Enrichissement d'un réseau sémantique multilingue

Julie Jouanneaux

Enrichissement d'un réseau sémantique multilingue

JOUANNEAUX Julie

Sous la direction de Pr Thomas LEBARBÉ et Dr Dominique NOËL

LLASIC (UFR Langage, Lettres, Arts du Spectacle, Information et Communication)

Département d'Informatique pour les Lettres, Langues et Langage

---

Mémoire de Master 2 - 20 crédits

Parcours : Industrie De la Langue

Année universitaire 2015-2016

Enrichissement d'un réseau sémantique multilingue

JOUANNEAUX
Julie

Sous la direction de Pr Thomas LEBARBÉ et Dr Dominique NOËL

LLASIC (UFR Langage, Lettres, Arts du Spectacle, Information et Communication)

Département d'Informatique pour les Lettres, Langues et Langage

---

Mémoire de Master 2 - 20 crédits

Parcours : Industrie De la Langue

Année universitaire 2015-2016

Remerciements

Mes remerciements les plus sincères à Dominique Noël pour m'avoir accompagnée lors de ce stage. Je remercie également mon camarade de travail Paul Labruyère pour avoir partagé avec moi ce stage avec humour. Je n'oublie pas tous mes collègues Expert System en particulier l'équipe de Recherche et Développement du bureau de Paris pour m'avoir accueillie, intégrée, écoutée, soutenue et encouragée comme il se doit ! Un grand merci à Thomas Lebarbé pour m'avoir suivie et conseillée lors de ces 6 mois.

Un remerciement un peu spécial à John Williams, Michael Giacchino, Tyler Bates, Fakear et Chopin pour ces heures de musique qui m'ont permis de me concentrer pendant toutes ces heures de travail parfois acharnées.

UNIVERSITÉ Grenoble Alpes

Déclaration anti-plagiat
Document **à scanner** après signature
Et **à intégrer** au mémoire électronique

DÉCLARATION

1. Ce travail est le fruit d'un travail personnel et constitue un document original.
2. Je sais que prétendre être l'auteur d'un travail écrit par une autre personne est une pratique sévèrement sanctionnée par la loi.
3. Personne d'autre que moi n'a le droit de faire valoir ce travail, en totalité ou en partie, comme le sien.
4. Les propos repris mot à mot à d'autres auteurs figurent entre guillemets (citations).
5. Les écrits sur lesquels je m'appuie dans ce mémoire sont systématiquement référencés selon un système de renvoi bibliographique clair et précis.

NOM : JOUANNEAUX

PRENOM : Julie

DATE : 07/09/2016 SIGNATURE : [redacted]

Sommaire

Introduction

Ce document est le résultat d'un travail de 6 mois au sein d'une entreprise de développement de solutions sémantiques : Expert System.

Je fais partie de la première équipe française à travailler sur le réseau sémantique du français de l'entreprise. L'objectif était de faire avancer le développement de cette ressource qui sert de base aux différents outils développés par l'entreprise.

Ma mission a été également d'effectuer une analyse de la qualité des désambiguïsations fournies par un moteur d'analyse sémantique de l'entreprise. Cette phase de test a pour but de définir les axes prioritaires de développement. Ces axes de développement permettront par la suite d'appliquer un plan d'action pour optimiser le développement de cette ressource attendue sur différents projets autour du français.

Ce document décrit, dans la Partie 1, les structures et les étapes de développement du réseau multilingue et des réseaux monolingues. La Partie 2 présente les tâches effectuées dans le cadre de ce stage. Enfin, les Partie 3 et 4 comportent la description d'une proposition de test et les conclusions des analyses effectuées.

Partie 1

-

Cadre de travail

Chapitre 1 - Présentation de l'entreprise

1. Généralités

Expert System est une entreprise éditeur de logiciels, basée en Italie (Modène) qui compte des antennes dans différents pays d'Europe (Espagne, Italie, France, Angleterre et Allemagne) et d'Amérique du Nord (États-Unis et Canada). Cette entreprise développe et commercialise des logiciels d'analyse sémantique. Elle propose des solutions sémantiques qui analysent le sens, la nature et la pertinence des contenus ciblés. En 2015, l'entreprise fusionne avec l'entreprise française Temis, éditeur de solutions en Text Mining[1], et de ce fait élargit son champ d'actions.

Fondée en 1989 à Modène en Italie, l'entreprise est listée à la Bourse italienne, AIM Italia, à partir de février 2014 et au mois de juillet de la même année elle apparaît dans le Magic Quadrant des entreprises de recherches de Gartner. Plus de 250 personnes y travaillent à ce jour.

Aujourd'hui, Marco Varone est le Président Directeur Général, Gilles Pouzenc est Directeur de la Technologie et Claudio Palmolungo est Vice-Président Exécutif d'Expert System.

Figure 1 : Logo de l'entreprise où le stage s'est déroulé.

La maison-mère d'Expert System France est basée à Paris. Ce bureau regroupe différentes branches de l'entreprise : la direction, le département commercial ainsi que les départements des Services, de l'Avant-Vente et de la Recherche & Développement. Un

[1] Text Mining : Le Text Mining est un domaine dans le traitement automatique des langues également appelé « Fouille de textes » qui rassemble des techniques de linguistique, du langage, de la sémantique, des statistiques ainsi que l'informatique.

bureau antenne se trouve à Grenoble. Il regroupe principalement les départements d'assurance qualité et du support. Le stage sur le projet « Sensigrafo FR » s'est déroulé au bureau de Paris dans le département de Recherche et Développement où se trouve la responsable du projet « Sensigrafo FR » en France.

2. Présentation de l'équipe de travail

Le projet Sensigrafo FR a débuté dans l'équipe italienne d'Alberto Bonnazi. Celui-ci est responsable du projet et expert sur le développement et la structure de ce type de réseau. C'est par conséquent auprès de lui que nous nous référons pour toutes questions. De plus, il nous fournit les instructions et le matériel de travail.

Dominique Noël est référente du projet Sensigrafo FR en France. Les éventuels échanges avec le responsable du projet sont effectués par elle. Une de ses missions est d'être l'intermédiaire entre le responsable du projet et les stagiaires. Elle se charge du bon déroulement du stage que ce soit au sein de l'entreprise ou du côté universitaire pour tout ce qui concerne les mémoires et les soutenances. Par ailleurs, d'autres missions d'entreprise lui sont affectées.

Paul Labruyère partage le travail des tâches présentées dans la Partie 2 de ce document dans le cadre de son stage de fin d'études.

- Alberto BONAZZI – Linguiste informaticien et responsable du projet « Sensigrafo étendu » en poste au bureau de Modène.
- Dr Dominique NOËL - Linguiste informaticienne et tutrice des stagiaires sur le projet « Sensigrafo FR » en poste au bureau de Paris.
- Paul LABRUYÈRE - Stagiaire, étudiant en Master 2 Lexicographie Terminographie et Traitement Automatique de Corpus (LTTAC) Lille 3 et stagiaire au bureau de Paris.

3. Contact

www.expertsystem.com – info@expertsystem.com

Chapitre 2 - Présentation du projet Sensigrafo

1. Généralités

Le stage se concentrait sur l'amélioration d'une ressource : le Sensigrafo FR. Plusieurs outils et interfaces ont été utilisés dans ce but. Nous allons premièrement présenter le contexte général.

Le « Sensigrafo » est le nom donné à un vaste projet de réseaux sémantiques. Il comprend différents niveaux, le « Sensigrafo étendu » et les « Sensigrafos monolingues ». Le « Sensigrafo étendu » est un réseau fondé sur une structure parallèle à la structure à nœuds de Wikipédia. C'est aussi un réseau sémantique multilingue composé de plusieurs réseaux sémantiques monolingues.

Ces réseaux monolingues ont pour appellation « Sensigrafo » suivi de la langue dont il est question. Par exemple pour le français l'appellation utilisée est « Sensigrafo français » soit « Sensigrafo FR » à l'écrit. Pour l'anglais l'appellation utilisée est « Sensigrafo anglais » écrit « Sensigrafo EN ». Nous utiliserons le terme « Sensigrafo monolingue » lorsque nous ne désignerons pas le réseau d'une langue en particulier. Lors du développement d'un Sensigrafo monolingue, ce dernier est lié, à certains niveaux, à un Sensigrafo monolingue pour lequel le développement est achevé. Cette mise en relation permet d'acquérir automatiquement certaines informations. C'est à ce titre que le Sensigrafo étendu est multilingue.

Il faut savoir, avant tout, qu'un Sensigrafo monolingue fait partie d'un moteur d'analyse sémantique. Ce moteur d'analyse sémantique peut être schématisé en deux branches : d'un côté se trouvent les règles et de l'autre les ressources.

Le côté des « règles » est composé d'un analyseur codé en C++ qui effectue des analyses grammaticales, sémantiques et syntaxique…

Les analyses grammaticales correspondent aux désambiguïsations des catégories grammaticales et des groupes grammaticaux. Les analyses sémantiques correspondent à la désambiguïsation sémantique des termes présents ainsi qu'à la reconnaissance des entités nommées. L'analyse syntaxique correspond à la reconnaissance des relations syntaxiques du type reconnaissance du sujet, de l'objet ou des compléments d'un verbe.

Le côté des « ressources » comprend un dictionnaire de flexions qui correspond au lexique et le Sensigrafo monolingue qui correspond à la sémantique du moteur.

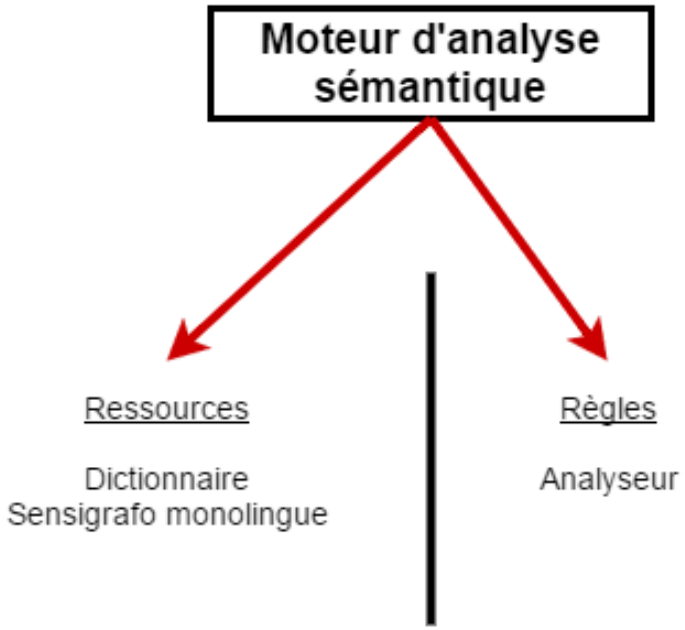


Figure 2 : Représentation schématique de la composition du moteur d'analyse sémantique d'Expert System

2. Sensigrafo monolingue

Les Sensigrafo « monolingues » sont des ontologies développées par Expert System depuis une quinzaine d'années. Une ontologie est un ensemble structuré des termes, concepts et liens existant, entre les informations, d'un domaine de connaissance donné. Une ontologie peut décrire un domaine très large comme une langue, ce qui est le cas ici, ou bien un domaine plus restreint comme les termes et concepts de la mécanique automobile. Deux Sensigrafos monolingues sont achevés, l'un pour l'italien (IT) et l'autre pour l'anglais (EN). Ces deux Sensigrafos monolingues ont été développés essentiellement manuellement. D'autres sont en cours de développement : le français (FR), l'allemand (DE), l'espagnol (ES), le chinois (ZH), le japonais (JA), le coréen (KO), le portugais (PT) et le russe (RU). Le développement d'un Sensigrafo monolingue se déroule en 4 grandes étapes : La constitution de la base dictionnairique, la mise en correspondance, la dévirtualisation et la compilation.

Les Sensigrafos monolingues sont composés d'éléments appelés « syncons ». C'est un terme utilisé dans les réseaux sémantiques d'Expert System pour désigner des entités conceptuelles regroupant toutes les informations sur un concept dans une langue. Dans la langue, une forme lexicale a une ou plusieurs acceptions c'est-à-dire des définitions, des utilisations. Une acception est le sens dans lequel un mot est utilisé. Pour désigner une forme

lexicale nous utilisons dans ce document le terme « lemme Sensi[2] » pour éviter les confusions avec le concept de « lemme[3] » en linguistique. La majorité des lemmes Sensi sont des mots seuls. Cependant, il est possible que certains lemmes Sensi soient formés de plusieurs mots comme des locutions tels que la locution adverbiale « à la fois ». D'autres lemmes Sensi commencent par la forme « cat. ». Ce ne sont pas des formes lexicales utilisées dans la langue pour faire appel à un concept. Ces lemmes Sensi commençant par « cat. » sont communs à tous les Sensigrafo monolingues. Ils sont utilisés pour catégoriser les syncons et ont un rôle important dans le développement des Sensigrafo monolingues. Il existe 11 syncons de catégories sémantiques de base, c'est-à-dire qu'ils n'ont pas de « père » sémantique. Les 11 catégories sont les catégories « État », « Objet », « Être humain », « Plante », « Animal », « Concept », « Groupe, ensemble », « Quantité », « Phénomène naturel », « Lieux » et « Temps ». Il existe plus de 200 catégories en plus des 11 catégories de base.

