interview nous présentons le découpage taxinomique réalisé grâce à Alceste à l'expert et nous lui demandons de nommer ou fusionner les différentes classes de ce découpage. Suite à cela une discussion est ouverte entre l'expert et le terminologue sur la meilleur façon de découper le domaine.

Le découpage taxinomique final sera utilisé pour le rangement automatique des documents dans la mémoire d'entreprise et pour l'organisation de la base de connaissance.

3. Pour un complément d'information : [Lapalut, 1995]

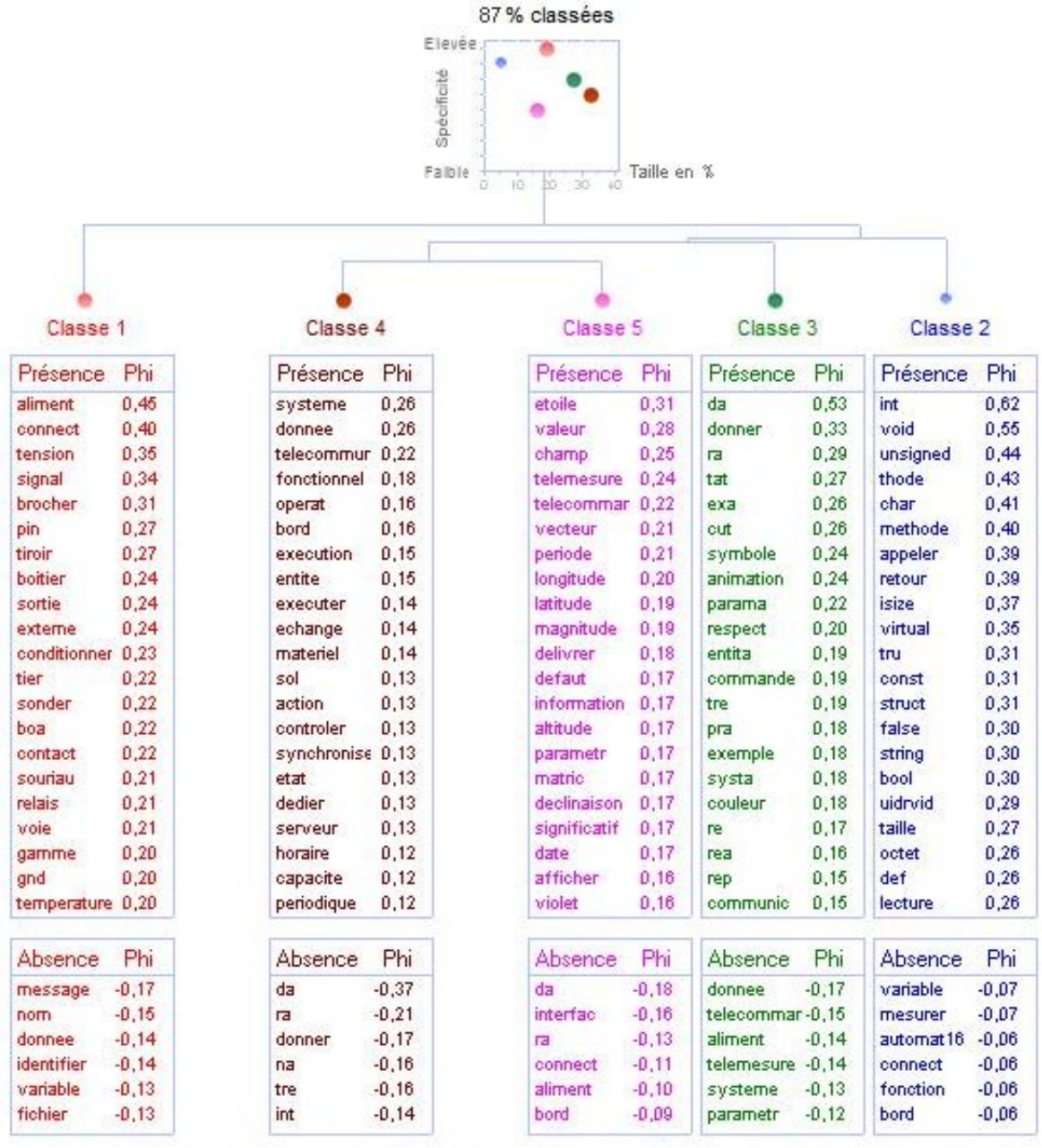

Figure 2.3 — Exemple d'un découpage statistique d'un corpus par Alceste

3 Travail sur le corpus

3.1 Le corpus de référence

L'acquisition de connaissances

"Dans les langues de métier, la prolifération conceptuelle correspond à une attention plus soutenue envers les propriétés du réel, à un intérêt mieux en éveil pour les distinctions qu'on peut y introduire" [Lévi-Strauss, 1962]

L'acquisition des connaissances d'un expert à partir de textes pose le problème de la composition d'un corpus représentatif de ses compétences. L'hypothèse derrière l'utilisation de documents textuels comme base de travail est qu' "*ils fournissent les éléments stables, consensuels et partagés d'un domaine*" ([Aussenac-Gilles *et al.*, 2012] citant les travaux de [Bourigault et Slodzian, 1999] ; [Condamines, 2003]). D'après [Aussenac-Gilles *et al.*, 2012], l'utilisation de textes pour extraire les connaissances est conditionnée par deux éléments clés : *la construction d'un corpus pertinent* et une contribution régulière d'*un spécialiste* du domaine pour assurer l'interprétation des résultats.

À la définition donnée par [Sinclair, 1996] [1], sur les corpus, j'aimerais rajouter la précision apportée par [Rastier, 2005] [2] : *un corpus est composé dans le but d'une gamme d'utilisation précise*. Au CNES, la composition d'un corpus d'étude est réalisée par un expert avec l'aide d'un terminologue lors d'une interview.

1. Un corpus est une collection de données langagière qui sont sélectionnées et organisées selon des critères linguistiques explicites pour servir d'échantillon du langage.[Sinclair, 1996, p.4]

2. "De l'archive au corpus de travail. [...]
1/ L'archive contient l'ensemble des documents accessibles. Elle n'est pas un corpus, parce qu'elle n'est pas constituée pour une recherche déterminée.
2/ Le corpus de référence est constitué par ensemble de textes sur lequel on va contraster les corpus d'étude.
3/ Le corpus d'étude est délimité par les besoins de l'application.
4/ Enfin le sous-corpus de travail en cours varie selon les phases de l'étude et peut ne contenir que des passages pertinents du texte ou des textes étudiés." [Rastier, 2005]

3.2 Interview

3.2.1 Méthodes

Une interview pour composer une taxinomie consiste selon [Vogel, 1988], à identifier, justifier, clarifier, contrôler.

Identifier : l'expert se prononce sur les différents grands objets composant son domaine et les documents qui y sont liés.

Justifier : le terminologue interroge l'expert et l'amène à méditer sur les identifications qu'il a faites, en lui demandant d'expliciter ses choix. La collecte de taxèmes n'est donc pas un moment passif.

Clarifier : Identifier les zones de compétences de l'expert au sein du domaine.

Contrôler : examiner les taxèmes variables ou arbitraires.

L'interview pour la composition d'un corpus repose sur des principes similaires, accomplis dans un ordre différent sur plusieurs séances. À partir de la **clarification**, nous délimitons les zones de compétence de l'expert, puis nous lui demandons d'**identifier** des documents représentatifs de son métier. La **justification** se déroule après le découpage statistique du domaine (voir chapitre 2.3). On peut être amené à demander à l'expert d'enrichir une classe peu peuplée en nous donnant de nouveaux documents ou en explicitant des taxèmes. Le **contrôle** est fait une fois que l'on a extrait, grâce à l'extracteur terminologique Talismane (voir chapitre 3.3), les concepts du corpus. Un entretien est alors réalisé pour nuancer les extractions (bruit) et corriger les erreurs dues à la composition du corpus.

Les Corpus utilisés par l'équipe de Gestion des Connaissances sont usuellement issus du cours de "*Techniques et Technologies des Véhicules Spatiaux*" (TTVS). Les parties étant en relation avec le domaine d'expertise, sont identifiées et serviront à constituer le corpus.

Le TTVS est un document didactique conçu pour la formation de nouveaux employés du CNES. C'est donc un document très riche en connaissances, écrit dans les différentes « langues de spécialités » correspondant aux différentes disciplines représentées dans ce cours. Par construction, les termes « importants » pour comprendre les domaines traités sont concentrés sur peu de pages. Ce document didactique permet d'enrichir facilement la base de connaissance, car les relations entre les termes (sorte de, partie de, conséquent de, ...) y sont explicitées.

3.2.2 Qu'est ce qu'un expert et un spécialiste ?

Un expert est une personne compétente dans un domaine, tirant son savoir d'une somme d'enseignements issus de l'expérience. Le savoir de l'expert est un savoir "privé", qui dépend des expériences qu'il a vécues, et qui lui appartient.

Contrairement à un expert, un spécialiste dispose d'un savoir "public", généralement sous-tendu par un ensemble de théories établies et par un faisceau d'expériences communes à une discipline.

En fait, le même individu est tour à tour spécialiste, informé de l'état de l'art dans son

domaine, et expert, utilisant un réseau de convictions intimes. L'opposition entre expert et spécialiste repose sur une opposition entre un modèle théorique et un modèle pratique de compétence.

3.3 Extraction terminologique.

Pour pouvoir exploiter un corpus dans le but d'enrichir une ontologie, il faut pouvoir extraire des informations terminologiques. L'un des problèmes que pose l'extraction d'information est d'identifier les relations entre termes (exemple de relation hyperonyme : le X est un Y). Ce problème, traité entre autres par [Malaisé *et al.*, 2004] et [Aussenac-Gilles et Condamines, 2009], présente de grandes variabilités de réalisations au sein d'un texte. Cette variation est encore plus notable entre différent genre textuel. De nombreux travaux ont traité de ce problème, notamment [Rebeyrolle et Tanguy, 2000] sur le repérage de structures énonciatives, qui présente l'intérêt de travailler sur un texte étiqueté et lemmatisé plutôt que sur du texte brut. On peut également évoquer le développement d'analyseurs syntaxique, suite aux travaux de [Rebeyrolle et Tanguy, 2000], comme Syntex [Bourigault *et al.*, 2005].

3.3.1 Talismane

Talismane est un analyseur syntaxique probabiliste développé par Assaf Urieli [3], à l'occasion d'une thèse démarré en 2009, à l'université Toulouse II, le Mirail, dans l'Équipe de Recherche en Syntaxe et Sémantique (ERSS). Écrit en java et disponible en ligne [4] sous licence GLP, Talismane fonctionne sur tous les systèmes d'exploitation et est intégrable à d'autres applications.

3.3.1.1 Une analyse en cascade

Pour réaliser une analyse syntaxique Talismane part d'un texte brut pour le segmenter en phrase, tokeniser les phrases, étiqueter les mots (catégorie morphosyntaxique) et parser [5] (repérage et étiquetage des dépendances syntaxiques entre les mots).

Les modules d'étiquetage morphosyntaxique et d'analyse syntaxique en dépendances de Talismane, sont tous deux entraînés sur le corpus annoté du French Treebank [Abeillé *et al.*, 2003]. Les deux modules se basent sur un classifieur par entropie maximale (MaxEnt) [6] (source : [Tanguy, 2012]).

3.3.1.2 Une analyse configurable

À Chaque étape de l'analyse en cascade correspond un module. Les modules sont configurables à la fois au niveau des traits et des règles. Les traits sont les informations sur les

3. http://univ-tlse2.academia.edu/AssafUrieli
4. http://redac.univ-tlse2.fr/applications/talismane.html
5. Le parsing de Talismane se base sur l'algorithme décrit dans [Nivre, 2008]
6. http://www.maxent2013.org

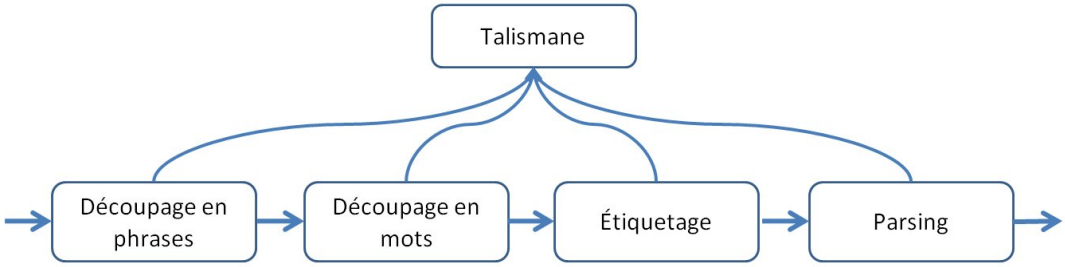

Figure 3.1 — Chaîne de traitement de Talismane

configurations rencontrées dont dispose l'algorithme pour prendre des décisions, alors que les règles sont des contraintes qui forcent ou interdisent des décisions locales.