Un « père » sémantique est un concept plus général qu'un autre concept par rapport à une relation sémantique donnée. Un « fils » sémantique est un concept moins général qu'un concept par rapport à une relation donnée. Par exemple, prenons le concept de « femme » en tant qu'être humain de sexe féminin, « femme » est un type d'être humain. La relation « est un type de » s'appelle hyperonymie. Le concept de femme, présent ici, est donc un fils du concept « être humain » selon la relation sémantique d'hyperonymie. À l'inverse, « personne » est un père « hyperonyme » de femme. La relation présente dans cet exemple est une relation hiérarchique conceptuelle. Une relation hiérarchique conceptuelle est une relation entre deux concepts dont l'un est plus générique que l'autre par rapport à une relation sémantique. Pour désigner un ensemble d'éléments liés par des relations sémantiques nous parlerons de chaîne hiérarchique conceptuelle. Cette relation de hiérarchie peut s'appliquer à partir de plusieurs relations sémantiques. La liste complète des relations sémantiques utilisées dans le projet Sensigrafo se trouve à l'annexe 1.

Des informations sur la nature, les caractéristiques, l'utilisation ou encore les relations sémantiques de ces syncons peuvent être renseignées. Ces informations sont utilisées lors des différentes désambiguïsations effectuées dans les logiciels où le moteur d'analyse sémantique est implanté.

---

[2] Lemme Sensi : le lemme Sensi est le terme utilisé dans le projet Sensigrafo FR pour désigner la forme lexicale de base permettant d'exprimer un concept.

[3] Lemme : En linguistique un lemme est la forme de base d'un mot. Par exemple, pour les verbes les infinitifs sont les lemmes.

Nous ne connaissons pas le nombre exact d'informations qu'il est possible de renseigner car toutes n'ont pas été abordées lors de ce stage. C'est pour cette raison que dans ce document nous ne présenterons que les informations avec lesquelles nous avons travaillé.

A. Le Sensigrafo FR

Le Sensigrafo FR ne définit que les mots dit « pleins » c'est-à-dire des noms, verbes, adverbes et adjectifs ainsi que des noms propres (prénoms, personnalités, organisations, entreprises, villes, pays, fleuves, massifs de montagne, évènements, etc…). Les mots grammaticaux[4] ne sont pas définis dans ce réseau.

Le réseau est constitué d'une base dictionnairique. Des définitions ont été extraites pour former les glossas qui forment la base des syncons. Les glossas sont les formulations descriptives des concepts.

Précédemment nous parlions du concept de lemmes Sensi dans un Sensigrafo monolingue. Nous utilisons le terme « lemme Sensi » mais certains sont en réalité des collocations ou des locutions comme « alors que » ou « d'abord ». Les noms propres sont souvent composés comme par exemple : la « Ligue des Champions ».

La figure 3 est une représentation schématique du Sensigrafo FR fondée sur des données présentes dans le Sensigrafo FR. Quatre syncons y sont représentés. Trois d'entre eux sont associés au lemme Sensi « hérisson » et le dernier syncon est associé au lemme Sensi « mammifère terrestre » ainsi qu'à la catégorie « cat. mammiferi terrestri ».

B. Relations Lemme/Syncon

Les syncons sont associés systématiquement et au minimum avec un lemme Sensi. Les syncons associés à un même lemme Sensi sont les acceptions possibles de ce lemme Sensi. La base du syncon est la formulation du concept qu'il porte nommé glossa. Le terme « café » peut désigner une boisson ou une couleur. Le concept de la boisson et le concept de la couleur, tous deux associés au lemme Sensi « café », sont deux concepts distincts et uniques donc ils formeront 2 syncons. Un syncon peut avoir plusieurs lemmes Sensi associés (figure 4) comme plusieurs syncons peuvent avoir le même lemme Sensi associé (figure 5).

[4] Mots grammaticaux : Les mots grammaticaux sont les mots des catégories grammaticales des articles et déterminants, pronoms, prépositions, conjonctions (de coordination et de subordination) et des interjections.

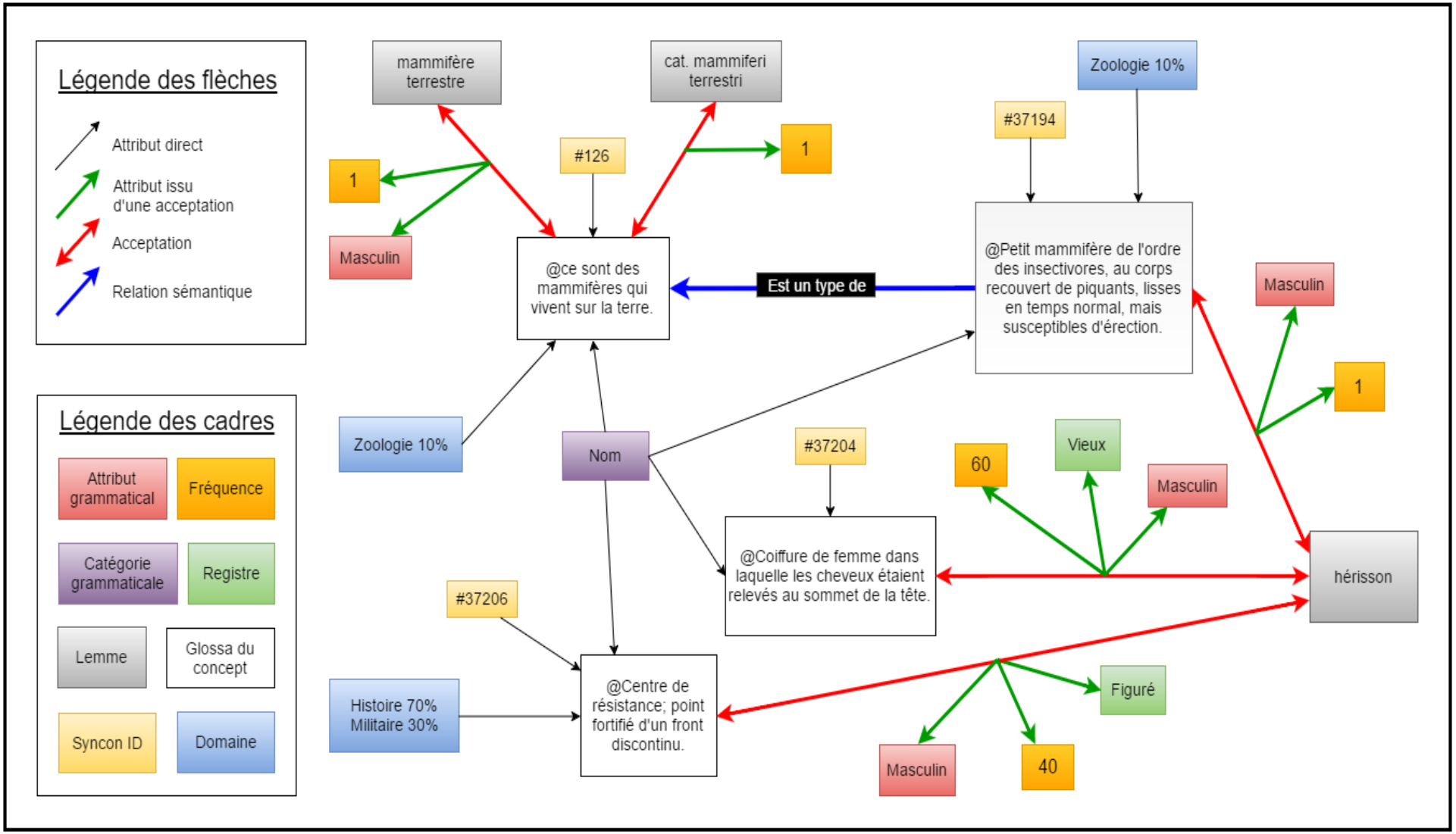


Figure 3 : Représentation schématique du Sensigrafo FR fondé sur des données extraites du Sensigrafo FR.

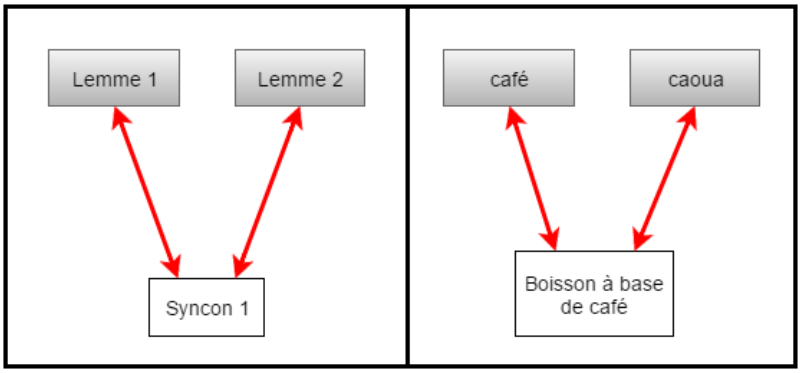


Figure 4 : À gauche se trouve une représentation schématique de l'association lemme Sensi/syncon où deux lemmes Sensi sont associés à un syncon. À droite se trouve le même schéma avec un exemple.

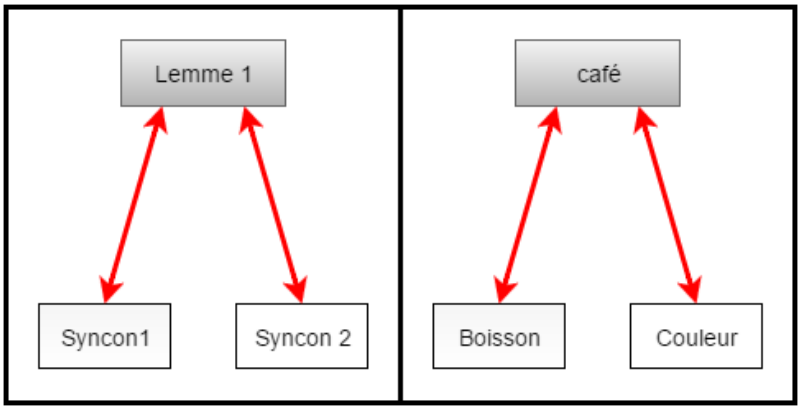


Figure 5 : À gauche se trouve une représentation schématique de l'association lemme Sensi/syncon où un lemme est associé à plusieurs syncons. À droite se trouve le même schéma avec l'exemple du terme « café » qui peut désigner le concept de la boisson caféinée ou bien le concept de la couleur de la teinte marron semblable à la couleur de la boisson caféinée à base de café.

Les différentes structures présentées dans les figures 4 et 5 peuvent se combiner. Ainsi, régulièrement, des structures analogues à la figure 6 seront rencontrées. Ce sont des structures où plusieurs syncons sont associés à un même lemme Sensi et l'un de ces syncons à un second lemme Sensi associé.

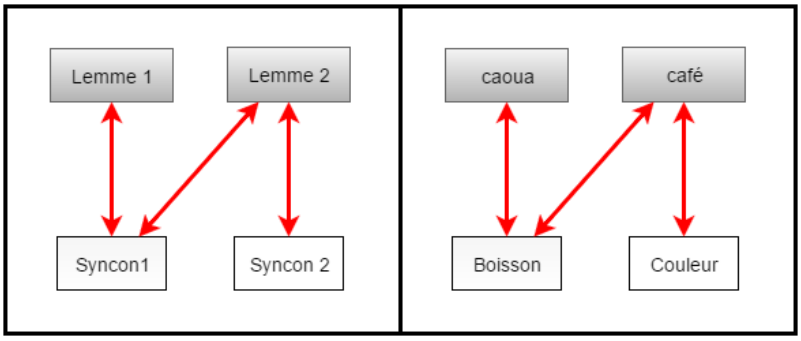


Figure 6 : À gauche se trouve une représentation schématique de l'association lemme Sensi/syncon où se fusionnent les précédentes structures. À droite se trouve le même schéma avec un exemple.

C. Les attributs de relations Lemme Sensi /Syncon

Nous avons donc les lemmes Sensi et les syncons. L'association de ces deux éléments produit de nouvelles informations (voir figure 7).