Avoir des traits et des règles définissables donne la possibilité de respecter des contraintes propres à un corpus spécifique.

Il existe un modèle de traits par défaut proposé par Talismane. Ce modèle utilise les traits classiques pour chacun des modules. C'est ce modèle standard que nous utilisons au CNES.

3.3.1.3 L'analyse syntaxique en dépendance

Il existe différentes méthodes d'analyse syntaxique : l'analyse syntagmatique[7] (par constituants) et l'analyse en dépendance[8]. Elles peuvent être guidées par les règles d'une grammaire (probabiliste ou non) ou être entraînées sur un corpus. Il existe deux principales techniques d'analyse syntaxique en dépendance : l'analyse par graphes [McDonald, 2006] et l'analyse par transitions [Nivre, 2008]. Talismane se base sur l'analyse syntaxique en dépendance par transition, il a été entraîné sur un corpus (French Treebank) et incorpore une recherche par faisceau[9].

3.3.1.4 Le faisceau

Durant son analyse en cascade (découpage en phrase, token, taggage, parsing) les analyses, faites selon la probabilité à un niveau bas, ne sont plus forcement pertinentes à un niveau haut. Pour éviter de traiter toutes les configurations possibles, l'analyse en faisceau (beam search) limite le nombre de configuration à examiner. Un faisceau de largeur élevée aura plus de chance de trouver la meilleure configuration, mais devra exécuter plus de calculs. Les modules se transmettront un nombre de configurations correspondant à la longueur du faisceau avec leurs probabilités associées.

7. Pour plus d'information voir : [Kow *et al.*, 2006], [Vanrullen *et al.*, 2006]
8. Pour plus d'information voir : [Nasr, 2004], [Kübler *et al.*, 2009], [Alfared *et al.*, 2011]
9. Pour plus d'information sur Talismane voir : [Urieli et Tanguy, 2013]

3.4 Résumé

Les corpus utilisés par l'équipe Gestion des connaissances sont issus du cours de "Techniques et Technologies des Véhicules Spatiaux" (TTVS). Les parties du TTVS composant le corpus sont sélectionnées par un expert.

À partir de ce corpus nous créons un taxinomie du domaine avec l'expert. Cette taxinomie sert pour le classement automatique des documents en mémoire d'entreprise et l'organisation de la base de connaissance (rappel chapitre 2.3).

Pour enrichir la base de connaissance et améliorer le classement automatique de documents nous avons besoins des informations contenues dans le corpus. Pour cela nous extrayons des listes de vocabulaire grâce à l'extracteur terminologique Talismane. Pour assurer la continuité avec le découpage taxinomique réalisé précédemment nous fournissons les parties du corpus correspondantes à Talismane.

Une fois les listes de vocabulaire extraites nous devons contrôler leurs qualités et retirer le bruit à l'aide d'une nouvelle interview avec un expert. Cette partie sera abordée lors du chapitre 4.

4 Comment déterminer qu'un terme est adéquat à une catégorie ou au domaine ?

Ce chapitre va porter dans un premier temps sur les travaux que j'ai effectué au CNES en expliquant les problèmes rencontrés et les solutions apportées lors de la création de listes présentables à un expert. Il sera suivi d'une explication sur comment les listes créées nous servent pour créer des filtres pour le classement automatique dans la base de connaissance tout en abordant la problématique de ce mémoire. Une troisième partie parlera des méthodes de présentation des entrées des listes créées à un expert. Ce chapitre se terminera sur quelques recommandations d'ordre général pour l'amélioration des procédures d'enrichissement des ontologies et de rangement des bases de connaissance

4.1 Cas traités

4.1.1 Cas simple : Rappel de la méthodologie

Les extractions au CNES se passent d'ordinaire en collaboration avec un expert sur son domaine d'expertise. Grâce à une interview menée par l'équipe de Gestion des Connaissances (GC) auprès de l'expert, les parties du cours de "Techniques et Technologies des Véhicules Spatiaux" (TTVS) étant en relation avec le domaine d'expertise, sont identifiées et serviront à constituer le corpus (voir chapitre 3.2).

Le corpus est alors découpé en classes sur la base d'une analyse statistique. Le CNES utilise l'outil statistique Alceste (voir chapitre 2.3) qui réalise une classification hiérarchique ascendante des textes soumis à l'analyse. Les classes extraites sont alors soumises à l'expert qui apporte sa connaissance pour les nommer ou les fusionner.

Les extractions terminologiques se déroulent à partir des classes ainsi constituées, grâce à l'extracteur terminologique Talismane (voir chapitre 3.3).

4.1.2 Cas complexe

Les cas qui ont fait l'objet de mon stage étaient plus complexes que les cas simples présentés ci-dessus. Ainsi deux cas se sont présentés :

Le premier, sur les Nacelles Pointées (domaine des Ballons) présentait la difficulté de ne comporter aucun document didactique. De plus, ce domaine n'est pas couvert par les cours du TTVS. Cette absence de document didactique s'explique par le fait que le domaine est nouveau et innovant. Il reprend cependant des techniques de nombreux autres domaines et les experts travaillant dessus n'ont pas encore eu le temps de rédiger des documents destinés à la formation.

Le second portait sur les antennes/radars. Il présentait la difficulté d'être un domaine très large et ayant de nombreuses spécialités : antennes actives, antennes passives, antennes à hélice, antennes à faisceaux fixes, antennes à faisceaux reconfigurables, antennes phasées, radar, télédétection, ... recouvrant un champ lexical commun important. Le TTVS couvre en partie seulement certaines de ces spécialités d'une part, mais d'autre part l'organisation des connaissances correspondantes ne suit pas nécessairement l'organisation en métiers adoptée par le CNES.

Il a donc fallu dans les deux cas utiliser de grands corpus constitués à partir de documents autres que le TTVS. Ces corpus ont été alimentés lors d'interviews en récupérant les documents représentatifs du domaine selon l'expert. Outre les problèmes liés aux différents genres de documents récupérés (retours d'expériences, notes techniques, documents de spécifications, assurance et qualité produit, publications CNES, brevets, ...), la taille des corpus a engendré des problèmes d'extraction avec Talismane.

Nous n'aborderons que le cas du domaine des Nacelles pointées. Le domaine des radars venait de commencer à être traité lors de mon début de stage au CNES. J'ai pu participer aux interviews servant à la composition du corpus tel que présenté dans le chapitre 3 et au découpage taxinomique du domaine à l'aide de l'outil Alceste tel que présenté dans le chapitre 2.3.

4.1.3 Exemples de problèmes rencontrés

Le nombre de syntagmes nominaux extraits par Talismane est très élevé par construction : le logiciel est capable d'extraire des syntagmes simples tels que « radar » ou plus complexes tels que « radar VARAN » ou « radar VARAN à antenne synthétique de longueur 1 m avec une résolution de l'ordre de 3 m ». On peut contrôler le nombre de syntagmes en imposant une fréquence minimale d'apparition.

Dans sa conception initiale le logiciel ne permettait que de contrôler la fréquence des têtes des syntagmes (« radar » dans les exemples donnés) et non la fréquence de leurs expansion (« radar VARAN », « radar VARAN à antenne synthétique » etc). Dans ces conditions une très forte fréquence imposée sur une tête ne contraignait que peu le nombre de ses expansions.

Lors des extractions sur les nacelles pointées, l'extracteur Talismane n'a pas réussi à supporter la taille importante des corpus. Ce dysfonctionnement a engendré des listes de syntagme très grandes (plus de 120 000 entrées) et mal construites. Les termes techniques et les syntagmes étaient souvent répétés de plus en plus tronqués jusqu'à leur tête.

La conception fonctionnelle de l'interface a donc dû être modifiée : offrir la possibilité de contrôler la fréquence de tous les syntagmes calculés par Talismane quelles que soient leur

longueur.

4.1.4 Méthodes utilisées

Les listes de syntagmes correspondant au découpage d'Alsceste et extraites pas Talismane sur le domaine des Nacelles Pointées sont donc inexploitables en l'état, il faut les nettoyer.

Pour que les syntagmes extraits par Talismane puissent être soumis à l'expert avant d'être codés sous forme d'une ontologie, il faut que cette liste ne dépasse pas 500 entrées, et que les syntagmes soient bien formés. Pour nettoyer ces listes nous avons recouru à plusieurs moyens.

4.1.4.1 Travailler sur la fréquence

Le premier moyen employé fut d'utiliser la fréquence dans le corpus des entrées de la liste. Cependant, avec autant d'entrées, pour réduire la liste à environ 500 entrées, il fallait mettre un seuil sur la fréquence très élevée (d'environ 50 occurrences).
Travailler sur la fréquence a l'avantage de ne garder que les syntagmes les plus représentés du corpus, par conséquent ceux qui sont également les plus représentés dans les autres textes du domaine en question. Toutefois l'utilisation de la fréquence possède l'inconvénient de ne pas faire ressortir de nombreux concepts et termes intéressants, du fait de leur fréquence inférieure au seuil déterminé.

Le seuillage en fréquence peut masquer des termes pouvant refléter les techniques innovantes ou des concepts émergents qui ne sont pas encore très intégrés dans les textes du domaine malgré leur importance. Il en est de même pour les termes reflétant les concepts et techniques "obsolètes", c'est-à-dire les termes qui étaient en usage dans le passé mais qui ne sont plus d'actualité. Ces termes seraient pourtant présents dans un corpus didactique. En perdant ces termes, on perd aussi un moyen de classement des documents anciens.

De plus, le fait d'avoir différents genres textuels tels que les documents de spécifications, où des boilers plates [1] (ou gabaries en français), fausse l'utilisation de la fréquence comme moyen de discrimination. Car les documents de spécifications se servent de phrases type, répétant le même vocabulaire de nombreuses fois.

La discrimination sur la fréquence n'était pas, dans les cas présents, le moyen le plus adapté, c'est pourquoi je fus amené à tester d'autres moyens.

4.1.4.2 Détection des majuscules

Le second moyen testé fut de rechercher les syntagmes comportant des majuscules. Pour réduire le nombre d'entrées des listes, j'ai voulu tester une approche "naïve" en partant du principe que les mots ayant une majuscule était des mots qui était mis en valeur (noms propres, sigles et acronymes, ...) et donc important. La réalisation fut faite grâce à un script ne sélectionnant que les entrées ayant au moins une majuscule. Cette méthode s'est avérée

1. phrases types à remplir

peu fructueuse, car elle sélectionne encore trop d'entrées (+- 30 000 entrées) souvent mal formées. Elle présente l'inconvénient majeur d'éliminer les groupes nominaux bien formés et pertinents au domaine.

En voulant améliorer la technique de repérage des mots en majuscule nous avons mis au point un nouveau script. Ce script parcourt les phrases retenues précédemment et cherche à ne sélectionner que les expressions commençant par un mot ayant une majuscule et finissant par un mot ayant une majuscule (voir figure 4.1). Une fois arrivé en bout de syntagme, on retire le mot en tête ayant une majuscule et on recommence. Chaque instance est alors rangée dans une table de hachage ainsi que sa fréquence.

Si à une fréquence égale, une des deux entrées est plus petite (en nombre de mots) et est compris dans l'entrée la plus grande, alors nous supprimons l'entrée la plus petite.

Exemple :
- Mot (Fréquence = 1)
- Mot de Passe (Fréquence = 1)

Le terme "Mot" n'est jamais utilisé en dehors de "Mot de Passe"

Au contraire, si une courte entrée est comprise dans plusieurs entrées plus longues, alors elle est intéressante, car elle représente une tête de syntagme.

Exemple :
- Mot (Fréquence = 2)
- Mot de Passe (Fréquence = 1)
- Mot clé (Fréquence = 1)

Le terme "**Mot**" est utilisé en en tant que tête de "**Mot** de Passe" et de "**Mot** Clé"

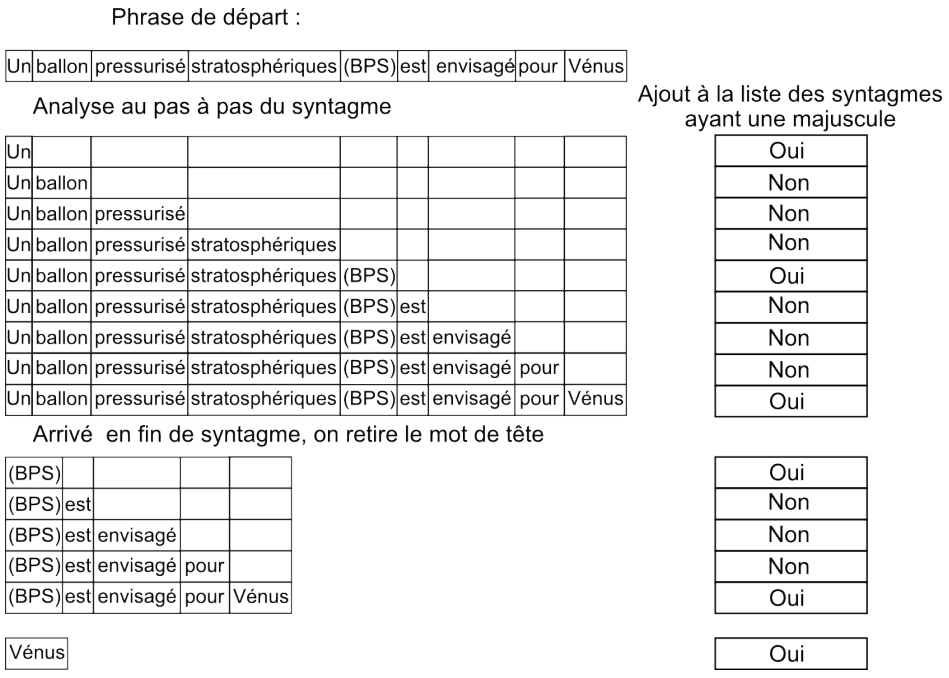

Figure 4.1 — Exemple de l'analyse d'un syntagme contenant des majuscules

Il y a eu une grande réduction de la taille de la liste, mais cette méthode a engendré trop

de bruit pour être exploitable.

4.1.4.3 Utilisation des têtes

En partant du constat que la méthode précédente a engendré trop de bruit et a éliminé des syntagmes bien formés potentiellement pertinents au domaine, j'ai essayé une troisième méthode. Cette méthode cherche à réutiliser les moyens mis en œuvre précédemment tout en éliminant les syntagmes mal formés. Dans cette méthode nous mettons de coté la recherche de mots ayant au moins une majuscule pour nous concentrer sur analyse des têtes des syntagmes. Cette approche se justifie par le fait que tous les groupes nominaux sont (en français) classés par leur tête (ex. satellite d'observation, satellite défilant, satellite géostationnaire, etc.) Nous avons alors modifié le script précédent pour détecter les têtes des termes et comparer leurs expansions.

phrase de départ				rajout à la liste des syntagmes
un	deux	trois	quatre	
analyse au pas à pas dans la phrase				
un				Oui
un	deux			Oui
un	deux	trois		Oui
un	deux	trois	quatre	Oui

Figure 4.2 — Principes de fonctionnement de recherche sur les têtes de syntagmes

Comme représenté dans la figure 4.2 le syntagme est découpé à partir de sa tête en ajoutant à chaque passage un token. L'objectif étant, comme dans la méthode précédente, d'obtenir une liste de termes et de syntagmes. Puis de comparer les éléments de la liste pour obtenir une fréquence pour chaque entrée. Ce qui permet de sélectionner les syntagmes ou termes importants (voir exemple de la méthode précédente 4.1.4.2).

Cette méthode est très lourde. Elle fonctionne sur un panel d'essais réduit mais ne fonctionne pas sur la liste extraite par Talismane de 120 000 entrées. Un espace de stockage RAM d'un ordinateur normal ne permet pas de la mettre en place.

Le script se heurtait au problème de la taille du corpus.

4.1.4.4 Taille des syntagmes

Enfin, le dernier moyen utilisé consiste à jouer sur la longueur, en nombre de mots [2], des syntagmes et termes. En effet, de nombreux syntagmes produits par l'extracteur Talismane étaient composés d'au moins 10 à 20 mots. En limitant le nombre de mots possibles dans les syntagmes et termes, on parvient à éliminer une grande majorité des syntagmes mal formés et on limite efficacement le nombre d'entrées sélectionnées.

Exemple :

2. un "mot" est ici pris au sens informatique : un token entre deux espaces

– *station de contrôle et de supervision* = 6 mots
– *connecteur db 15 femelle* = 4 mots

Une phrase entière et bien formé est peu intéressante pour devenir une entrée dans la base de connaissance et est peu pertinente pour permettre un classement automatique. C'est pourquoi définir un seuil sur le nombre de mots des entrées est intéressant.

4.1.5 Solution adoptée

Comme solution nous avons utilisé un seuil sur le nombre de mots des entrées pour réduire la liste à des entrées intéressantes et nous avons utilisé un seuil sur la fréquence d'apparition de cette liste réduite pour éliminer les syntagmes et termes mal formés.

En utilisant à la fois un seuil sur les fréquences d'apparition des entrées et leurs tailles en nombre de mots, une liste réduite de 600 entrées a pu être constituée sans avoir à utiliser des fréquences élevées pour le seuillage.

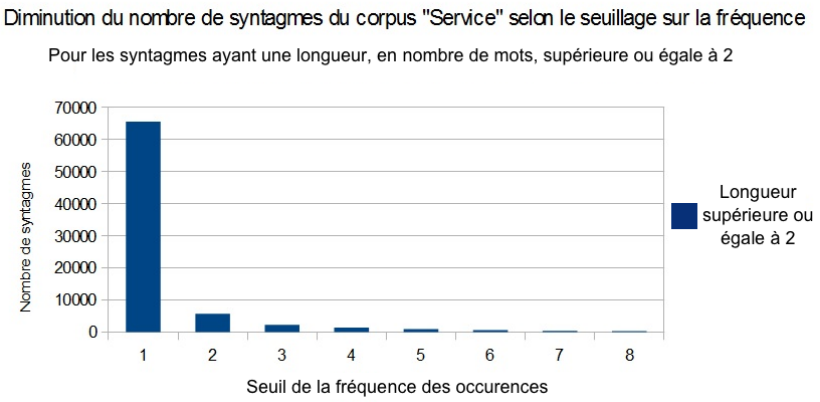

Figure 4.3 — Application de filtres sur une liste de syntagmes (1)

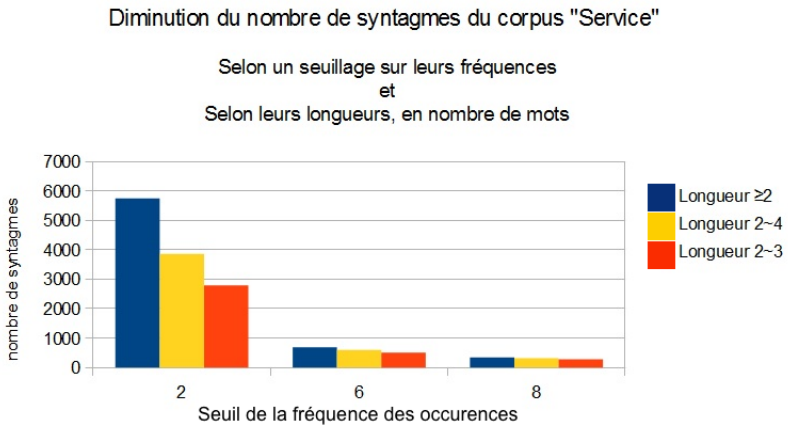

Figure 4.4 — Application de filtres sur une liste de syntagmes (2)

La fréquence et la taille retenues ont été de 6 en fréquence et d'une longueur entre 2 et 3

mots. La liste ainsi obtenue comportait certes, toujours des termes mal formés, mais au vu de sa taille, un nettoyage manuel était envisageable.

Ce travail automatisé sur le corpus des Nacelles Pointées à été réalisé manuellement en parallèle par la terminologue du CNES. C'est le travail de cette terminologue qui a été retenu pour continuer à travailler sur le domaine des Nacelles Pointées.

Il est intéressant de remarquer que les deux listes présentaient de très nombreux termes communs. Mon travail servira éventuellement au service de la Gestion des Connaissances lors d'un cas futur similaire au domaine des Nacelles Pointées.

4.2 La sélectivité

Le rangement de la mémoire d'entreprise se fait via la plateforme ATIC (rappel ATIC sert également comme plateforme pour le moteur de recherche sur la mémoire d'entreprise, voir chapitre 1.5). La mémoire d'entreprise est structurée selon les métiers et projets du CNES (rappel chapitre 2.2.2). Ces métiers et projets sont découpés en sous-domaines suite à un dialogue entre un expert apportant sa vision de la taxinomie du domaine et un terminologue apportant le découpage du corpus du domaine fait statistiquement par Alsceste (rappel chap 2.3). Lorsqu'un document est versé en mémoire d'entreprise il faut qu'il arrive dans le domaine, sous-domaine où catégorie lui correspondant.

Les Listes obtenues précédemment seront, une fois nettoyées, présentées à un expert pour pouvoir déterminer quels sont les termes appartenant au domaine et aux découpages en sous-domaine. Les termes ainsi choisis seront rajoutés à la base de connaissance.

Ces listes servent également pour l'enrichissement des filtres permettant le rangement automatique des documents versés en mémoire d'entreprise. La sélection de termes permettant ce filtrage est donc importante et sera développée dans les paragraphes suivants.

4.2.1 Pourquoi sélectionner ?

Une fois une liste de syntagmes obtenue, il faut déterminer la sélectivité des syntagmes retenus au regard des sous-domaines et du domaine d'expertise traité.

En effet, la constitution du corpus à partir de documents du domaine en question ne garantit pas que les syntagmes extraits lors de la création des listes ne soient pas utilisés dans d'autres domaines « métiers » du CNES (voir figure 4.5). On peut prendre en exemple les unités de mesure telles que la masse, la taille, le volume, ... qui sont utilisées dans de nombreux domaines. De même un terme appartenant à un sous-domaine donné peut se retrouver dans un sous-domaine voisin. (voir figure 4.6)

Pour pouvoir déterminer la sélectivité des termes d'un domaine vis-à-vis d'un autre domaine ou d'un sous-domaine, il faut mener à bien des analyses contrastives entre le corpus qu'on a constitué et d'autres corpus disponibles, au premier rang desquels celui du TTVS[3].

Ces études contrastives serviront à la sélection de termes employés en tant que filtre de

3. Techniques et Technologies des Véhicules Spatiaux

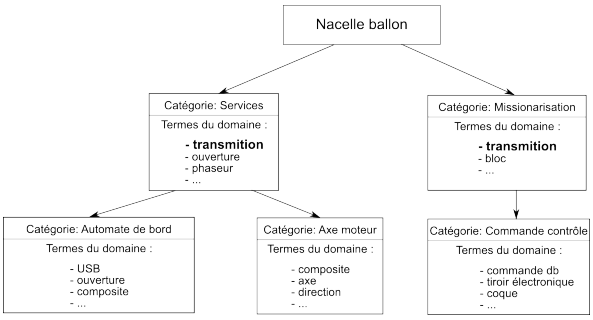

Figure 4.5 — Un terme utilisé dans le domaine "Services" peut aussi apparaitre dans le domaine "Missionarisation"

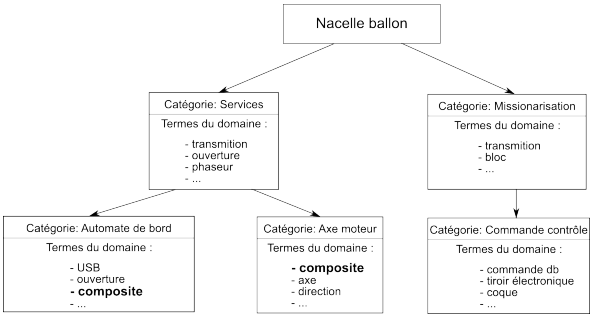

Figure 4.6 — Un terme utilisé dans le sous-domaine "Automate de bord" peut ne pas être utilisé par un sous-domaine voisin, "Axe moteur"

classement des documents dans la mémoire d'entreprise.

4.2.1.1 Classement automatique de la mémoire d'entreprise

Le moteur de recherche, associé à la plateforme ATIC, de la mémoire d'entreprise permet à l'utilisateur de rechercher selon les domaines et sous-domaines du CNES (voir chapitre 1.5). Pour cela, il faut que les documents de la mémoire d'entreprise soient classés. ATIC possède un système de classement percolant [4] des documents. ATIC classe les documents à l'aide la reconnaissance des mots-clés servant au filtrage dans le contenu des documents. Lorsqu'un document est déposé à la racine d'ATIC, il descendra dans l'arborescence du classement selon les mots-clés qu'il possède. S'il ne possède pas de mot-clé pour descendre à un classement plus bas, il restera au niveau où il est situé actuellement, même si ce niveau est un hyperonyme. (voir figure 4.7)

C'est pourquoi il est important de distinguer les filtres de chaque niveau de l'arborescence de rangement, des domaines aux sous-domaines et aux catégories inférieures. Avant mon stage au CNES, les filtres du domaine comprenaient tous les filtres des sous-domaines en plus des termes couramment utilisés appartenant au domaine. Grâce à mes observations sur la sélectivité, cette façon de faire a changé.