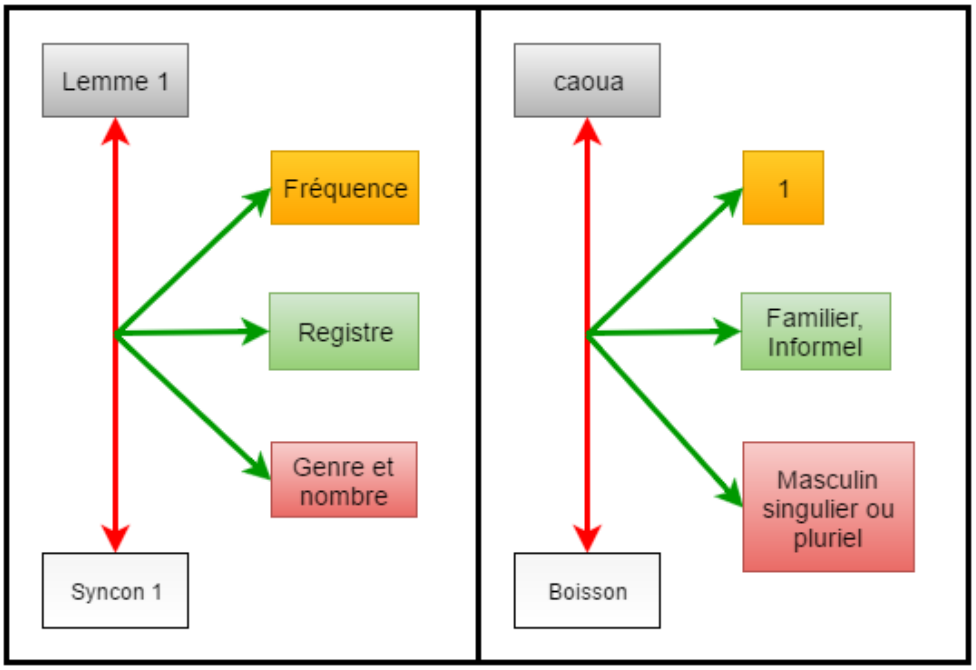


Figure 7 : À gauche se trouve une représentation schématique d'attributs possibles pour une association lemme Sensi/syncon. À droite se trouve le même schéma avec un exemple.

1. Le registre

Les registres Sensi[5] sont des registres[6] qui permettent d'avoir une information complémentaire. La liste complète des registres Sensi disponibles dans le Sensigrafo FR se trouve à l'annexe 2. Par exemple, l'utilisation du mot « caoua » pour désigner le concept de la boisson appartient au registre familier contrairement à « café » qui appartient au registre standard.

[5] Registre Sensi : terme utilisé dans le projet Sensigrafo FR pour désigner un type d'information relative au niveau de langue associé aux syncons.
[6] Registre : terme désignant un mode d'expression relatif à une situation d'expression particulière. Il est possible de choisir certains termes ou structures grammaticales adaptés selon la situation en présence.

2. Le genre et le nombre

Au même niveau se trouve le genre[7]et le nombre[8] s'il est restrictif pour une acception. Prenons l'exemple de la marque de cigarettes « gauloises ». Une « Gauloise » est un terme utilisé pour désigner une cigarette de la marque « Gauloises » fondée en 1910. Ainsi le syncon du concept de la « cigarette » associé au lemme Sensi « gaulois » aura le genre féminin uniquement car le concept de la cigarette de la marque « Gauloise » n'existe pas au masculin. Le nombre n'est pas restrictif dans cet exemple.

3. La fréquence

La fréquence représente la fréquence d'utilisation d'une acception pour un lemme Sensi donné dans la langue courante.

Par exemple, le mot « hérisson » est plus souvent utilisé en français standard pour désigner le petit mammifère que pour désigner l'outil de ramonage.

La fréquence est une échelle de 1 à 100 appliquée aux différents syncons d'un terme. Un lemme Sensi doit obligatoirement avoir une acception ayant une fréquence de 1. Cette fréquence de 1 veut dire que c'est l'acception la plus courante dans la langue française standard. Le reste des acceptions est échelonné sur 5 groupes à partir d'indications flexibles.

- Groupe 1

Le groupe 1 représente la « première acception » qui aura une fréquence de 1 ainsi que les acceptions les plus connues, les acceptions les plus évidentes pour la grande majorité des locuteurs d'une langue.

- Groupe 1 bis

Le groupe 1 bis est composé des acceptions connues par la majorité des gens ainsi que les acceptions qui sont utilisées dans des locutions connues où le contexte n'est pas nécessaire pour la compréhension du sens.

- Groupe 2

Dans le groupe 2 se trouvent les acceptions bien connues mais où le contexte est nécessaire pour la compréhension du sens.

[7] Genre : Le genre informe si le terme et son acception en présence sont au féminin ou au masculin.
[8] Nombre : Le nombre informe si le terme et son acception en présence sont au singulier ou au pluriel.

- Groupe 3 A

Le groupe 3 A regroupe les acceptions dites techniques et spécifiques comme une acception spécifique à un domaine et peu connue du grand public.

- Groupe 3 B

Le groupe 3 B comprend les acceptions archaïques et obsolètes.

Pour illustrer cette notion de fréquence prenons l'exemple du terme « hérisson ». Le Sensigrafo éditable compte 11 syncons associés à ce lemme. Le tableau 1 reprend sur chaque ligne les différents syncons avec leur glossa, fréquence et domaines. La dernière colonne donne le groupe auquel elle a été affectée et pourquoi.

L'écart entre les fréquences attribuées doit être relatif par rapport au groupe dans lequel ils se placent. L'écart doit être d'un minimum de 5 entre les groupes « 1 » et « 1 bis » ainsi que pour les groupes « 1 bis » et « 2 ». Le groupe 3A commence à l'indice 30. Cela dit s'il y a trop de syncons dans les groupes précédents il est possible de repousser cette limite. Il en est de même pour le groupe 3 B. Il débute en théorie à l'indice 60. Si dans un même groupe il y a peu de syncon il est possible de les espacer assez largement comme dans l'exemple présenté dans le tableau 1 où les acceptions du Groupe 1 bis ont un écart de 5. La fréquence s'applique donc en fonction des autres acceptions présentes et il est important de noter que la fréquence s'applique, dans cette logique, aux acceptions de même classe grammaticale.

L'exemple du tableau 1 ne donne que des syncons avec l'attribut « nom ». Logiquement, pour le lemme « observateur » la fréquence sera appliquée sur les acceptions avec la classe grammaticale « nom », puis sur les acceptions avec la classe grammaticale « adjectif » sans prendre en compte les fréquences des acceptions appliquées aux autres classes grammaticales.

D. Relations Syncon/Syncon

1. Les relations sémantiques

Les relations sémantiques sont des relations données entre syncons (voir la figure 8). Ainsi le syncon désignant « la partie du corps humain se trouvant à l'extrémité du membre supérieur » (associé au lemme Sensi « main ») est un méronyme[9](qui peut se

[9] Méronyme : Un méronyme est un terme qu'on peut définir comme étant une « partie de » d'un autre concept. La relation sémantique associée s'appelle la méronymie.

traduire par « est une partie de ») du syncon « membre supérieur du corps humain » (associé au mot « bras »). Ainsi, on peut dire que le concept de main présent ici est un méronyme du concept de corps humain mais il est également possible de dire que le concept de main ici est un fils du concept de corps humain selon la relation de méronymie.

Glossa	Fréquence	Domaine	Groupe
@Petit mammifère de l'ordre des insectivores, au corps recouvert de piquants, lisses en temps normal, mais susceptibles d'érection.	1	Zoologie	Cette acception est le « Premier sens » donc elle fait partie du Groupe 1 et aura la fréquence de 1.
@Tige garnie de lames flexibles en métal, disposées en étoile, et servant à ramoner les cheminées.	5	Pas de domaine	Cette acception est technique mais connue donc elle fait partie du Groupe 1 bis.
@[égouttoir, porte-bouteilles]Tige garnie de chevilles où l'on place les bouteilles à égoutter.	10	Terme technique	Cette acception est technique mais connue donc elle fait partie du Groupe 1 bis.
@Personne d'un caractère, d'un abord difficile.	15	Psychologie	Cette acception est connue mais peu utilisée donc elle fait partie du groupe Groupe 1 bis.
@Roue dentelée. | Grappin à quatre becs. | Assemblage de pointes de fer garnissant le sommet d'un mur, d'une grille, d'une clôture, pour empêcher l'escalade.	30	Marine	Cette acception est technique donc elle fait partie du groupe 3 A .
@Élément mobile d'un réseau barbelé, formé d'un quadrilatère de fils de fer barbelés.	32	Pas de domaine	Cette acception est technique donc elle fait partie du groupe 3 A .
@[herse]Rouleau* garni de pointes pour écraser les mottes de terre.	34	Agriculture	Cette acception est technique donc elle fait partie du groupe 3 A .
@Organe distributeur du semoir d'engrais.	36	Agriculture	Cette acception est technique donc elle fait partie du groupe 3 A .
@Centre de résistance; point fortifié d'un front discontinu.	40	Histoire, Militaire	Cette acception est technique et historique à la fois. Elle fait partie du groupe 3 B.
@Engin formé d'une poutre hérissée de pointes de fer.	50	Pas de domaine	Cette acception est technique et historique à la fois. Elle fait partie du groupe 3 B.
@Coiffure de femme dans laquelle les cheveux étaient relevés au sommet de la tête.	60	Pas de domaine	Cette acception est technique et historique à la fois. Elle fait partie du groupe 3 B.

Tableau 1 : Exemple d'application de fréquences pour les acceptions du nom « hérisson ».

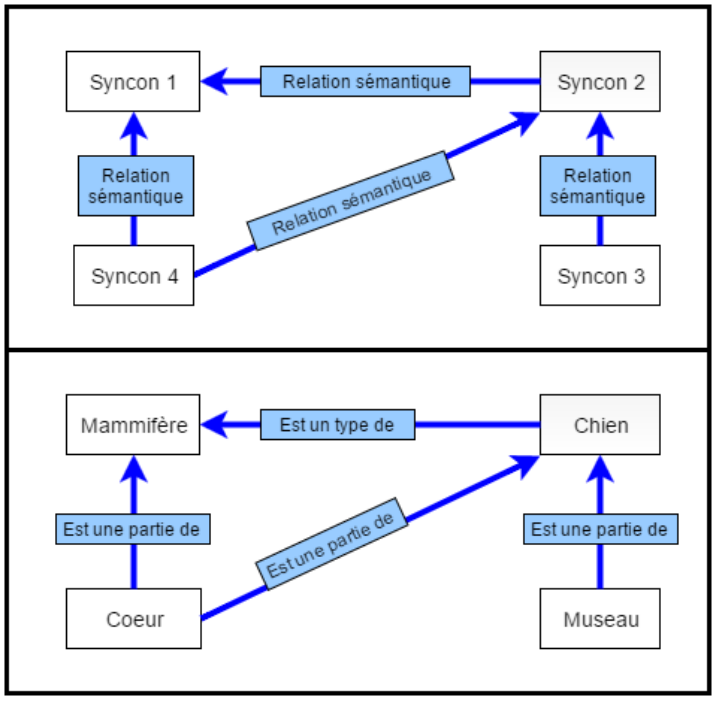


Figure 8 : En haut se trouve une représentation schématique de structure de relations sémantiques possibles. En bas se trouve le même schéma avec un exemple.

E. Attributs directs des syncons

Les syncons sont des unités sémantiques uniques. Il existe trois informations directement associées à un syncon (voir figure 9). Il y a le syncon ID, la catégorie grammaticale et le domaine sémantique. Cette dernière information est facultative contrairement aux deux autres.

1. Les syncon IDs

Chaque syncon est unique et associé à un identifiant unique appelé « syncon ID ». Ces identifiants sont formés d'un croisillon suivi d'un nombre. Il existe deux types de syncon ID : « l'identifiant standard » et « le nouvel identifiant ». Le « QClient » est l'interface d'édition du Sensigrafo FR. Les syncons présents dans cette interface ont un seul Syncon ID. Certaines étapes de développement peuvent provoquer des changements au niveau des identifiants. C'est pour cela qu'un nouvel identifiant est attribué à ce moment. Ainsi les interfaces présentant le Sensigrafo FR après ces étapes présentent des syncons avec deux identifiants sous la forme :

Lemme Sensi : glossa

[Catégorie_Grammaticale] #Nouvel_Identifiant W[#Identifiant_Standard]

Par exemple :

Arbitre : @personne chargée d'arbitrer une rencontre sportive, d'en contrôler la régularité.

[Noun] #666 W[#856]

2. Les catégories grammaticales

Chaque syncon ne peut être associé qu'à une seule catégorie grammaticale. Dans le Sensigrafo FR, uniquement les classes « nom », « adjectif », « adverbe », « verbe » et « nom propre » sont représentées.

3. Les domaines

Le domaine est une information sur le champ sémantique d'un syncon. Quand le terme « hérisson » est utilisé pour faire référence au petit mammifère, le domaine en présence est bien la zoologie par contre lorsque ce terme est utilisé pour désigner l'outil de ramonage le domaine en présence cette fois est le domaine de terme technique.

La liste complète des domaines représentés dans le Sensigrafo FR est disponible à l'annexe 3. Lorsqu'un syncon est détecté à la désambiguïsation, ses domaines associés sont extraits et permettent la désambiguïsation d'autres termes via le contexte sémantique.

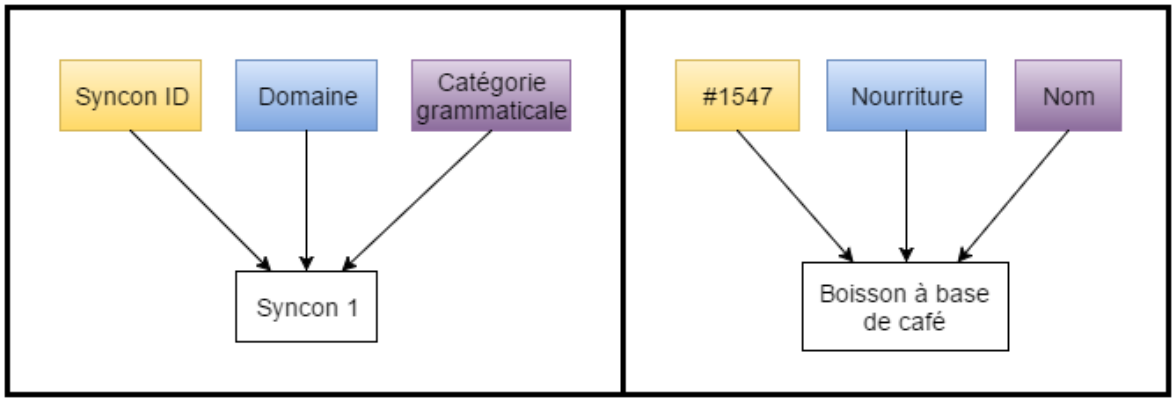


Figure 9 : À gauche se trouve une représentation schématique des attributs directs d'un syncon. À droite se trouve le même schéma avec un exemple.

F. Informations par catégorie grammaticale

La catégorie grammaticale d'un mot est une information élémementaire. Mais cette information d'apparence simple est en fait source d'un grand nombre d'indicateurs. La catégorie grammaticale donne des indications permettant de définir les catégories grammaticales des autres mots grâce à des règles établies. Dans la phrase « *Je montre la lune*

du doigt. » après avoir désambiguïsé le premier mot « *je* », comme étant un pronom, il est possible d'affirmer que le mot « *montre* » est un verbe. Dans ce cas il ne peut être un nom car un pronom ne peut être suivi d'un nom dans une formulation grammaticale. Par la suite les catégories grammaticales permettent de reconnaître les groupes grammaticaux mais aussi les relations syntaxiques. Certaines informations permettent de donner des indications pour la désambiguïsation sémantique. Nous présentons quelques une de ces informations attribuées par catégorie grammaticale.

4. Informations associées aux noms

a) Attributs

Pour les syncons de la catégorie grammaticale « nom commun[10]», il est possible de donner un attribut entre les 5 disponibles : lieu, temps, solide, liquide, gaz. Si un nom a pour attribut « temps » l'information portée par cet attribut permettra de ne pas sélectionner un verbe qui s'applique uniquement à des noms portant l'attribut « solide ».

Prenons l'exemple du lemme Sensi « passer ». Un syncon associé représente le concept de passer physiquement comme dans l'énoncé « *Un train passe dans la gare.* ». Ce syncon porte l'attribut de sujet des noms portant l'attribut « solide ». Un autre syncon pour le lemme Sensi « passer » représente le concept de « passer » utilisé dans un contexte de durée comme dans l'énoncé : « Le temps passe vite. ». Dans ce cas, ce syncon porte l'attribut de sujet de noms portant l'attribut « temps ».

b) Élément suivant

Des informations supplémentaires peuvent aussi être renseigné comme le fait qu'il est probablement positionné avant un nom propre ou un numéro.

Par exemple, le syncon du lemme Sensi « monsieur », dont le concept est le titre donné aux hommes, a une probabilité importante de se placer avant un nom propre comme dans l'exemple : « *Monsieur De Montmiraille n'est pas disponible pour le moment.* »

[10] Noms communs : les noms communs sont les termes qui désigne en général, une personne, un animal ou une chose par opposition au noms propres qui eux désigne des personnes, animaux ou choses en particulier

5. Informations des adjectifs

c) Positionnement

Les syncons de la catégorie des adjectifs peuvent avoir un attribut pour dire s'ils peuvent se présenter avant ou après un nom ou si les deux positions sont possibles. Prenons l'exemple de « sale ». Voici deux exemples dans lesquelles se trouve le mot « sale » en tant qu'adjectif.

« C'est un sale type. »

« C'est un type sale. »

Ils ne désignent pas le même concept. Dans la première phrase « sale » pourrait être associé au syncon exprimant le concept de qualificatif de ce qui est nuisible, méchant, ou désagréable. Ce syncon à l'attribut de positionnement « avant » un nom.

Dans la seconde phrase le mot « sale » pourrait être associé au syncon qui décrit quelqu'un qui néglige les soins de propreté hygiénique élémentaire. Ce syncon porte l'attribut « après » un nom.

d) Classes des noms cibles

Il y a aussi la possibilité de définir la classe des noms qu'un verbe qualifie. Le tableau 2 présente les différentes classes de nom disponibles dans le Sensigrafo FR. Ainsi, il est possible de définir si un adjectif s'applique uniquement à des personnes ou des plantes pour des sujets concrets.

Par exemple pour lemme « liquide », l'acception adjectivale « *Dont la faible cohésion a pour conséquence une mobilité plus ou moins grande des molécules qui, obéissant à la loi de la pesanteur, coulent* ou tendent à couler ; spécialt (cour.), qui est dans cet état aux températures ordinaires.* » aura l'attribut « fluide », qui se trouve dans le tableau 2. Cela signifie que ce syncon s'applique uniquement à des syncons nominaux ayant comme catégorie sémantique [11]« fluide ».

[11] Catégories sémantique : Les catégories sémantiques dans ce document représentent les catégories conceptuelles permettant de regrouper des concepts. La liste des catégories de bases se trouvent dans le tableau 2.

<table>
<tr><th colspan="2">Sujet concret</th><th>Concept</th></tr>
<tr><th>Sujet animé</th><th>Sujet inanimé</th><td rowspan="2">Temps
Mesures
Oscillations
Activités
- Actions
- Huma
Propriétés
Émotions
Évènements
- Occurrences
Communications
- Textes
Processus
Connaissance
États</td></tr>
<tr><td>Personne
Groupe
- Groupe d’humain
- Groupe d’animaux
- Groupe d’objet
- Groupe de plantes
Animaux
Plantes
Partie de
- Corps humain
- Plantes
- Corps d’animaux</td><td>Objets
- Dispositif
- Instrument
- Argent
- Véhicule
- Contenant
Lieux
- Travail architectural
- Bâtiment
Substance
- Solide
- Fluide
• Boisson
- Gaz
Nourriture
Vêtements
Énergie</td></tr>
<tr><th colspan="3">Verbes</th></tr>
<tr><th>Domaine 1</th><th>Domaine 2</th><td></td></tr>
</table>

Tableau 2 : Tableau présentant les différentes classes de noms pouvant être marquées pour les possibilités d’application d’un adjectif ou d’un verbe.

6. Informations des adverbes

e) Attributs

Pour les syncons ayant la classe grammaticale adverbe il est possible de définir s’ils expriment une manière, une quantité, un doute, un temps, une affirmation, une cause, une place, une négation ou s’ils apparaissent dans les syntagmes prépositionnels ou conjonctifs.

Prenons l’exemple du lemme Sensi « doucement » dont le syncon associé est « de manière douce ». Ce syncon a l’attribut « manière » car il décrit une manière dont une chose est effectuée.

f) Positionnement

La position peut elle aussi être précisée. 6 possibilités sont proposées : avant ou après un nom, un verbe ou un adjectif. Par exemple, les adverbes de quantité et de temps,

quand ils sont monosyllabiques, se placent après le verbe comme par exemple, « trop » ou « tard » :

« Il est arrivé **tard**. »

« Nous avons **trop** de responsabilités. »

7. Informations des verbes

g) Attributs

Les syncons de la catégorie des verbes ont une liste d'attributs disponibles qui permet d'exprimer si le verbe est transitif, transitif pronominal, intransitif, intransitif pronominal, impersonnel, réflexif, réflexif réciproque, copulatif ou encore si c'est un verbe de modalité ou auxiliaire. La liste se trouve à l'annexe 4. Ils permettent de désambiguïser des termes. Ainsi, un verbe peut avoir un syncon avec l'attribut transitif et un syncon avec l'attribut intransitif, par exemple : pour le verbe « arrêter » il existe 18 syncons dont 3 d'entre eux se trouvent dans le tableau 3.

Exemple 1
ID : #114762
Glossa : « Arrêter (de...). Interrompre, cesser ce qu'on est en train de faire. »
Attribut : Intransitif
Exemple 2
ID : #114764
Glossa : S'interrompre ou finir (processus, action). | Cesser de couler. | Cesser de passer.
Attribut : Transitif pronominal
Exemple 3
ID : #114748
Glossa : « Interrompre ou faire finir (une activité, un mouvement, un processus, une évolution). »
Attribut : Transitif.

Tableau 3 : Tableau présentant des exemples d'attributs différents entre des syncons associés au même lemme Sensi.

Ainsi, la forme du verbe « arrêter » ne pourra pas être désambiguïsée avec le syncon de l'exemple 2 s'il n'apparaît pas dans une structure pronominale.

h) Classes des substantifs cibles

Il est possible d'indiquer la catégorie d'appartenance des substantifs potentiellement sujets de ce syncon. Ce sont avant tout les catégories principales (voir Tableau 2) qui sont renseignées mais il est possible d'ajouter quelques syncons pour des usages spécifiques. Le processus est le même pour les compléments d'objet direct.

Pour les compléments d'objet indirect il est possible d'indiquer si ce dernier est un fils des catégories Chose, Personne, Animal, Concept ou Lieu.

Prenons le lemme Sensi « donner » dont le syncon verbe décrit l'action de céder, offrir sans rien recevoir en retour. Ce syncon porte l'information que son sujet peut être un fils de la catégorie Personne ou de la catégorie Groupe de personnes. Les compléments d'objet direct de ce verbe doivent être fils d'au moins une des catégories suivantes : Animal, Plante, Objet, Boisson, Aliments, Vêtements. Les compléments d'objets indirects peuvent être fils d'une des deux catégories Personne ou Animal.

À ce jour, le Sensigrafo FR compte 247 748 lemmes Sensi pour 261 026 Syncons. Les noms représentent 39.7% des syncons avec 103 742 entrées, les verbes représentent 9.6% avec 24 986 entrées, les adjectifs représentent 11.4% avec 29 864 entrées. Les adverbes représentent, eux, 1.4% avec 3 759 entrées et les noms propres représentent 37.8% avec 98 674 entrées.

Nous connaissons à présent la composition du Sensigrafo FR et nous savons que c'est un réseau qui est constitué d'un certain nombre d'informations. Nous allons voir par la suite comment certaines de ces informations sont renseignées automatiquement par des stratégies de développement.

3. Constitution d'une base dictionnairique

La confection de la base dictionnairique est la première étape dans le développement d'un réseau monolingue Sensigrafo. La base dictionnairique représente un lexique de la langue ciblée par le développement et ce sont les syncons qui représentent les différents concepts de la langue instanciés par des formes lexicales. Cette étape consiste à extraire des définitions avec leur forme lexicale d'instanciation. D'autres informations peuvent être extraites dans le même temps selon les standardisations de structurations présentes dans les ressources utilisées.

4. Mise en correspondance

La mise en correspondance est une étape très importante dans le développement d'un Sensigrafo monolingue. Cette étape consiste à associer un syncon d'une langue cible avec un syncon d'une langue source. La langue cible est la langue pour laquelle le Sensigrafo est en cours de développement contrairement à la langue source qui, elle, est la langue pour laquelle le développement du Sensigrafo est terminé. De plus, cette langue source est utilisée pour le développement de la langue cible. Cette mise en correspondance doit prendre en compte toutes les informations décrites précédemment : attributs de relations lemme Sensi/syncon, relation syncon/syncon, attributs directs de syncon, informations par catégories grammaticale etc. Toutes ces informations doivent être prises en compte lors de cette étape car une fois qu'un syncon de la langue cible est associé à un syncon de la langue source, le premier calque toutes les informations du second pour se les approprier. Le calque est produit lors de l'étape de développement suivante, la dévirtualisation. Il s'approprie également les relations hiérarchiques conceptuelles du syncon de la langue source et ainsi les chaînes hiérarchiques conceptuelles également. En théorie, il est possible de remonter une chaîne hiérarchique jusqu'à une des 11 catégories de base.