4. Les documents présents à un niveau N, éligibles à un niveau N+1 disparaissent du niveau N et se retrouvent au niveau N+1.
C'est à dire que les documents peuvent basculer d'un domaine à un sous-domaine, ou d'un sous-domaine à un niveau plus bas. C'est une distribution en cascade, où les documents descendent dans l'arborescence depuis la

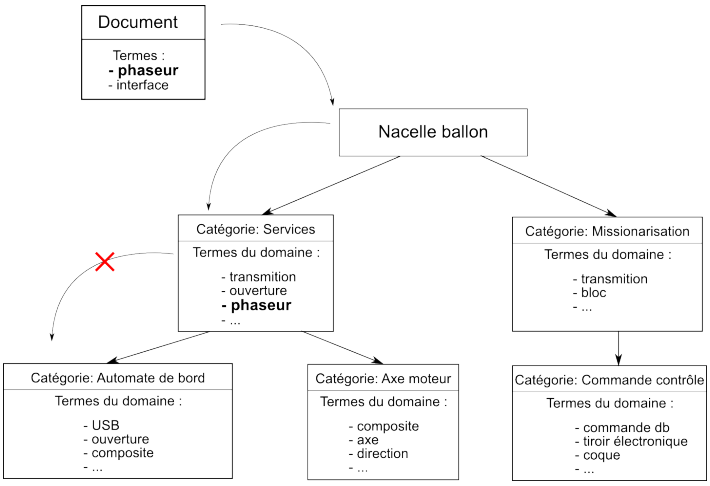

Figure 4.7 — Percolation : le document s'arrête au niveau 1 (Service) et ne descend pas plus bas

Pour augmenter l'efficacité du basculement des documents à classer, il n'y a pas besoin que les filtres du niveau supérieur soient les mêmes que ceux des niveaux inférieurs (voir figure 4.8).

Un terme spécifique à un sous-domaine peut permettre d'augmenter la précision de distribution du document, alors que ce même terme utilisé au niveau supérieur aurait engendré du bruit en ramenant des documents n'appartenant pas au domaine (voir figure 4.9 et 4.10). Cela s'explique par le fait que les documents, une fois arrivés à un niveau, sont déjà identifiés comme appartenant au domaine correspondant à ce niveau. Les critères pour déterminer s'ils appartiennent au domaine sont alors remplis.

4.2.1.2 Le scoring

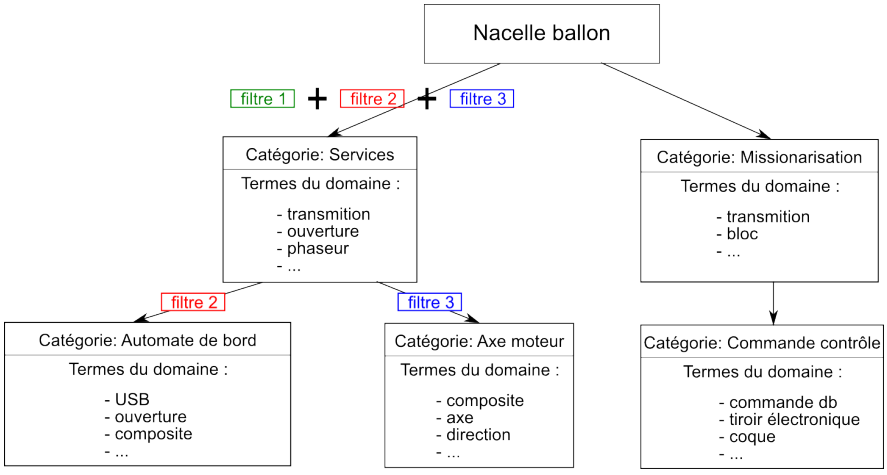

Figure 4.8 — Ancien système de filtre

Actuellement les documents sont distribués dans l'arborescence selon les mots-clés

racine vers une catégorie de plus en plus précise.

qu'ils possèdent. Ce qui représente un système de décision de distribution booléen. Si aucun mot-clé n'est présent le document ne bascule pas et reste au niveau où la décision a eu lieu. Si un mot-clé est présent il y a basculement du document dans tous les niveaux inférieurs de l'arborescence ayant ce mot-clé comme filtre. Ce qui implique qu'un document peut être référencé dans plusieurs domaines et sous-domaines différents.

Une des suggestions que j'ai proposé est d'utiliser un système de décision basé sur un score attribué aux documents pour chaque niveau inférieur. Ce système de score, où scoring, attribue au document un score plus ou moins élevé pour chaque niveau inférieur en fonction du nombre de mot-clé correspondant entre leurs filtres et le contenu du document. Ce score permettrait de basculer (ou non) les documents dans un niveau inférieur unique.

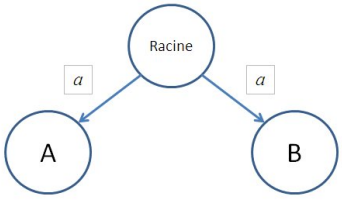

Figure 4.9 — Terme "*a*" non exclusif aux domaines **A** et **B**

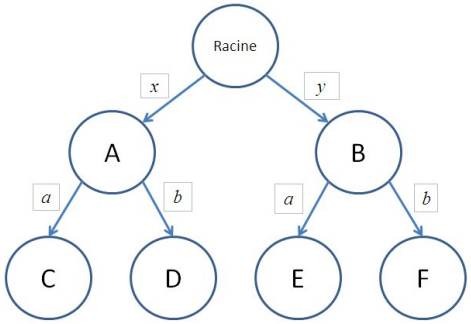

Figure 4.10 — Le terme "*a*" est exclusif aux sous-domaines **C** et **E**

On peut se représenter l'arborescence de classement comme une chaîne de Markov, les documents passent d'état (niveau de l'arborescence) en état, les transitions se passent alors par les filtres (termes clés) et l'état précédent n'influence pas l'état suivant. (voir figure 4.11 [5])

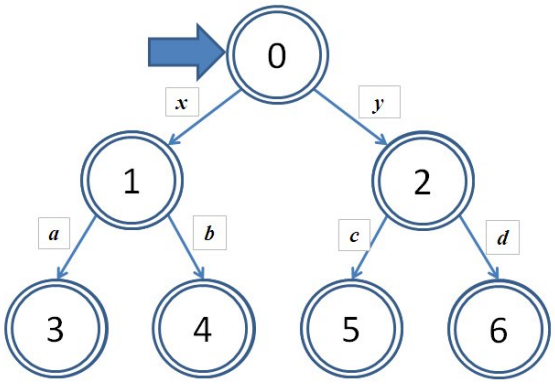

Figure 4.11 — Représentation de l'arborescence par un automate

Chaque état est un état final à cause de la percolation. De l'état **1** on ne peut que (i) passer à l'état **3**, (ii) passer à l'état **4** ou (iii) sortir. Les recherches à partir de l'état présent ne sont

5. Pour schématiser j'ai représenté l'arborescence sous forme d'un arbre binaire, bien entendu un domaine peut avoir plus de deux sous-domaines et un sous-domaine peut lui même avoir des niveaux inférieurs, voir figure 1.4

pas rendues plus précises par des éléments d'informations supplémentaires concernant le passé. C'est-à-dire, qu'à l'état **1**, le choix entre la transition "*a*" ou la transition "*b*" n'est pas influencé par le choix précédent de la transition "*x*".

Exemple : Soit un document D contenant les mots-clés <aabaabxxxybccccccccccccc>
- Ce document en partant de l'état **0** aura un scoring de trois "*x*" pour un "*y*". Il prendra donc la transition *x* pour aller en état **1**.
- D à un scoring à l'état **1** de quatre "*a*" pour deux "*y*". Il prendra la transition "*a*" pour arriver à l'état **3**.
- En partant de l'état **1**, le document D n'a pas de chemin pour aller à l'état **0, 2, 5 ou 6**.
- Si D n'avait eu ni de "*a*" ni de "*b*", il serait resté à l'état **1**.

4.2.2 Études contrastives

Une étude contrastive correspond à la recherche de concordance d'un ensemble de termes donnés par rapport à un corpus précis dans le but d'explorer un phénomène linguistique. La concordance est, pour un terme donné, le répertoire des exemples d'utilisation de ce terme dans un corpus. Les concordanciers sont des outils capables de rechercher dans des textes les concordances et d'extraire de nombreuses autres informations statistiques.

Les études contrastives étaient exécutées au CNES avec le concordancier AntConc, pour chaque terme. AntConc est un logiciel en libre accès disponible sur internet et a été développé par Laurence Anthony [6], professeur à l'Université de Waseda (Japon).

Lors du stage nous avons automatisé ce processus pour pouvoir prendre en entrée une liste de termes et extraire la fréquence de ces termes dans un corpus cible ainsi que leurs contextes. Le corpus cible varie en fonction de ce que l'on veut contraster. Les différences de fréquences d'utilisation des termes dans le corpus étudié et le corpus cible sont notre principal critère de sélectivité. La liste des termes potentiellement caractéristiques étant à notre disposition par le travail d'extraction précédent, l'obtention de la fréquence d'utilisation des termes de la liste dans chacun des corpus est aisé.

Lors de la détection d'un terme, pour compter sa fréquence d'apparition, nous extrayons également une partie de la phrase dans lequel il est employé afin de pouvoir vérifier le contexte d'utilisation. Ce script permet une automatisation partielle. Il fonctionne à l'aide d'expressions régulières et apporte un gain de temps significatif.

4.2.2.1 Termes caractéristiques d'un domaine

Pour savoir si un terme est caractéristique à un domaine il faudrait pouvoir faire une analyse en concordance, en prenant en compte tous les autres domaines du CNES. Ce qui est impossible, car un corpus regroupant tous les autres domaines du CNES, excluant le domaine étudié courant, n'existe pas. S'il existait, nous aurions une différente stratégie d'extraction des connaissances pour l'enrichissement de la base de connaissance.

Nous avons eu recours au corpus didactique du TTVS recouvrant essentiellement les domaines du CNES relatifs aux satellites et mentionnant les termes les plus courants de chaque

6. `http://www.antlab.sci.waseda.ac.jp/`

domaine. Grâce au script de concordance présenté plus haut, nous avons pu mener une étude contrastive en faisant ressortir comme contexte les chapitres et modules où les termes apparaissaient. À ce corpus a été rajouté, pour le domaine des Nacelles Pointées, le corpus de "l'informatique bord", domaine voisin au sous-domaine des Services. Cet ajout permet de différencier deux domaines *a priori* proches. Les termes sélectifs du domaine peuvent être des termes sélectifs des sous-domaines également. Dans ce cas-là ils apparaîtront aux deux niveaux de filtrages différents. (voir figure 4.12)

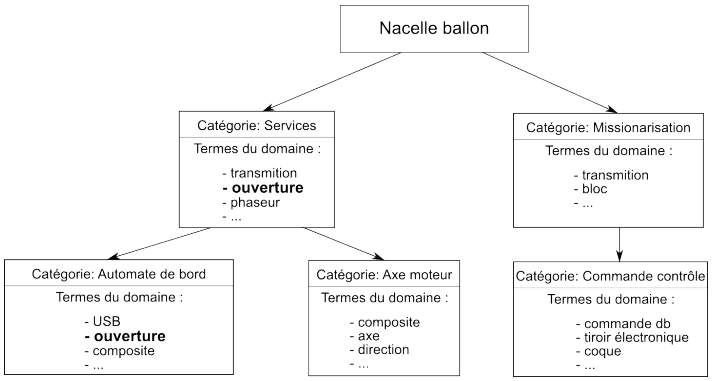

Figure 4.12 **—** Le terme a est sélectif de deux niveaux

Si les termes sont spécifiques au domaine, mais pas aux sous-domaines, ils ne seront présents qu'au niveau de filtrage le plus haut, et seront considérés comme des termes généraux.

4.2.2.2 Termes caractéristiques d'un sous-domaine

Pour savoir si un terme est spécifique à un sous-domaine, il faut l'analyser en concordance avec les corpus des autres sous-domaines étant au même niveau dans l'arborescence et appartenant au même domaine. Cela permet de se rendre compte de son exclusivité à un sous-domaine ou non.

Dans le cas où le terme n'apparaît pas dans d'autres sous-domaines de même niveau, il est alors caractéristique du sous-domaine dont il est issu et peut servir comme filtre pour celui-ci. Cette sélectivité est due au fait qu'une fois qu'un document est basculé au niveau du domaine, la distribution (ou non) de ce document aux sous-domaines se passe à l'intérieur de ce domaine. Ce qui veut dire que le document ne peut pas être basculé vers des sous-domaines n'appartenant pas au domaine le contenant. Par conséquent, si, dans un domaine, un terme est exclusif à un sous-domaine il permet alors de le caractériser et lorsqu'un document arrive au niveau du domaine il peut servir de critère de filtrage pour le faire basculer au niveau du sous-domaine si le document contient le terme.

Alors, un terme couramment utilisé par d'autres domaines peut devenir sélectif d'un sous-domaine spécifique. C'est ainsi que les termes "pointage", "axe de tangage" ou "filtre de Kalman" ne sont pas exclusifs au domaine des Nacelles Pointées, mais à l'intérieur de ce domaine ils sont exclusifs au sous-domaine "Missionnisation, commandes contrôle" par rapport aux autres sous-domaines de même niveau. (rappel : figure 4.9 et 4.10)

Il se peut qu'un terme appartienne à deux sous-domaines différents. Actuellement Le

terme est retenu en fonction de sa pertinence vis à vis des deux sous-domaine et du nombre de sous-domaine du domaine (un terme commun à deux sous-domaines sur deux n'aura pas la même valeur qu'un terme commun à deux sous-domaines sur quinze). Si le terme est retenu comme filtre pour les deux sous-domaines, le document serait versé dans l'un et l'autre.

Si on tient compte à nouveau du mécanisme de scoring et en considérant qu'il y a également d'autres sous-domaines où il n'est pas représenté, on peut décider de conserver le terme ainsi présent dans les deux sous-domaines. [7]

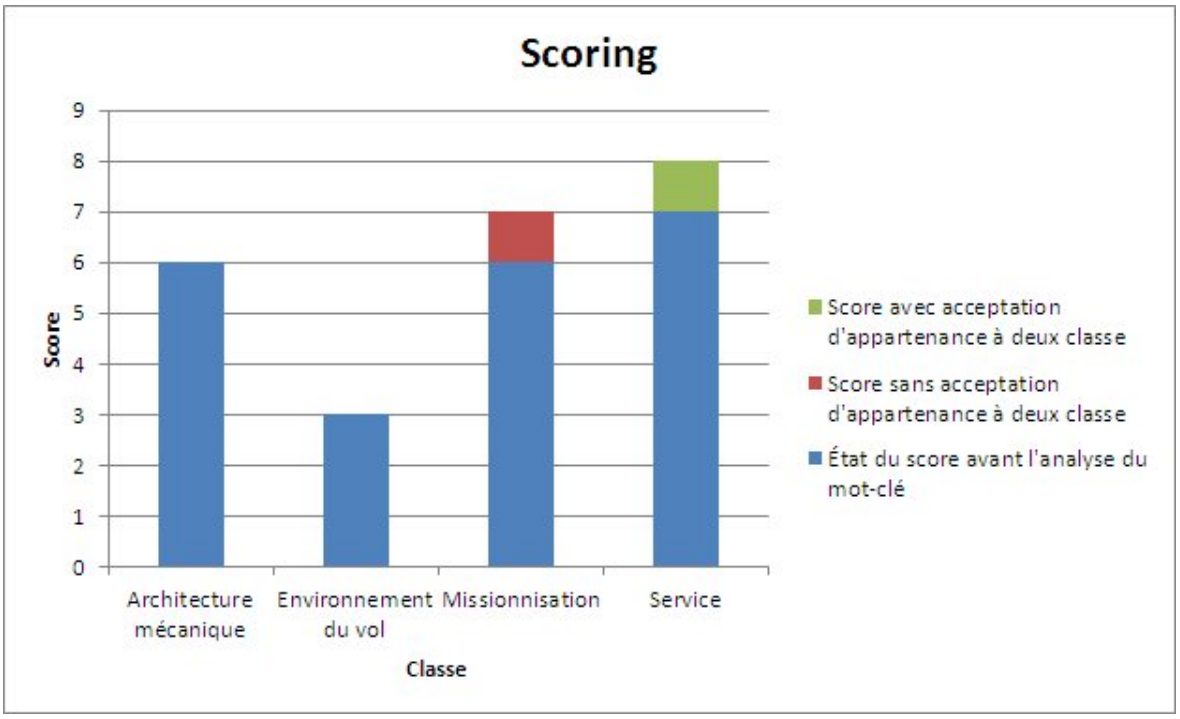

Figure 4.13 — Exemple de distribution d'un score

Les deux sous-domaines voient ainsi leur classement (scoring) se renforcer si le document contient le mot-clé recherché. Cela permet de faire progresser leurs positions dans le classement par rapport aux autres sous-domaines. (voir figure 4.13)

Si le document obtient un score égal pour deux sous-domaines, on peut envisager plusieurs systèmes de résolutions. Premièrement, si les scores sont égaux le document peut-être important pour les deux sous-domaines et donc indexé dans les deux. Secondement nous pouvons envisager une mise à l'écart du document pour être consulté et identifié par un expert ultérieurement. Enfin nous pouvons envisager un poids différent pour chaque terme filtrant et un calcul logarithmique du score pour accentuer les résultats.

7. Les études menées n'ont pas rapporté de cas où un terme appartiendrait à plus de deux sous-domaines différents. Cela est peut-être dû au découpage statistique du corpus en sous-corpus par l'outil Alceste.

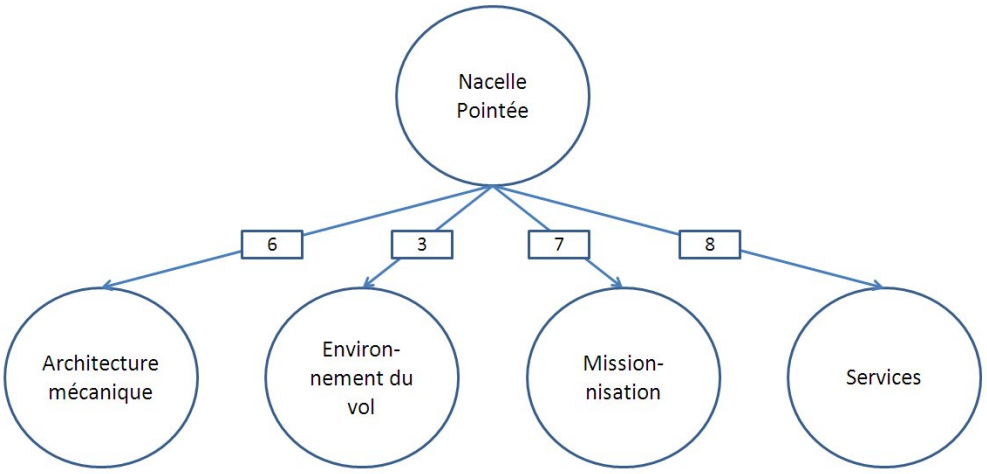

Figure 4.14 — Le document prendra le chemin au score le plus élevé

4.2.2.3 Critère de sélection

Comme vu précédemment les études contrastives ont été automatisées sur les parties qui nous intéressent (fréquence et contexte).

Lorsque l'on se place au niveau du corpus du domaine, on l'étudie en concordance avec le TTVS et si possible avec d'autre corpus estimés proche. Cela permet de faire ressortir les termes propres au corpus.

Pour les sous-domaines, on les étudie en concordance avec le corpus du domaine dont ils sont issus. Cela permet de voir si les termes étudiés sont propres au sous-domaine, à plusieurs sous-domaines ou à tous les sous-domaines. De même les catégories inférieures aux sous-domaines seront étudiées en concordance avec l'ensemble supérieur auquel elles sont rattachées.

La représentativité des termes extraits peut-être calculée. Si on dispose de la fréquence d'apparition d'un terme dans le corpus étudié et un corpus cible nous pouvons en déduire l'indice de représentativité du terme sur le corpus étudié. Nous avons donc ajouté au script de concordance un calcul d'indice de représentativité. L'indice de représentativité va de proche 0, indiquant que le terme n'est presque pas représenté dans le sous-corpus, à 1 indiquant que le terme est exclusif au sous-corpus.

La représentativité correspond au nombre de fois qu'un terme est présent dans un sous-domaine par rapport au nombre de fois où il est présent dans le corpus.
Soit a un terme de la liste des syntagmes appartenant au sous-domaine
$f(a)$: la fréquence de a dans le corpus composé avec l'expert
$f'(a)$: la fréquence de a dans le sous-domaine.

$$\frac{f'(a)}{f(a)}$$

Exemple de calcul :
Si *filtre de Kalman* apparait 15 fois dans le corpus "Nacelles pointées" composé avec l'expert et

15 fois dans le sous-domaine "Missionnisation, commandes contrôle" alors $\frac{15}{15} = 1$, l'indice sera de 1 et indiquera que *filtre de Kalman* est exclusif au sous-domaine "Missionnisation, commandes contrôle".

En prenant comme critère de sélection les termes ayant un indice de représentativité entre 0,75 et 1, nous devrions en théorie améliorer les résultats des filtres permettant le classement automatique. Cependant nous n'avons pas pu tester la valeur de cette fenêtre d'indice en condition réelle. Pour tester cette fenêtre il aurait fallu utiliser un panel d'essai de documents annotés comme appartenant à tel ou tel sous-ensemble et réaliser un rangement automatique en utilisant les termes répondant à la fenêtre de représentativité comme filtres. Ces tests n'ont pu être réalisés par manque de temps.

Nous avons préféré nous servir de cette fenêtre d'indice 0,75 à 1 comme d'un critère supplémentaire et continuer à regarder au cas par cas, selon la fréquence d'apparition dans le corpus étudié et selon sa fréquence d'apparition dans le corpus cible.

4.2.3 En résumé

Pour implémenter des filtres permettant le rangement automatique des documents de la mémoire d'entreprise, il faut sélectionner des termes caractéristiques des catégories de sa taxinomie.

Après les extractions terminologiques avec Talismane et un nettoyage post traitement (vue en partie 4.1), nous obtenons une liste réduite de termes pour le domaine et chaque sous-domaine. Les termes de ces listes sont caractéristiques de l'ensemble dont ils sont issus s'il ne sont pas ou peu retrouvables dans d'autres ensembles de même niveau.

Pour cela nous avons recours à des études contrastives (voir partie 4.2). Les sous-domaines sont étudiés en contraste avec le domaine dont ils sont issus. Les domaines sont étudiés en contraste avec les cours du TTVS. Lors de mon stage au CNES j'ai proposé l'utilisation d'un indice de représentativité qui serait un critère quantitatif et permettrait une automatisation du procédé d'étude contrastive.

Nous pouvons résumer les différentes relations de caractérisation entre un terme et le corpus dont il est extrait.

1. **Un terme peut être caractéristique d'un domaine et d'au moins un de ses sous-domaines.**
   Il sera alors gardé comme filtre sur les deux niveaux.
2. **Un terme peut être caractéristique d'un domaine, mais d'aucun de ses sous-domaine.**
   Le terme permet un filtrage au niveau supérieur mais est diffus dans les niveaux inférieurs.
3. **Un terme peut ne pas être caractéristique d'un domaine, mais être caractéristique d'au moins un de ses sous-domaines.**
   Le terme peut être employé comme filtre a un niveau inférieur. Au niveau supérieur il engendrerait du bruit en renvoyant des documents n'appartenant pas au domaine.
4. **Un terme peut n'être caractéristique ni d'un domaine ni de ses sous-domaines.**
   Ce terme ne sera pas extrait lors de l'extraction avec Talismane.

5. **Un terme peut être caractéristique d'un seul sous-domaine.**
   Le terme permettra une très bonne sélection des documents appartenant ou non au sous-domaine.
6. **Un terme peut être caractéristique de plusieurs sous-domaines.**
   Le terme pourra servir de filtre dans certain cas pour le classement automatique des documents dans les deux catégories. Si le nombre de sous-domaines total est trop proche du nombre de sous-domaines pour lequel ce terme est caractéristique, le terme sera inutilisable comme critère de filtrage. (Le terme n'aura pas la même valeur s'il apparait dans deux sous-domaines sur quinze ou s'il apparait dans deux sous-domaines sur trois)
7. **Un terme peut être caractéristique de tous les sous-domaines.**
   C'est le même cas que le cas N°2, le terme ne pourra être utilisable comme critère de filtrage au sein des sous-domaine.

4.3 Stratégie de présentation à l'expert

Après le travail de concordance nous avons deux listes. La liste ayant été obtenue par Talismane servant à l'enrichissement de la base de connaissance et la liste des termes permettant le classement automatique en mémoire d'entreprise obtenue après une étude contrastive. Les résultats obtenus pourraient être exploitables et intégrés dans la base de connaissance d'ATIC ou comme filtre pour la mémoire d'entreprise. Cependant ils peuvent également être entachés d'erreurs dues au corpus de référence choisi.

Pour pouvoir valider les résultats nous recourrons à des interviews avec l'expert pour confirmer que les termes appartiennent bien au domaine. L'expert est chargé de réduire le bruit en retirant les termes qui, bien que présents dans le corpus, ne sont pas représentatifs du domaine (ou simplement anecdotiques, ponctuels). L'équipe de Gestion des Connaissances est capable de dire si un terme est adapté pour le filtrage ou non, mais en tant que néophytes du domaine étudié, nous ne sommes pas à même de juger de la pertinence des termes vis-à-vis du domaine considéré. C'est à l'expert d'en juger.