Prenons l'exemple du concept du chat de la race des chartreux. Il est possible d'affirmer :

- Un chartreux est un type de chat.
    - Un chat est un type de félin.
        - Un félin est un type de carnivore.
            - Un carnivore est un type de mammifère.
                - Un mammifère est un type de vertébré.
                    - Un vertébré est un type d'animal.

Le concept d'animal est l'un des 11 syncons « catégorie de base ». Cet exemple est un exemple de chaîne conceptuelle hiérarchique.

Le fait de calquer toutes ces informations permet de ne pas avoir à les ajouter lors de la conception de la base dictionnairique et permet ainsi d'optimiser le temps de travail.

Dans la figure 10 se trouve une représentation schématique comparant l'état du Sensigrafo EN et le Sensigrafo FR avant la mise en correspondance avec un exemple. Le Sensigrafo FR a une base dictionnairique stable mais n'a pas encore de relation sémantique entre les syncons contrairement au Sensigrafo EN et certaines informations relatives aux lemmes Sensi et syncons sont manquantes. À cette étape du développement du Sensigrafo

FR, pour obtenir un réseau complet, les relations sémantiques, les chaînes hiérarchiques ainsi que les informations associées aux lemmes Sensi et aux syncons doivent être renseignées. Pour éviter d'ajouter manuellement ces informations, les syncons du Sensigrafo FR vont être associés avec des syncons du Sensigrafo EN comme représenté dans la figure 11. Les relations sont les mêmes entre les différents syncons peu importe la langue puisque c'est une mise en correspondance par concept prenant en compte les aspects sémantiques mais également grammaticaux. Lorsque la mise en correspondance est effectuée, les relations vont être calquées automatiquement lors de la dévirtualisation. Un Sensigrafo en développement est « construit » à l'aide d'un Sensigrafo terminé. Ainsi, une fois que les lemmes et leur syncons sont implantés, l'étape d'association de correspondance de concept est effectuée. Cette action est également appelée « mapping ».

D'après les figures 10 et 11, il est possible de dire que le Syncon 1 de la langue 1 appelé (A) est l'équivalent du syncon 1 de la langue 2 (B). Le syncon 2 de la langue 1 (C) est l'équivalent du syncon 2 de la langue 2 (D). **Si** (A) et (B) ont un lien (X) **et** (A) est l'équivalent de (C) et (B) l'équivalent de (D) **Alors** (C) et (D) ont la même lien (X) que (A) avec (B). Ainsi, après la mise en correspondance, le syncon du Sensigrafo FR associé au lemme Sensi « cat. animal » deviendra un hyperonyme du syncon du Sensigrafo FR associé au lemme Sensi « lion ». Ceci est un exemple des effets de la mise en correspondance. Il est important de garder à l'esprit que les différentes informations d'un syncon de la langue source mis en lien avec un syncon de la langue cible seront calquées lors de la dévirtualisation. Par exemple le syncon du Sensigrafo EN associé au lemme Sensi « lion » a l'attribut « solide ». L'attribut « solide » défini les éléments solides comme le bois ou un être humain par opposition aux attributs « liquide », « gaz », « lieu » et « temps ». Une fois mis en correspondance avec le syncon du Sensigrafo FR associé au lemme Sensi « lion » suivi de la dévirtualisation, ce dernier aura également l'attribut « solide ».

La mise en correspondance est une étape partiellement automatisée à ce jour. Environ 100 000 syncons ont été mis en correspondance de manière automatique entre le Sensigrafo FR et le Sensigrafo IT.

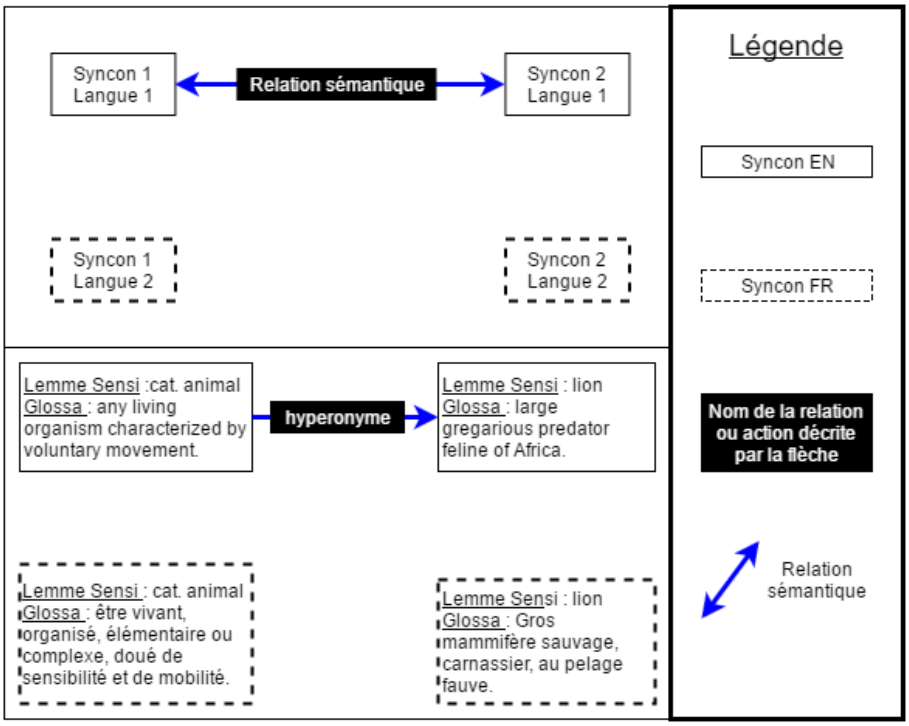


Figure 10 : L'encadré du haut est une représentation schématique de l'état du Sensigrafo EN et du Sensigrafo FR avant la mise en correspondance ; L'encadré du bas est un exemple reprennant la représentation du haut.

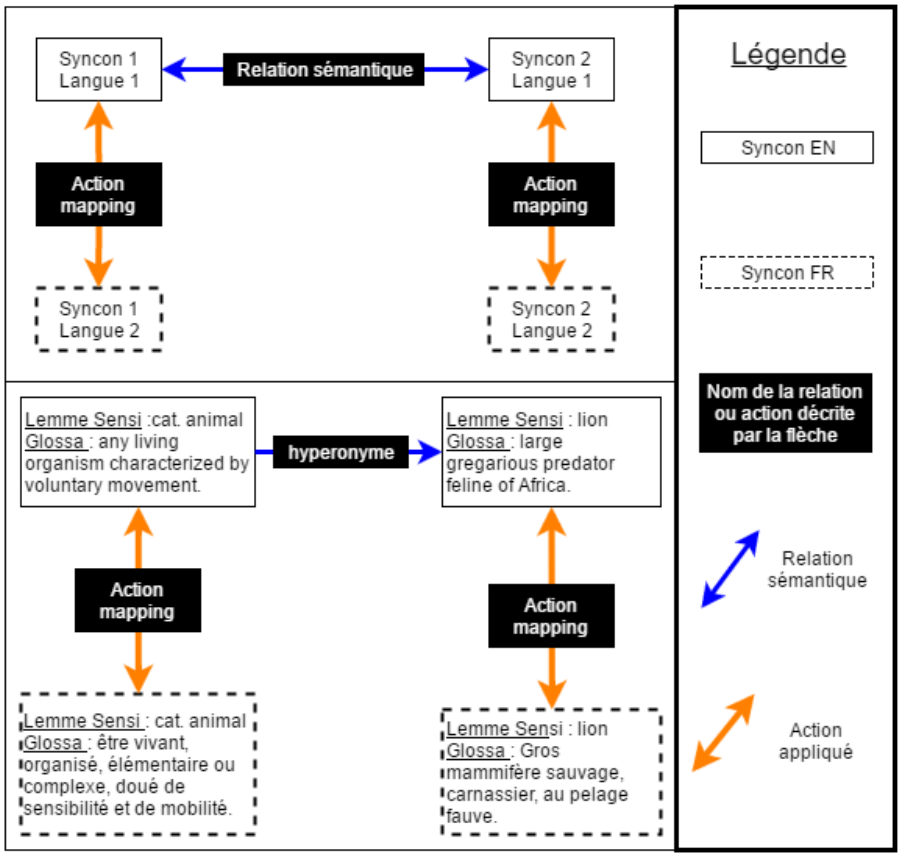


Figure 11 : L'encadré du haut est une représentation schématique de la mise en correspondance du Sensigrafo EN et du Sensigrafo FR ; L'encadré du bas est un exemple reprenant la représentation du haut.

5. Dévirtualisation

La dévirtualisation est une étape dans le développement d'un réseau Sensigrafo monolingue qui est effectuée après la mise en correspondance. Cette étape s'applique une fois que la base dictionnairique de la langue cible est constituée. Avant tout, il faut savoir qu'il existe en réalité deux Sensigrafos monolingues pour chaque langue :

- Le Sensigrafo éditable : C'est la référence pour les linguistes qui réalisent différentes tâches comme l'ajustement des fréquences ou la mise en correspondance. Il est géré en ligne sur le réseau interne d'Expert System. L'interface d'édition est le QClient.

- Le Sensigrafo fixe : Il possède toute la cohérence et l'exhaustivité nécessaires pour fonctionner dans les systèmes Expert Système. C'est une base de données autonome utilisée par le moteur d'analyse sémantique implantée dans les outils.

La dévirtualisation est la transformation d'un Sensigrafo éditable en Sensigrafo fixe. Il produit un réseau qui fonctionne à partir de données non connectées et abstraites. La dévirtualisation suit plusieurs stratégies pour résoudre certains cas problématiques rencontrés. Ces stratégies sont mises en place lors de la mise en correspondance pour que la dévirtualisation s'applique correctement même aux situations problématiques. Quatre stratégies sont présentées :

- Le **clonage de réseau** pour éviter qu'un syncon d'une langue source soit mis en correspondance avec plusieurs syncons d'une langue cible. Cette stratégie est l'approche classique appliquée dans la plupart des cas.
- Le **regroupement de synonymes** permet d'éviter le surnombre de syncons.
- Le **saut de nœuds vides** pour résoudre les problèmes de chaînes hiérarchiques conceptuelles incomplètes dans une langue cible.
- Le **mapping imprécis** pour résoudre le cas où aucune mise en correspondance n'est possible pour un syncon d'une langue cible ; c'est-à-dire qu'un syncon d'une langue cible est trop différent pour être mis en correspondance de manière cohérente avec un syncon de la langue source, en tant qu'équivalent.

A. Le clonage de réseau

Un syncon d'une langue cible mis en correspondance avec un syncon d'une langue source hérite des attributs et des liens de ce dernier. Les relations hiérarchiques conceptuelles

sont incluses. Le clonage de réseau est le résultat de l'action d'héritage des informations effetuée sur le réseau mis en correspondance, lors de l'étape précédente de mise en correspondance. La figure 12 est une représentation schématique de cette action d'héritage.

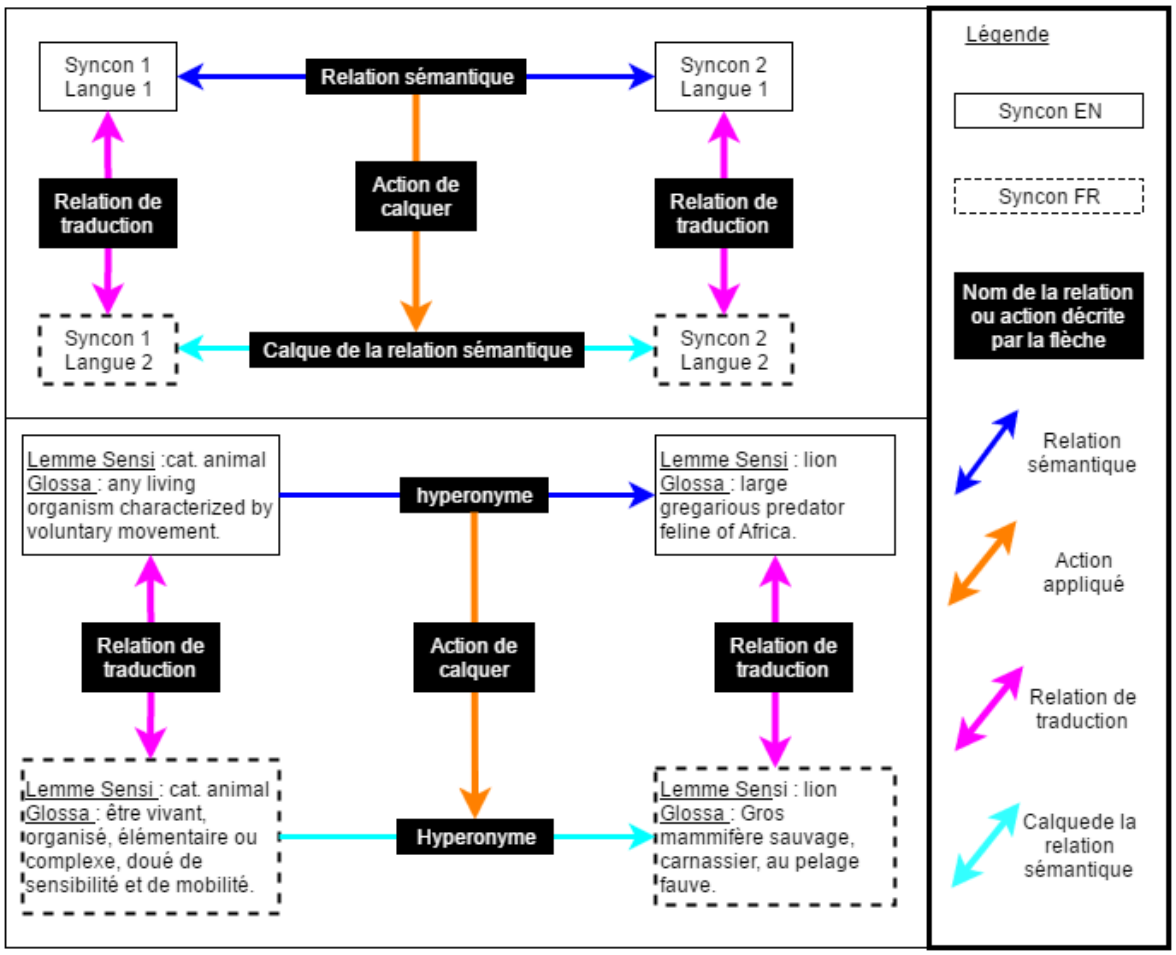


Figure 12 : L'encadré du haut est une représentation schématique des actions qui suivent la mise en correspondance du Sensigrafo EN et du Sensigrafo FR ; L'encadré du bas est un exemple reprennant la représentation du haut.

B. Le regroupement de synonymes

Tous les syncons d'une langue cible mis en correspondance, de manière identique à un syncon de la langue source sont fusionnés en un seul syncon. Le glossa et l'identifiant gardé sont ceux du premier syncon mis en relation avec le syncon de la langue source. Par exemple dans le Sensigrafo éditable FR deux syncons existent : l'un est associé au lemme Sensi « jeunot » et le second est associé au lemme Sensi « gars ». Les figures 13 et 14 sont les captures d'écran de ces deux syncons effectuées dans le QClient. « Jeunot » est associé à l'identifiant standard #39462 et « gars » à l'identifiant standard #34269.

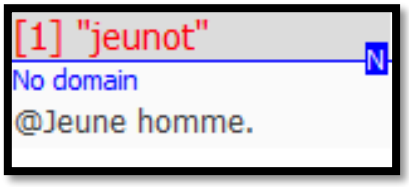


Figure 13 : Capture d'écran du syncon portant le lemme Sensi « jeunot » et l'identifiant #39462 dans le Sensigrafo éditable.

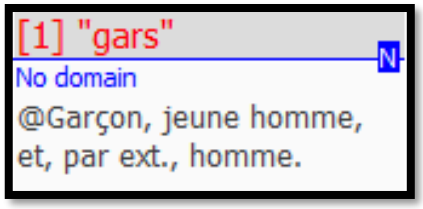


Figure 14 : Capture d'écran du syncon portant le lemme Sensi « gars » et l'identifiant #34269 dans le Sensigrafo éditable.

Dans le Sensigrafo FR fixe un syncon existe avec les lemmes Sensi « gars » et « jeunot ». Les identifiants associés à ce syncon sont : [Noun]#25856 W[#34269]. L'identifiant standard correspond à l'identifiant du syncon de « gars » dans le QClient.

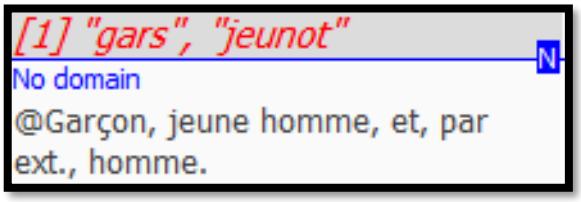


Figure 15 : Capture d'écran du syncon portant les lemmes Sensi « gars » et « jeunot » et dont le nouvel identifiant est #25856 et l'identifiant standard est #34269 dans le Sensigrafo fixe.

La conclusion de cet exemple est que le syncon associé à « gars » et le syncon associé à « jeunot » ont été mis en correspondance avec un syncon identique du Sensigrafo EN. Le syncon associé à « gars » a été mis en relation avec le syncon du Sensigrafo EN avant le syncon associé à « jeunot ».

Le regroupement de synonymes implique que le nombre de syncons dans le Sensigrafo fixe est moins important que dans le Sensigrafo éditable. Le regroupement de synonymes implique également que les différents Sensigrafos fixes ont des structures hautement similaires mais des syncons très différents.

C. Le saut de nœuds vides

Le saut de nœud est une stratégie de développement pour les cas où une chaîne hiérarchique conceptuelle n'est pas complète dans une langue cible. Une chaîne hiérarchique conceptuelle incomplète empêche l'héritage d'un certain nombre d'information. Cette stratégie permet, une fois que la mise en correspondance a été effectuée, d'appliquer l'héritage des informations également dans les cas où les chaînes hiérarchiques ont un maillon manquant. Pour l'exemple qui suit nous choisissons de prendre la relation sémantique d'hyperonymie mais il est possible d'appliquer cette stratégie avec n'importe quelle relation sémantique définie à l'annexe 1 de ce document.

Prenons le cas d'un syncon (A) dans la langue cible. Il est mis en correspondance avec un syncon (B) dans la langue source, dont le père « hyperonyme » (C) n'a pas encore été mis en correspondance avec un syncon de la langue cible. Aucun père hyperonyme n'est disponible pour (A) dans la langue cible comme le représente la figure 16.

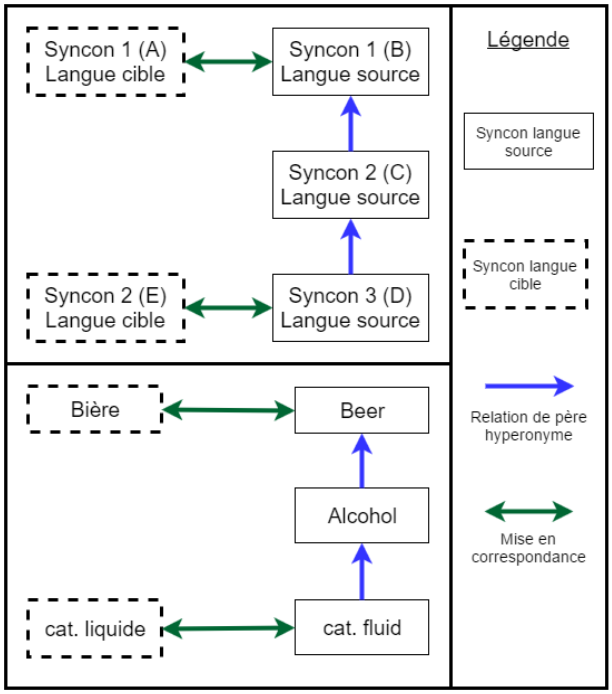


Figure 16 : Représentation schématique d'une situation où la chaîne hiérarchique conceptuelle de la langue cible compte un nœud vide.

Cependant, il est possible que dans la langue source, (C) possède un père hyperonyme. En effet, (D) est un père hyperonyme de (C). De plus il est mis en correspondance avec le syncon (E) dans la langue cible. Donc (E) va être mis en correspondance avec (C) en tant père hyperonyme dans la langue cible (voir figure 17).

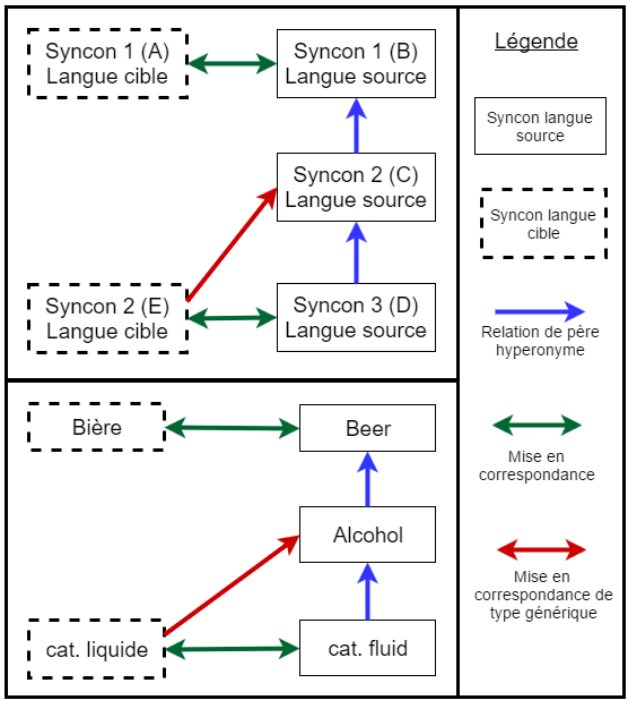


Figure 17 : Représentation schématique d'une situation où la chaîne hiérarchique conceptuelle de la langue cible compte un nœud vide et où une mise en correspondance a été ajouté entre les syncon (E) et (C).

Ainsi, (E) va apparaître comme père hyperonyme de (A). Le nœud vide se situait entre les syncons (A) et (E). L'héritage des informations et des relations va pouvoir être effectué (voir figure 18).

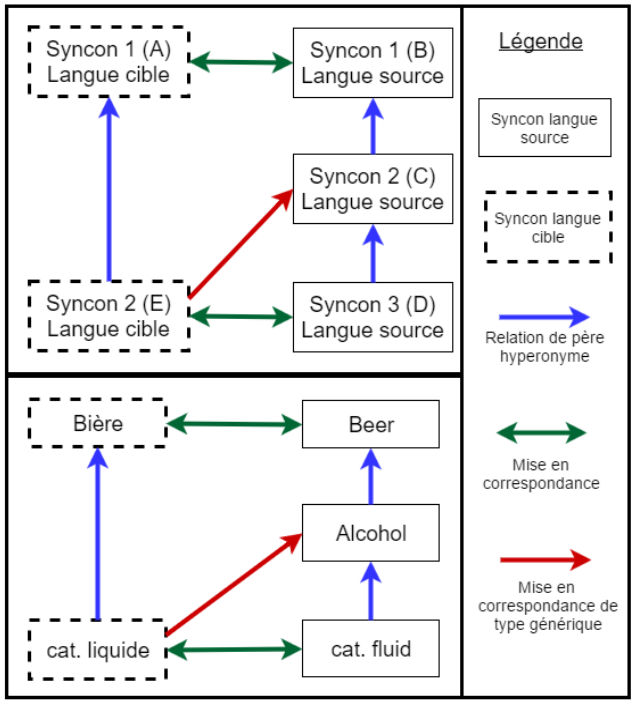


Figure 18 : Représentation schématique d'une situation où la chaîne hiérarchique conceptuelle de la langue cible compte un nœud vide et l'ajout d'une mise en correspondance permet de reconstruire la chaîne hiérarchique conceptuelle.

Il est possible de sauter plusieurs nœuds vides à la suite. Toutes les catégories sont mises en correspondance en premier lieu et les nœuds de complétion sont produits de manière périodique.

D. La mise en correspondance imprécise

Il est possible qu'un syncon d'une langue cible n'ait pas d'équivalent conceptuel dans la langue source. Ce phénomène peut provenir d'un champ sémantique trop large ou au contraire trop restreint pour considérer une mise en correspondance conceptuelle correcte.

Dans le cas d'un champ sémantique restreint, il est possible qu'un syncon générique existe dans la langue source : c'est une forme de catégorisation. Il est rare de rencontrer un cas qui ne peut pas être catégorisé. Dans ce cas, le syncon générique est considéré comme un fils hyperonyme d'un syncon de la langue source. Lors de la mise en correspondance, il est possible d'indiquer que le syncon de la langue source est un générique du syncon de la langue cible. Cette information fournie par la mise en correspondance permet d'appliquer l'héritage d'informations lors de la dévirtualisation (voir figure 19 et 20).