C'est ainsi que l'on a pu retirer des termes tels que "bouée" du corpus des Antennes/Radar. Le terme s'était glissé dans le corpus via un retour d'expériences traitant d'émission satellite $\longleftrightarrow$ mer en visant des radars portés par des bouées. L'utilisation ponctuelle de ce terme n'en faisait pas un terme du domaine.

Ce point de vue apporté par l'expert permet d'améliorer la constitution de la liste des termes filtrants et d'éviter de rapporter des documents n'appartenant pas au domaine. Il facilite aussi l'enrichissement de la base de connaissance en éliminant des entrées.

Lorsque l'expert s'exprime sur l'appartenance ou la non-appartenance d'un terme il s'exprime souvent à l'aide d'images, par métonymies [8] et synecdoques. [9]. L'équipe de la Gestion des Connaissances profite des explications de l'expert, pour récupérer également les

8. Métonymie : figure rhétorique, procédé par lequel on exprime un concept au moyen d'un terme désignant un autre concept qui lui est uni par une relation nécessaire (la cause pour l'effet, le contenant pour le contenu, le signe pour la chose signifiée).

9. Synecdoque : Est un cas particulier de métonymie. Il s'agit d'une figure de rhétorique qui consiste à

nouveaux termes associés aux termes analysés pour la construction de l'ontologie (base de connaissance). Ces termes ne seront pas utilisés comme filtres, mais pour l'enrichissement de la base de connaissance. Bien que faisant partie d'un entretien ultérieur, les rapports entre les termes sont notés lorsque l'expert donne une explication.

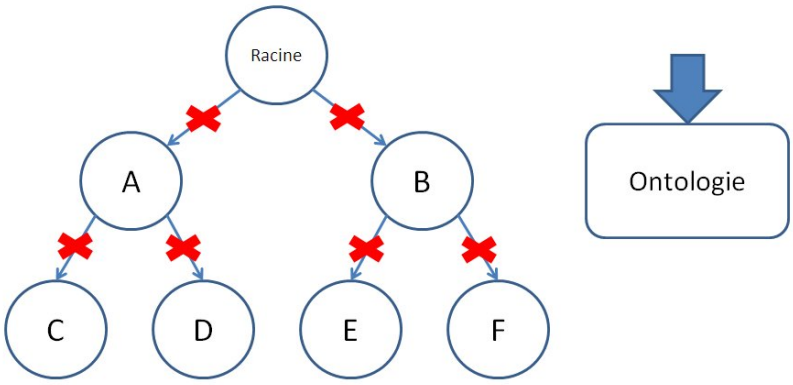

Figure 4.15 — Un terme n'étant sélectif d'aucun niveau est gardé pour la base de connaissance

4.3.1 Comment présenter ?

Pour que l'expert n'ait pas à faire d'efforts cognitifs trop importants en examinant les termes de son domaine, nous avons choisi de présenter les termes regroupés par sous-domaine. L'expert peut alors choisir si un terme appartient à un sous-domaine, au domaine ou s'il n'appartient ni à l'un ni à l'autre.

Pour faciliter le travail de l'expert, nous avons à disposition un panel de contextes pour chaque terme grâce au concordancier Antconc. Car comme le font remarquer [Martinet, 1960, §2.8] : "*Un élément linguistique n'a réellement de sens que dans un contexte et une situation donnés*" et [Leeman, 1996, §23][10], "*C'est l'inscription du mot dans un contexte qui fait surgir son interprétation*". L'expert possède alors toutes les clés pour pouvoir prendre une décision. Autrement dit, il s'agit de tenir compte à chaque fois de la distribution des termes dans leurs contextes.

L'examen des relations sémantiques doit être fait en tenant compte de l'effort demandé à l'expert. Parcourir une liste de 500 termes ou plus, un à un, pour les relier les uns aux autres n'est pas envisageable. Cela est trop fatiguant cognitivement parlant et cela prendrait également trop de temps à un expert pas forcément disponible.

4.3.1.1 Présentation par sous-domaine

Pour examiner de manière intuitive les termes à relier, nous les présentions jusqu'à présent par sous-domaine. En les posant sur un tableau blanc (voir image 4.16), en donnant un

prendre le plus pour le moins, la matière pour l'objet, l'espèce pour le genre, la partie pour le tout (ex : les mortels pour les hommes, le fer pour une épée, une voile pour un navire)

10. Le "sens" et l' "information" chez Harris

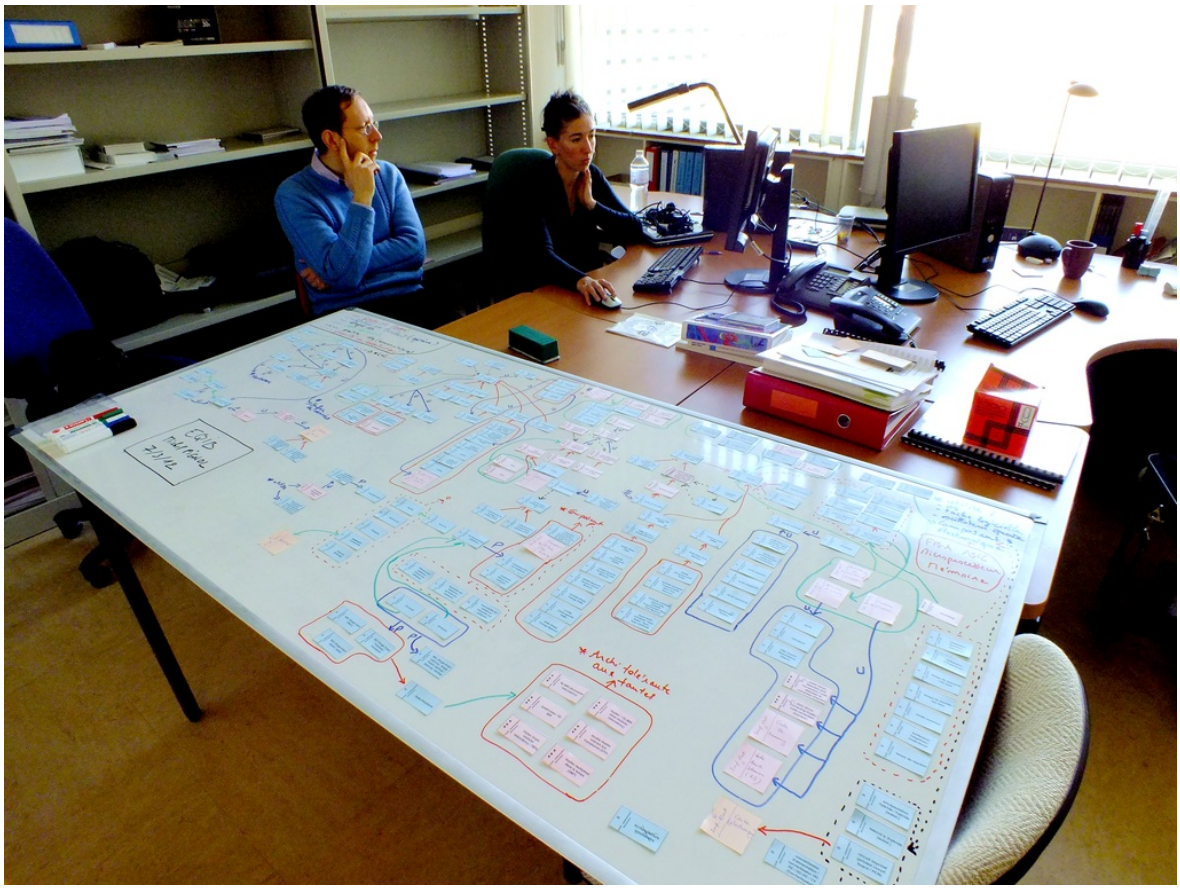

Figure 4.16 — Tableau, cartes et réseaux (1)

nouveau terme (sous la forme d'une carte), nous demandions à l'expert de relier ce terme aux termes déjà présentés, tout en faisant apparaître les liens sémantiques tels que « partie de », « sorte de », « conséquent », ... (voir image 4.17)

En traitant par sous-domaines les termes examinés, nous divisons le nombre de termes à relier entre eux. Ce faisant, nous perdons certainement des relations sémantiques entre termes de différents sous-domaines.

Pour pallier à cela une partie de mon travail a été de proposer des stratégies de présentation. Ces stratégies ont été envisagées pour être intégrées aux futures interviews de la gestion des connaissances.

4.3.1.2 Présentation par significativité

Tout d'abord nous envisageons de présenter les termes sélectifs à la fois d'un sous-domaine et du domaine, puis les termes sélectifs uniquement de l'un des deux, et finalement de présenter les termes non-sélectifs.

Les termes non-sélectifs sont présentés, car même s'ils ne permettent pas de filtrer ils appartiennent au vocabulaire du domaine et peuvent être sémantiquement importants (voir figure 4.15). En présentant les termes de cette manière, nous réduisons aussi le nombre de termes à montrer simultanément à l'expert, nous mettons en valeur notre travail et nous

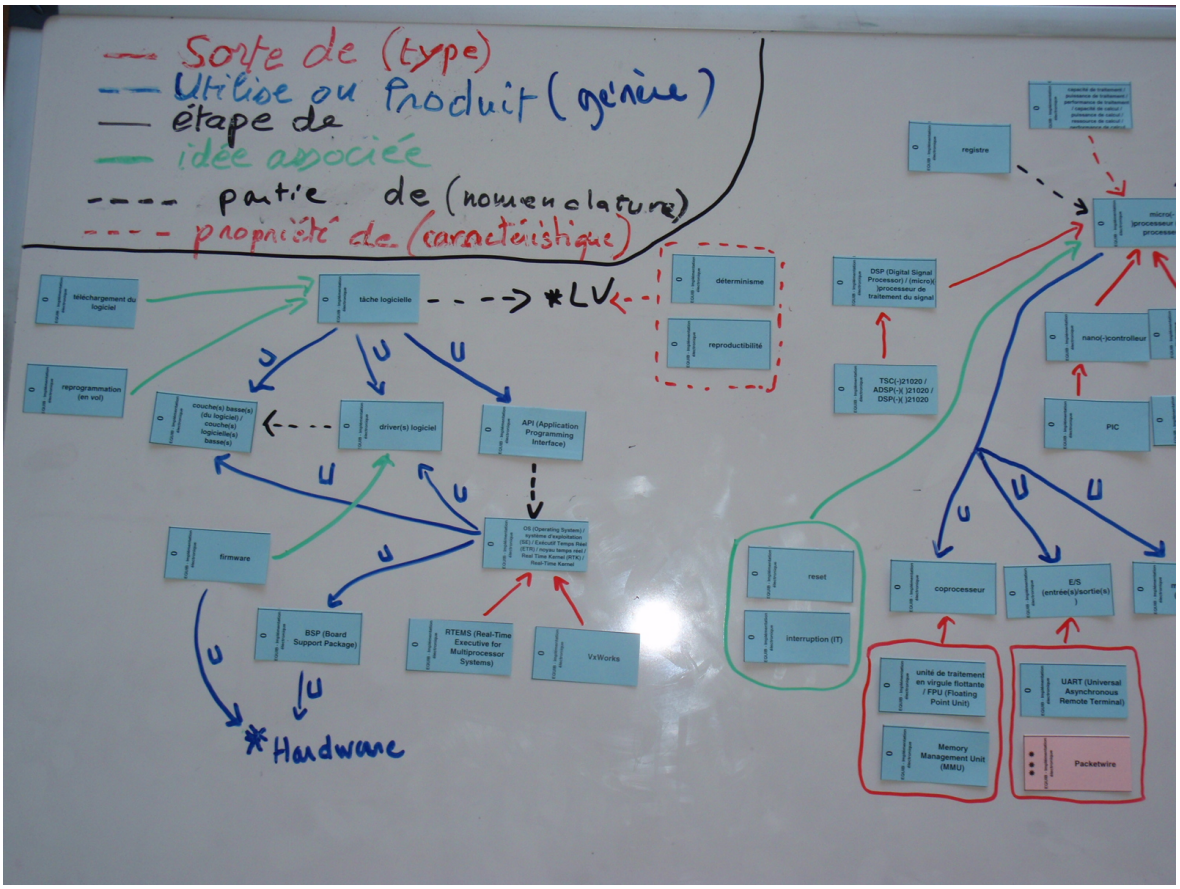

Figure 4.17 — Tableau, cartes et réseaux (2)

abaissons les barrières entre les sous-domaines.

4.3.1.3 Présentation à l'aide de la cooccurrence

La seconde proposition de présentation que nous sommes en train de développer est de regrouper les termes par cooccurrences. C'est-à-dire qu'au lieu de présenter les termes un à un, nous envisageons de présenter des groupements de termes se situant à proximité les uns des autres dans le corpus. Pour cela un script Perl a été développé pour pouvoir détecter les cooccurrents. Cette proximité est calculée selon la fréquence de rencontre entre deux termes à une distance en nombre de mots donnés (mot au sens informatique, un token entre deux espaces).