Prenons le cas où le champ sémantique du syncon de la langue cible est restreint comme le cas du dépaysement. Le concept de dépaysement en français est une *émotion ressentie lors d'un changement d'habitude ou d'environnements*. Le dépaysement désigne également les sentiments qui apparaissent lors d'une immersion dans un environnement différent de l'environnement habituel ou d'origine mais également dans les environnements inconnus. Ce phénomène peut, ainsi, intervenir lors de changements de lieu de vie. Aucun syncon ne définit ce concept dans le Sensigrafo EN éditable.

Cependant, comme indiqué dans la définition, le dépaysement est une émotion. Donc notre syncon est un type d'émotion. Le syncon qui définit le concept d'émotions existe en anglais. Alors notre syncon « dépaysement » du Sensigrafo FR éditable est mis en correspondance avec le syncon « émotion » du Sensigrafo EN éditable. Cette mise en correspondance précise que le syncon du Sensigrafo EN éditable est un générique du syncon présent dans le Sensigrafo FR éditable. Cette mise en correspondance imprécise permet d'appliquer un héritage lors la dévirtualisation.

Le cas d'un champ sémantique large n'est pratiquement jamais rencontré. Nous ne détaillerons pas ce cas.

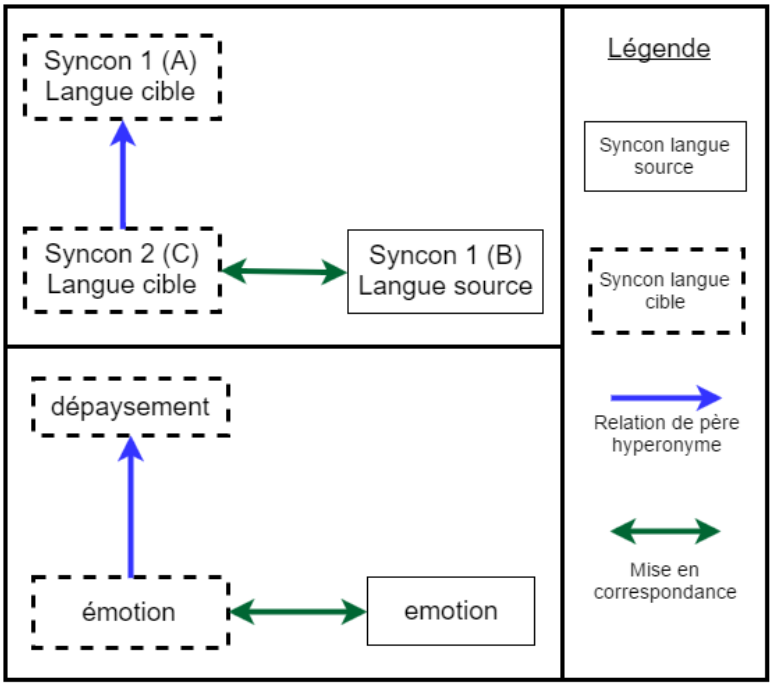


Figure 19 : L'encadré du haut est une représentation schématique d'une situation d'absence d'équivalent conceptuel entre une langue source et une langue cible. L'encadré du bas est un exemple reprenant la représentation du haut.

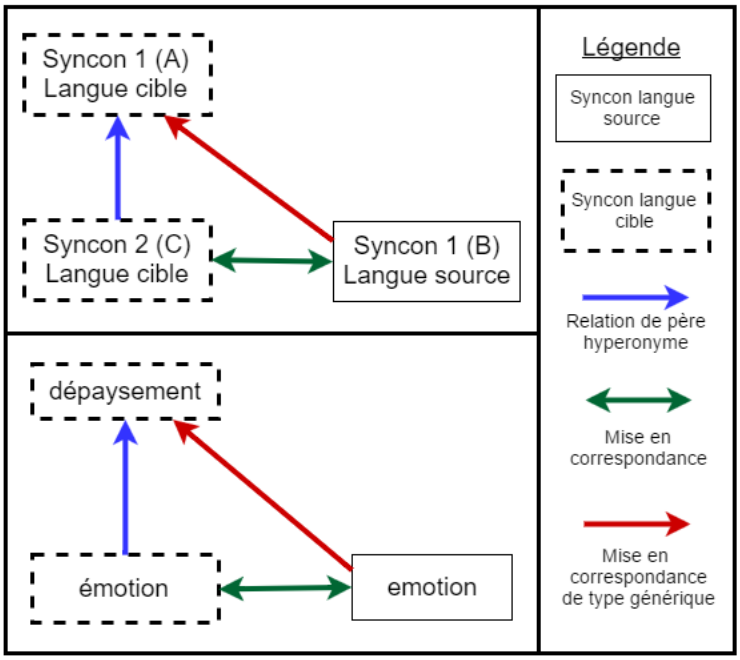


Figure 20 : L'encadré du haut est une représentation schématique d'une situation d'absence d'équivalent conceptuel entre une langue cible et une langue source résolu par la stratégie de mise en correspondance imprécise ; L'encadré du bas est un exemple reprenant la représentation du haut.

La dévirtualisation est permet d'obtenir un réseau sémantique fixe qui pourra être implanté dans le moteur d'analyse sémantique qui lui-même sera implanté dans des outils de traitement sémantique.

6. La Compilation

La compilation consiste à produire un fichier exécutable comprenant la dernière version du Sensigrafo monolingue fixe. Nous savons que lors de cette étape, plusieurs ressources sont utilisées en plus du Sensigrafo monolingue pour enrichir le contenu d'information.

Des bases de données géographiques sont utilisées pour extraire automatiquement des lemmes du domaine de la géographie et les syncons de noms propres sont constitués à partir d'extractions issues d'encyclopédies libre d'utilisation.

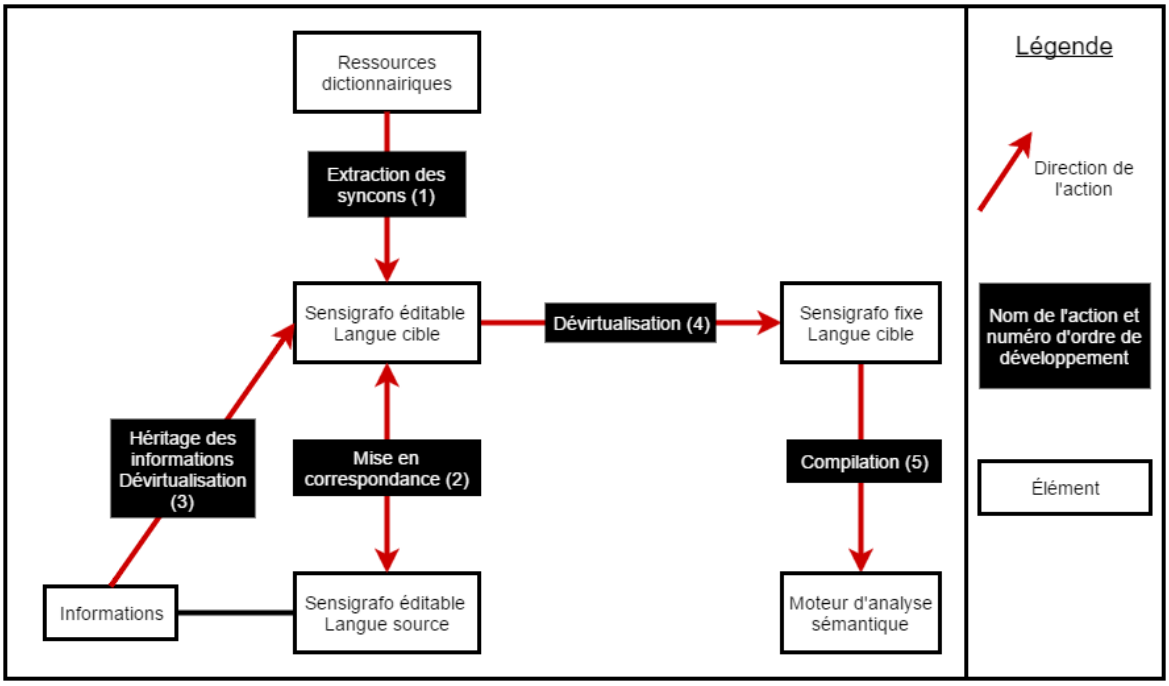


Figure 21 : Représentation schématique des relations entre les Sensigrafo éditables, fixes et les moteurs d'analyses sémantique.

La figure 21 est une représentation schématique du développement d'un Sensigrafo monolingue. La constitution de la base dictionnairique s'effectue par l'action (1) de l'extraction des syncons des ressources dictionnairiques vers le Sensigrafo éditable. La seconde étape de la mise en correspondance (2) s'effectue entre le Sensigrafo éditable de la langue cible et celui de la langue source. La dévirtualisation compte deux actions. La première action est l'héritage des informations (3), autrement dit le calque des informations entre les syncons mis en correspondances. La seconde action de la dévirtualisation est la dévirtualisation proprement dite, c'est-à-dire la transformation d'un réseau virtuel en un réseau concret et effectif qui pourra être implanté lors de la compilation (5) dans un moteur d'analyse sémantique. Ce moteur d'analyse sémantique est utilisé dans les différents outils de l'entreprise tel que le Cogito Desambiguator ou le XTagger.

7. Les outils utilisés

A. Logiciel - COGITO Desambiguator

Le « Cogito Desambiguator » est un logiciel interne[12] et local[13], développé par Expert System, qui permet d'obtenir une désambiguïsation complète d'un texte. Ce logiciel utilise le moteur d'analyse sémantique dans lequel le Sensigrafo FR fixe est implanté. Il fournit une désambiguïsation grammaticale, sémantique et syntaxique. Les différentes désambiguïsations sont effectuées par un analyseur codé en C++ qui prend en compte les différentes informations présentes dans le Sensigrafo FR présentées dans la Partie 1 ; Chapitre 2 ; 2. Sensigrafo monolingue ; points B., C., D., E. et F.. Toutes ces informations sont croisées pour des désambiguïsations optimales car ces dernières s'influencent les unes les autres. Le Cogito Desambiguator est un outil de développement. Il permet d'observer les désambiguïsations effectuées par le moteur d'analyse sémantique et ainsi de calculer la qualité de ce dernier.

La figure 22 est une capture d'écran du Cogito Desambiguator. Dans l'encadré 1 se trouve le texte traité. L'encadré 2 renseigne des différents domaines reconnus lors de l'analyse dans le texte traité. Dans l'encadré 3 se trouve l'analyse syntaxique. Cette analyse se fonde par rapport à un verbe reconnu dans le texte auquel sont affectés les différents éléments (sujet, objets et compléments). L'encadré 4 contient l'analyse grammaticale et sémantique. Dans l'encadré (a) se trouve les catégories grammaticales des mots. Le nombre qui se trouve à côté de la catégorie grammaticale fait référence au nombre de syncons associé à la forme lexicale. Dans cette figure 22, il est indiqué que le lemme reconnu pour le mot « chien » compte 8 syncons. Dans l'encadré (b) se trouve les regroupements des groupes grammaticaux ainsi que leur nature. Les dénominations complètes de ces sigles se trouvent dans le lexique des abréviations page 98. Au-dessus de ces encadrés se trouvent différents boutons d'actions.

[12] Interne : Un logiciel interne désigne un logiciel existant et utilisé uniquement au sein d'une entreprise.
[13] Local : Un logiciel en local désigne un logiciel fonctionnant de manière autonome sur un poste informatique par opposition à un logiciel en réseau qui est connecté aux serveurs.

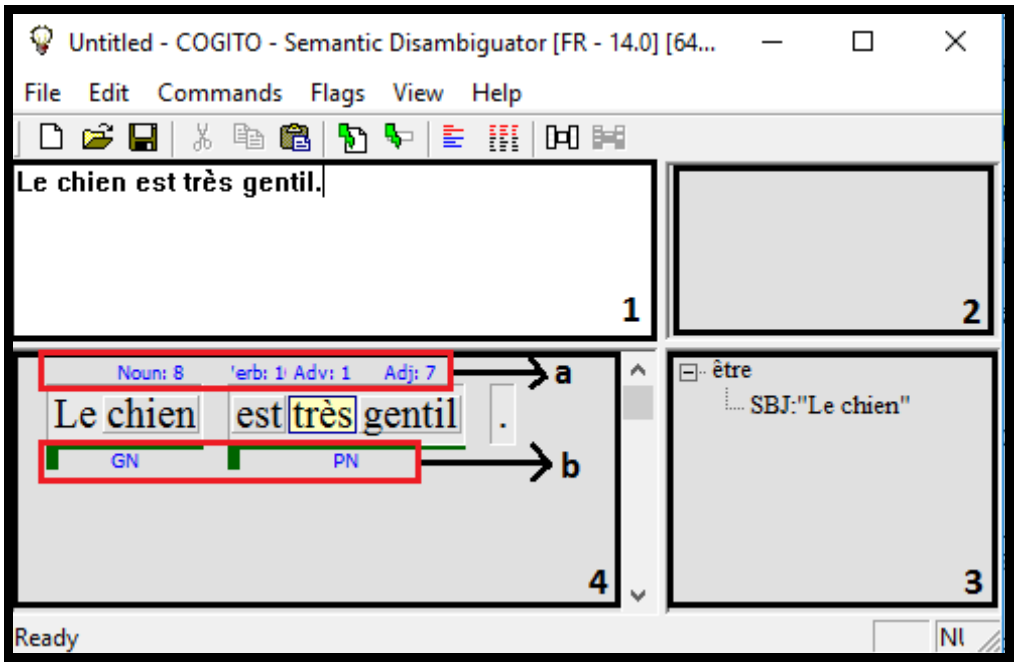


Figure 22 : Capture d'écran du Cogito Desambiguator avec la phrase « Le chien est gentil. » désambiguïsée.