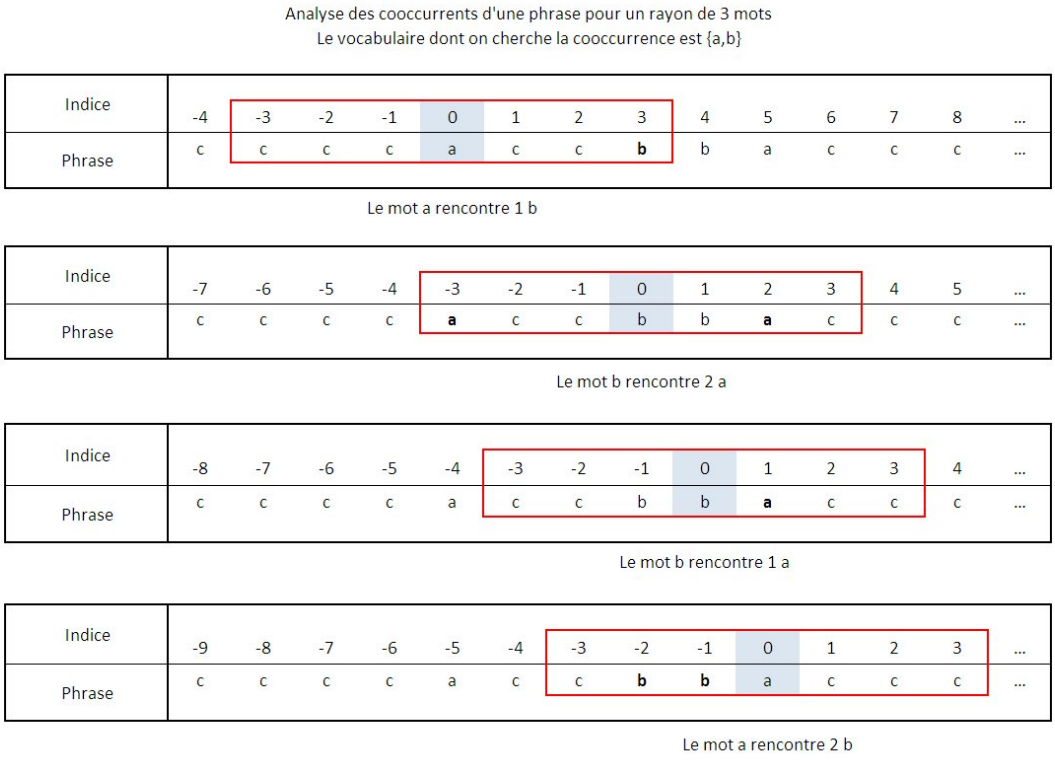

Figure 4.18 — Exemple d'analyse d'une phrase pour la liste de vocabulaire {a,b} avec une fenêtre de 3 mots

On va chercher dans le corpus les coocurrences entre les termes de la liste servant à l'enrichissement de la base de connaissance. Pour chaque occurrence d'une entrée de la liste dans le corpus on comptera une coocurrence avec une autre entrée si elle apparait dans son voisinage. Ce voisinage est déterminé par une distance choisie par l'utilisateur. Une fois que le comptage d'autres entrées de la liste dans la fenêtre a eu lieu, le programme passe à l'occurrence suivante du mot courant. Quand le programme arrive au bout de la liste, il aura compté les coocurrences pour chaque entrée de la liste permettant le regroupement des termes proches.

Cette méthode de regroupement des termes pourra être combinée avec les autres stratégies de présentation.

4.4 Recommandations

Pour pouvoir améliorer l'enrichissement d'ontologies, de nombreuses solutions peuvent être envisagées. Durant mon stage j'ai eu l'occasion de me rendre compte qu'une automatisation plus grande du travail du terminologue peut être faite. Les logiciels de Traitement Automatique des Langues existants actuellement ont tendance à avoir des analyses robustes et spécialisées. On pourrait :

- Améliorer les outils existants :
  - **Alceste**, qui pourrait faire gagner du temps :
    - En permettant de télécharger sous forme de fichier texte les classes qu'il produit.

    - En permettant de fusionner les classes en les déplaçant (clic and drop)
    - En gérant l'analyse de corpus plus grand. (j'ai rencontré des problèmes de taille de corpus sur le domaine des radars)
  - **Talismane**
    - Également en gérant l'analyse de corpus plus grand.
    - En proposant des méta-données comme les coocurrences ou la taille des syntagmes extraits, ...
    - En ayant un entraînement de son étiqueteur et de son analyseur syntaxique sur un corpus spécialisé correspondant au domaine que l'on veut traiter (French Treebank est un corpus journalistique)
- Automatiser la chaîne de traitement en créant un méta-programme, qui lancerait tour à tour les différents logiciels que l'on utilise (Alceste, Talismane, AntConc).
- Automatiser toute la chaîne en faisant un outil qui :
  - Fait le découpage du corpus et attend une validation.
  - Extrait des syntagmes (sur des classes fournies).
  - Rajoute des méta-données (fréquence, analyse contrastive, longueur du syntagme...).
  - Sélectionne une liste de termes à présenter à l'expert grâce aux méta-données.
  - Propose un réseau de cooccurrent / un réseau sémantique.

Développer de nouveaux outils :

- Fusionner les trois logiciels : Un logiciel qui permette de découper un corpus, extraire les termes selon le découpage, donner des informations sur le contexte.
- Développer une plateforme dédiée, permettant des interactions entre l'expert et le terminologue. La possibilité en sélectionnant un terme d'avoir toutes les informations aidant à la prise de décision de l'expert, son réseau sémantique interactif auquel on puisse rajouter, retirer, modifier les liens, ...

Discussion et conclusion

Au fil de ce mémoire nous avons tenté de répondre à la problématique en expliquant aux chapitres 2 et 3 le contexte de création des listes. Au chapitre 4 nous avons dans un premier temps abordé les contraintes du nettoyage et réduction de la taille des listes (chapitre 4.1). Dans un second temps, nous avons cherché au chapitre 4.2 à définir la sélectivité d'un terme pour une catégorie en vue du classement automatique des documents. Nous avons également cherché à montrer la façon actuelle de faire du CNES (chapitre 4.2.1), mon apport (chapitre 4.2.1.1) et les possibilités (scoring) existantes (chapitre 4.2.1.2). Enfin nous avons apporté des éléments de réponse à la problématique sur la pertinence des termes filtrant en abordant les études contrastives au chapitre 4.2.2.

Dans cette dernière partie nous entamerons tout d'abord une discussion sur les ontologies avant de conclure sur mon expérience au CNES.

Discussion

La conception d'une ontologie passe par de nombreuses étapes. Ces étapes sont possiblement automatisables, mais des problèmes peuvent arriver à la suite de la composition du corpus de base. La composition du corpus doit donc être considérée comme l'étape la plus importante. L'exploitation de ce corpus demande une supervision par un expert pour que les résultats soient contrôlés. En effet, comme dit au chapitre 3, le rôle de superviseur de l'expert a deux aspects :

- Déterminer des termes filtrants : Le travail des terminologues permet de savoir quel terme est caractéristique de telle et telle catégorie, au sein d'un corpus donné. Il n'est pas garanti que la représentativité des classes de ces termes soit transposable, lorsque l'on remplace le corpus conçu avec l'expert par le corpus du domaine. C'est pourquoi l'expert doit donner son avis sur l'appartenance ou non d'un terme au domaine.
- Spécifier les liens sémantiques de l'ontologie et l'enrichir : l'expert est celui qui est dépositaire de la connaissance technique, sans lui nous ne pourrions pas créer d'ontologie acceptable par la communauté technique.

La sélectivité

Plus tôt, nous avons présenté une hypothèse de travail sur la sélectivité des termes filtrants (voir chapitre 4.2). Selon nous, cette sélectivité est spécifique au niveau de l'arbores-

cence sur lequel se déroule le test filtrant. Elle n'est ni influencée par les tests filtrants précédents, ni par les tests filtrants suivants. À chaque fois que le document avance dans l'arborescence, il change d'ensemble (Racine > ballon > nacelle pointée > commandes contrôle), c'est ce nouvel ensemble qui doit être pris en compte pour la création de filtres destinés aux niveaux suivants. Cette méthodologie nous semble juste, cependant nous n'avons pas pu l'évaluer. Ne disposant pas de corpus déjà annoté, nous n'avons pas pu comparer les performances de classement de cette méthode et de la méthode utilisée auparavant. Un corpus d'une centaine de documents annotés aurait été souhaitable. Les critères de performance relatifs à l'évaluation d'un classement sont le rappel et la précision (voir figure 1.1).

Bilan du travail d'automatisation

Lors de mon stage au CNES nous avons essayé d'automatiser de nombreuses étapes du travail du terminologue.

- Confronté à une extraction terminologique donnant une liste de termes importante, nous avons dû rajouter des informations sur les termes extraits pour nous permettre de réduire cette liste. Les critères utilisés ont été la *fréquence du terme* dans le corpus et *la longueur*, en nombre de mots, de ce dernier (voir chapitre 4.1.5). Ce processus de rajout d'information a été réalisé automatiquement. Il a permis, dans le cas le plus remarquable, de passer d'une liste de 120 000 termes à une liste de 500 termes.
- Les études contrastives ont été partiellement automatisables. La représentativité des termes extraits peut être calculée pour les sous-domaines. Nous avons réalisé un script qui calcule un indice de représentativité : selon le nombre de fois qu'un terme est présent dans un sous-domaine par rapport au nombre de fois où il est présent dans le corpus (voir chapitre 4.2.2.3). La sélection d'une liste de termes selon leurs indices de représentativité pourrait être mis en œuvre automatiquement.
- Concernant les études contrastives des termes du corpus par rapport aux autres domaines, nous avons automatisé la recherche de fréquence et l'extraction du contexte (nom du chapitre) pour le TTVS.
- Pour permettre une visualisation plus aisée, nous avons tenté de réaliser un regroupement des termes selon un réseau de cooccurrences.
- ...

Les étapes automatisables sont pour la plupart en rapport avec la manipulation du corpus. Celles sur lesquelles nous avons travaillé lors de ce stage correspondent principalement à la manipulation des listes de termes extraits.

Réflexions et Perspectives

La question de la variation

"Pour énoncer sa culture, l'informateur utilise le jeu linguistique dont il dispose, et fixe les règles de ce jeu en fonction du contexte qu'il choisit" [Vogel, 1988]

Deux experts du même domaine peuvent avoir deux points de vue différents sur le découpage taxinomique du domaine. De même, si on ré-interroge un expert, son découpage

taxinomique peut être différent. N'étant pas qualifié pour se prononcer sur les connaissances de l'expert, le terminologue se doit d'être le plus objectif possible lors des interviews. À partir d'un corpus récupéré auprès de l'expert, le terminologue peut proposer un découpage taxinomique différent (voir chapitre 2.3). Ce découpage sera comparé avec le découpage de l'expert lors d'une interview.

Les différences de point de vue, entre deux experts ou avec un expert interrogé deux fois, sont dues à des expériences différentes, à une maturité des réflexions ou à un contexte différent.

"*la langue ne préexiste pas à la parole : elle est apprise en son sein, et la compétence des sujets évolue au cours de leurs pratiques effectives*" [Rastier, 2005]

Cette citation sous-entend que chaque individu, ayant évolué dans des milieux différents, possède un jeu linguistique unique et évolutif.

On peut revenir à l'hypothèse Sapir-Whorf précédemment citée (voir chapitre 2.2.1) : "*Human beings [...] are very much at the mercy of the particular language which has become the medium of expression for their society*" (les être humains sont à la merci du langage devenu le modèle d'expression de leur société) qui rajoute une dimension culturelle aux différences de points de vue sur le découpage taxinomique de leurs domaines. (différence de notation, d'unité de mesure, ...)

On comprend que différents individus ayant une culture propre et un moyen de l'exprimer basé sur leurs expériences, aient potentiellement différents points de vue sur un même sujet.

Par contre, si on se base sur leurs connaissances et que l'on demande des textes représentant leur domaine, les documents sélectionnés par différents expert ne varieront pas énormément. Les conditions d'énonciation ou de cognition n'impliquent pas un changement des connaissances de l'expert.

De même, même si l'on remarque des disparités au niveau du découpage taxinomique, lors des interviews pour l'enrichissement de la base de connaissance, les relations entre termes, étant basées sur les connaissances de l'expert, n'auront que peu de variation.

Vers une ontologie dynamique ?

Les ontologies peuvent évoluer, car les concepts qui y sont formalisés ne sont pas figés. En effet, les domaines des spécialités représentées par les ontologies évoluent. De nouvelles techniques se créent, de nouveaux outils sont utilisés ... Le besoin de modification d'une ontologie peut aussi venir des utilisateurs, car leurs besoins et les applications de l'ontologie changent avec le temps. Il peut aussi venir d'autres points de vue sur le domaine conceptualisé. Ces changements peuvent être basiques (ajouts, effacements ...) ou complexes (déplacement, fusion ...). [Stovanovic, 2004] définit diverses notions nécessaires pour l'évolution des ontologies, leur consistance, et propose une taxinomie de leurs changements.

Aux problèmes posés par la variation, lors de la constitution du corpus et de l'ontologie, et par l'évolution des concepts, se pose la question de l'évaluation des ontologies. Pour évaluer la qualité d'une construction d'une ontologie ou ses évolutions il faudrait établir des

critères sur le vocabulaire, la qualité des relations, l'organisation. À ce jour, de nombreuses techniques d'évaluation ont été mises au point, mais il n'y a pas de consensus sur une méthode particulière. Les ontologies étant composées en vue d'une application, les moyens adéquats pour les évaluer peuvent varier. En techniques d'évaluations nous pouvons citer :

- La comparaison d'une ontologie avec d'autres ontologies. [Maedche et Staab, 2002]
- Des mesures quantitatives sur la qualité de la structure [Ning et Shihan, 2006] (densité des relations, peuplement, équilibre des hiérarchies, proximité des concepts ...)
- Des mesures quantitatives de correspondance avec un corpus du domaine. [Brewster *et al.*, 2004]
- L'évaluation de l'application prévue pour l'ontologie. [Maynard *et al.*, 2006]
- L'évaluation par un expert. [Hernandez, 2005]
- L'évaluation par le comportement des utilisateurs. [Kalfoglou et Hu, 2006]

Pour répondre à ces problèmes, des projets tels que DYNAMO [11] (DYNAMic Ontology for information retrieval) utilisent des systèmes multi-agents pour maintenir (ou faire évoluer) une ontologie. Les systèmes multi-agents (SMA) [Ferber, 1999] sont composés d'entités actives et autonomes ayant un objectif individuel et interagissant dans un environnement commun. DYNAMO utilise deux modules, un pour la construction et la maintenance d'ontologie, l'autre pour l'annotation, l'indexation et la recherche d'information. DYNAMO est utilisé conjointement avec un utilisateur (terminologue) pour la constitution et l'évolution d'ontologies. De nombreux travaux sont en cours de réalisation sur ce projet. Le plus récent, réalisé par Zied Sellami lors de sa thèse, [Sellami, 2012], à l'université Paul Sabatier (Toulouse), cherche à gérer dynamiquement les ontologies à l'aide d'un système multi-agents adaptatif (AMAS), voir [Georgé *et al.*, 2011]. Les AMAS sont des SMA dont les agents sont dotés de capacités d'auto-organisation et de coopération entre eux, les rendant capables de s'adapter aux changements de leur environnement. Les AMAS reposent également sur le principe d'émergence et utilisent un Atelier de DEveloppement de Logiciels à Fonctionnalité Émergente (ADELFE) [12].

Conclusion

Mon stage au CNES portait sur le développement d'une base de connaissance à partir de textes. Lors de ce stage nous avons travaillé sur deux domaines : les Nacelles Pointées (dont les extractions par Talismane posaient problème) et les Radars (domaine qui venait d'être entamé).

Par ces deux domaines j'ai pu suivre l'enrichissement de la base de connaissance pratiquement de A à Z. Sur le domaine Radar j'ai pu participer aux interviews de composition du corpus, utiliser Alceste pour découper le corpus et participer aux interviews sur la taxinomie du domaine. Sur le domaine Nacelle Pointée j'ai pu proposer des solutions de réduction des listes de termes et travailler jusqu'à la première interview pour le contrôle de leurs qualités avec l'expert.

J'ai eu à travailler sur la création de corpus, leurs nettoyages pré-traitement, l'utilisation

11. http://www.irit.fr/dynamo/
12. http://www.irit.fr/ADELFE/

de l'outil Alsceste, Antconc et Lexico3. J'ai été amené à proposer des solutions pour réduire le nombre d'entrées d'une sortie mal formée de Talismane. J'ai pu proposer une méthode de filtrage différente. J'ai également été amené à travailler sur les méthodes de présentation des listes aux experts.

Je n'ai malheureusement pas eu l'occasion, lors de mon stage, d'implanter des triplets dans la base de connaissance.

Je pense que ce stage m'a été bénéfique. J'ai pu découvrir ce qu'est la gestion de connaissance et approfondir mes connaissances sur les ontologies. J'ai apprécié les ontologies et j'aimerais pouvoir travailler à nouveau dessus par la suite.

Bibliographie

Anne ABEILLÉ, Lionel CLÉMENT et François TOUSSENEL : Building a treebank for french. *Dans Treebanks*, pages 165–187. Springer, 2003.

Ramadan ALFARED, Denis BÉCHET, Alexander DIKOVSKY *et al.* : Cdg lab : a toolbox for dependency grammars and dependency treebanks development. *Dans DEPLING 2011 (International Conference on Dependency Linguistics)*, 2011.

Nathalie AUSSENAC-GILLES, Jean CHARLET, Chantal REYNAUD *et al.* : Les enjeux de l'ingénierie des connaissances. *Information, Interaction, Intelligence-Le point sur le i (3)*, 2012.

Nathalie AUSSENAC-GILLES et Anne CONDAMINES : Variations syntaxiques et contextuelles dans la mise au point de patrons de relations conceptuelles. *Filtrage sémantique dans les textes. Approches symboliques*, 4:109–149, 2009.

Jean-Paul BENZÉCRI : coll.(1973), l'analyse des données. *Tome*, 2:l, 1973.

Didier BOURIGAULT, Cécile FABRE, Cécile FRÉROT, Marie-Paule JACQUES, Sylwia OZDOWSKA *et al.* : Syntex, analyseur syntaxique de corpus. *Dans Actes des 12èmes journées sur le Traitement Automatique des Langues Naturelles*, 2005.

Didier BOURIGAULT et Monique SLODZIAN : Pour une terminologie textuelle. *Terminologies nouvelles*, 19(1999):29–32, 1999.

Christopher BREWSTER, Harith ALANI, Srinandan DASMAHAPATRA et Yorick WILKS : Data driven ontology evaluation. 2004.

Anne CONDAMINES : Sémantique et corpus spécialisés : Consitution de bases de connaissances terminologiques, mémoire d'hdr en sciences du langage. *Carnets de Grammaire, ERSS, Toulouse, France*, 2003.

Eugenio COSERIU : *Sistema, norma e" parola"*. Paideia, 1969.

Jacques FERBER : *Multi-agent systems : an introduction to distributed artificial intelligence*, volume 1. Addison-Wesley Reading, 1999.

Jean-Pierre GEORGÉ, Marie-Pierre GLEIZES et Valérie CAMPS : Cooperation. *Dans* Giovanna DI MARZO SERUGENDO, Marie-Pierre GLEIZES et Anthony KARAGEORGOS,

éditeurs : *Self-organising Software*, Natural Computing Series, pages 193–226. Springer, http ://www.springerlink.com, 2011. URL http://www.springerlink.com/content/r444353816j80851/fulltext.pdf.

Thomas Robert GRUBER *et al.* : A translation approach to portable ontology specifications. *Knowledge acquisition*, 5(2):199–220, 1993.

Benoît HABERT, Adeline NAZARENKO et André SALEM : *Les linguistiques de corpus*, volume 1. Armand Colin, 1997.

Nathalie HERNANDEZ : *Ontologies de domaine pour la modélisation du contexte en recherche d'information*. Thèse de doctorat, Université Paul Sabatier-Toulouse III, 2005.

Harry HOIJER : The sapir-whorf hypothesis. *Language in culture*, pages 92–105, 1954.

Yannis KALFOGLOU et Bo HU : Issues with evaluating and using publicly available ontologies. 2006.

Maurice George KENDALL : and stuart, alan 1967. *The advanced theory of statistics*, 2:278–345, 1967.

Eric KOW, Yannick PARMENTIER, Claire GARDENT *et al.* : Semtag, the loria toolbox for tag-based parsing and generation. *Dans TAG+ 8 (The Eighth International Workshop on Tree Adjoining Grammar and Related Formalisms)*, 2006.

Sandra KÜBLER, Ryan MCDONALD et Joakim NIVRE : Dependency parsing. *Synthesis Lectures on Human Language Technologies*, 1(1):1–127, 2009.

Stéphane LAPALUT : Text clustering to support knowledge acquisition from documents. 1995.

Danielle LEEMAN : Le «sens» et l'«information» chez harris. *Linx. Revue des linguistes de l'université Paris X Nanterre*, (8):209–220, 1996.

Claude LÉVI-STRAUSS : La pensée sauvage. 1962. *Paris, Plon*, 1962.

Alexander MAEDCHE et Steffen STAAB : Measuring similarity between ontologies. *Dans Knowledge engineering and knowledge management : Ontologies and the semantic web*, pages 251–263. Springer, 2002.

Véronique MALAISÉ, Pierre ZWEIGENBAUM et Bruno BACHIMONT : Detecting semantic relations between terms in definitions. *Dans COLING*, pages 55–62, 2004.

André MARTINET : Eléments de linguistique générale. 1960.

Diana MAYNARD, Wim PETERS et Yaoyong LI : Metrics for evaluation of ontology-based information extraction. *Dans International World Wide Web Conference*, 2006.

Ryan MCDONALD : *Discriminative learning and spanning tree algorithms for dependency parsing*. Thèse de doctorat, University of Pennsylvania, 2006.

Alexis NASR : Analyse syntaxique probabiliste pour grammaires de dépendances extraites automatiquement. *Habilitation à diriger des recherches, Université Paris*, 7:57, 2004.

Huang NING et Diao SHIHAN : Structure-based ontology evaluation. *Dans e-Business Engineering, 2006. ICEBE'06. IEEE International Conference on*, pages 132–137. IEEE, 2006.

Joakim NIVRE : Algorithms for deterministic incremental dependency parsing. *Computational Linguistics*, 34(4):513–553, 2008.

Natalya Fridman NOY et Deborah Louise MCGUINNESS : Développement d'une ontologie 101 : Guide pour la création de votre première ontologie. *Université de Stanford, Stanford, Traduit de l'anglais par Anila Angjeli. ht tp ://www. bnf. fr/pages/infopro/normes/pdf/noDevOnto. pdf*, 2000.

François RASTIER : Enjeux épistémologiques de la linguistique de corpus. *G. Williams (éd.), La*, 2005.

Josette REBEYROLLE et Ludovic TANGUY : Repérage automatique de structures linguistiques en corpus : le cas des énoncés définitoires. *Cahiers de grammaire*, 25:153–174, 2000.

Max REINERT : Présentation du logiciel alceste à l'aide d'un exemple. *Psychologie et Éducation*, 10:2, 1986.

Edward SAPIR : Language : An introduction to the study of speech, 1921.

Zied SELLAMI : *Gestion dynamique d'ontologies à partir de textes par systèmes multi-agents adaptatifs*. Thèse de doctorat, Université de Toulouse, Toulouse, France, juillet 2012.

John SINCLAIR : Preliminary recommendations on corpus typology. *EAGLES Document TCWG-CTYP/P (available from http ://www. ilc. pi. cnr. it/EAGLES/corpustyp/corpustyp. html)*, 1996.

Ljiljana STOVANOVIC : *Methods and tools for ontology evolution*. Thèse de doctorat, Karlsruhe, Univ., Diss., 2004, 2004.

Alan STUART et Keith ORD : *Kendall's Advanced Theory of Statistics : Volume 1 : Distribution Theory*. Numéro vol. 1 ;vol. 1994 *dans* Kendall's advanced theory of statistics. Wiley, 2009. ISBN 9780340614303. URL `http://books.google.fr/books?id=tW18thQWJQIC`.

Ludovic TANGUY : *Complexification des données et des techniques en linguistique : contributions du TAL aux solutions et aux problèmes*. Thèse de doctorat, Université Toulouse le Mirail-Toulouse II, 2012.

Assaf URIELI et Ludovic TANGUY : L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions : études de cas avec l'analyseur talismane. *Actes de TALN*, 2013.

Tristan VANRULLEN, Philippe BLACHE, Jean-Marie BALFOURIER *et al.* : Constraint-based parsing as an efficient solution : Results from the parsing evaluation campaign easy. *Proceedings of LREC 2006 (Language Resources and Evaluation)*, pages 165–170, 2006.

Claude VOGEL : *Génie cognitif*. Masson, 1988.

Benjamin Lee WHORF, John Bissel CARROLL et Stuart CHASE : *Language, Thought and Reality, Selected Writings of Benjamin Lee Whorf. Edited... by John B. Carroll. Foreword by Stuart Chase*. Mass., 1956.