Il existe une visualisation du réseau Sensigrafo FR fixe qui permet de consulter les syncons disponibles. Ainsi, lorsqu'une erreur de sémantique est détectée, il est possible de consulter le Sensigrafo FR fixe pour confirmer si le syncon correspondant au concept dans le texte, instancié par une certaine forme lexicale, existe bien dans le réseau.

Cet outil a été utilisé principalement lors de la tâche de test de performance du Sensigrafo FR en tout début de stage dont la description est établie au Chapitre 1 de la Partie 2 de ce document.

B. Logiciel - Xtagger

Le Xtagger est un logiciel local développé par Expert System. Il est également appelé « Cross Tagger ». Ce logiciel utilise le moteur d'analyse sémantique dans lequel le Sensigrafo FR fixe est implanté. Cet outil permet de visualiser les désambiguïsations produites par le moteur d'analyse sémantique. Il permet également d'extraire les informations issues de ces désambiguïsations et par conséquent établir des statistiques sur les performances. Les deux fonctionnalités disponibles que nous avons manipulées lors du stage porte sur l'étiquetage des entités nommées et l'étiquetage des erreurs de désambiguïsation.

Les étiquetages effectués et sauvegardés produisent trois documents :

- Un document au format RTF : ce fichier comprend le texte traité et les étiquettes des entités nommées.

- Un fichier au format log : ce document comprend toutes les informations relatives aux informations issues de désambiguïsations, c'est-à-dire, les catégories grammaticales, les groupes grammaticaux, les identifiants des syncons reconnus, les relations syntaxiques ainsi que l'intégralité des erreurs de désambiguïsation étiquetées. Pour les erreurs de désambiguïsation, il est renseigné plusieurs informations : le type de l'erreur, la gravité de l'erreur, la position dans le texte, la catégorie grammaticale (ou le type d'entité nommée lorsqu'une entité nommée est reconnue), le lemme reconnu pour la forme lexicale, la forme lexicale et le numéro de syncon reconnu.
- Un document au format Txt : ce document comprend le texte d'origine tel qu'il a été inséré lors du premier traitement.

Plusieurs versions du XTagger ont été développées par Expert System lors du stage apportant ainsi différentes fonctionnalités. La fonction pour l'étiquetage des erreurs est arrivée au cours du stage par exemple.

1. Étiquetage des entités nommées

L'outil permet d'obtenir un premier étiquetage automatique des entités nommées. Il est possible par la suite de modifier cet étiquetage à partir du XTagger. La figure 23 est une capture d'écran du XTagger pour l'étiquetage des entités nommées. Dans l'encadré 1 se trouve les boutons des types d'entités nommées possibles. La liste des entités nommées se trouve dans le tableau 15 et leurs descriptions se trouvent à l'annexe 5. Dans l'encadré 2 de la figure 23 se trouve les entités étiquetées dans le texte se trouvant dans l'encadré 3. Cette figure 23 montre que le texte compte deux entités : « Julie » qui est une entité du type « NPH » (Personne) dont le syncon associé est celui avec l'identifiant #129409. Ce syncon a comme glossa « nom propre féminin ». La seconde entité nommée présente dans le texte est « Grenoble » qui est étiquetée en tant que « GEO » (entité de géographique administrative) avec le syncon portant l'identifiant #13000620. Ce syncon a comme glossa « chef-lieu en Rhône-Alpes (France/Europe) ».

Le sujet de l'étiquetage des entités nommées dans le XTagger est détaillé dans la Partie 2 ; Chapitre 3.

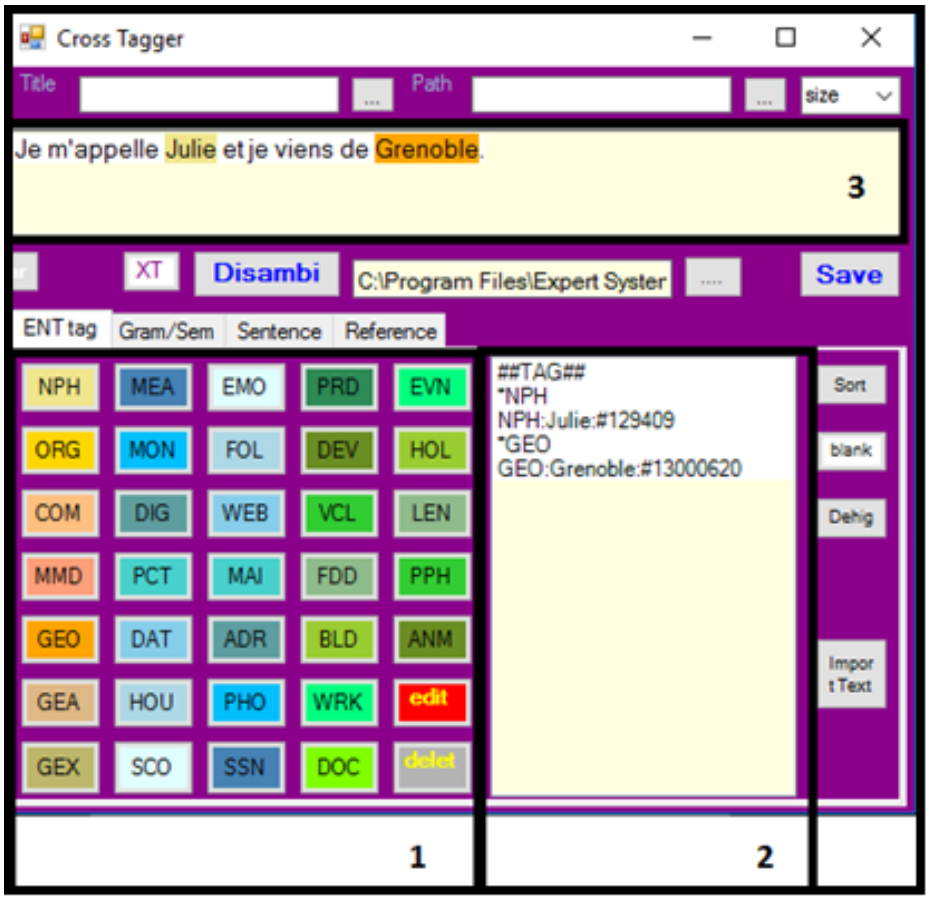


Figure 23 : Capture d'écran de la partie pour l'étiquetage des entités nommées du XTagger version 1.

2. Étiquetage des erreurs

L'étiquetage des erreurs de désambiguïsation est une tâche uniquement manuelle puisqu'elle permet d'indiquer les erreurs produites par le moteur d'analyse sémantique. Pour la partie de l'étiquetage des erreurs, il est possible de se référer à la figure 24. L'encadré 1 de cette figure renferme le texte traité. Dans l'encadré 2 se trouvent les types d'erreurs qu'il est possible d'étiqueter et les 3 niveaux de gravité (léger = gravité 1, lourd = gravité 2, horrible= gravité 3) de ces erreurs sont présentes dans l'encadré 3. Les erreurs ne concernent que des erreurs de sémantique, de catégorie grammaticale et d'entités nommées. Les types d'erreurs et les niveaux de gravité sont détaillés dans la partie 2, chapitre 5 de ce document. Quand un mot dans l'encadré 1 est sélectionné, l'encadré 4 affiche la catégorie grammaticale et les lemmes du syncon avec lequel il a été désambiguïsé. L'encadré 5 indique le glossa de ce syncon. Ces encadrés permettent une visualisation rapide des désambiguïsations grammaticale et sémantique faites sur le texte. L'encadré 6 affiche les erreurs qui ont été étiquetées par l'annotateur. Par exemple dans la figure 24, l'adjectif « préféré » a été désambiguïsé avec le syncon de type verbe portant les lemmes d'instanciation « distinguer »,

« préférer » et « aimer ». Alors, l'élément sera étiqueté « 2:Missing Concept:préféré,75742 ».

Le deux correspond à la gravité 3, « Missing Concept » signifie que le syncon correct n'existe pas, « préféré » est l'élément concerné tel qu'il est reconnu dans le texte et 75742 est l'identifiant standard du syncon reconnu.

Le dernier encadré numéro 7 renseigne le nombre d'éléments présents dans le texte qui ont été désambiguïsés c'est-à-dire les mots pleins. La figure informe que 6 éléments ont été désambiguïsés (« fromage », « préféré », « Paul », « est », « boulette » et « Avesnes »). Les éléments « Avesnes » et « Paul » sont surlignés car ils ont été reconnus comme étant des entités nommées.

Le sujet de l'étiquetage des erreurs dans le XTagger est détaillé dans la Partie 2 ; Chapitre 5.

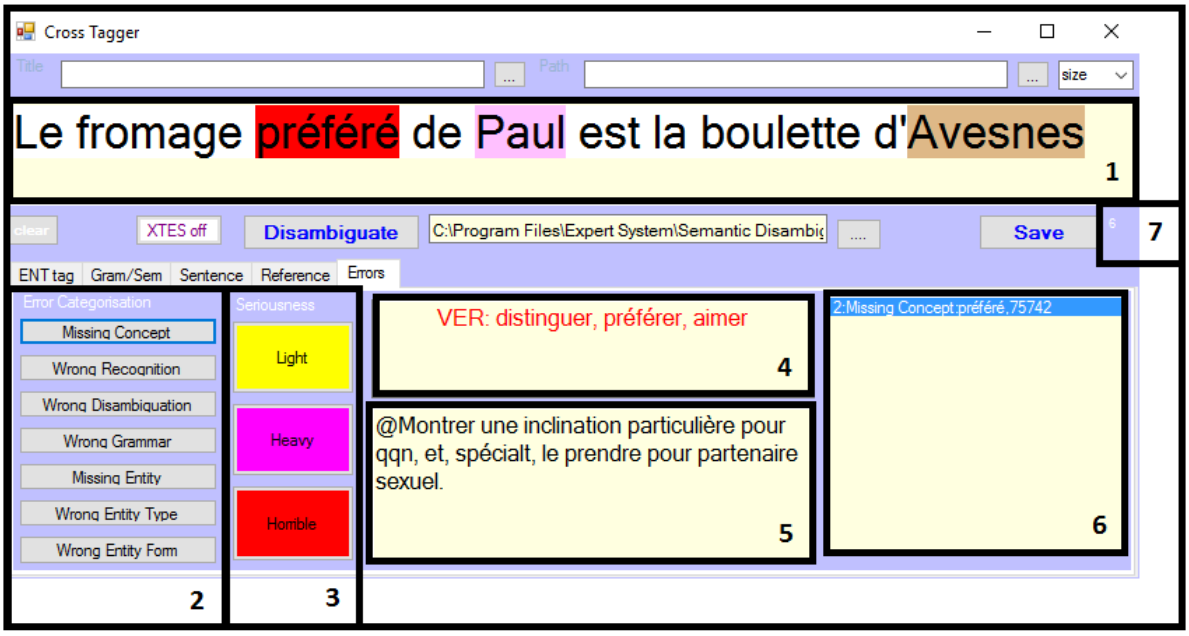


Figure 24 : Capture d'écran de la partie pour l'étiquetage des erreurs de désambiguïsation du XTagger version 2.

C. Interface - Qclient

Le Qclient est une interface en réseau[14], développée par Expert System, qui permet de naviguer et de modifier les données manuellement dans le Sensigrafo éditable selon la langue sélectionnée. Différentes actions sont possibles :

[14] En réseau : Une interface en réseau signifie que les modifications sont prises en compte directement sur le serveur par opposition au modèle en local où un logiciel peut être autonome sur un poste informatique.

- Modifier les informations comme la fréquence, les domaines, les attributs, les registres.
- Lier des syncons dans une même langue avec des relations sémantiques.
- Ajouter ou supprimer un syncon.
- Mettre en correspondance un syncon d'une langue cible et un syncon d'une langue source.
- D'autres actions sont possibles mais elles n'ont pas été abordées lors de ce stage.

La figure 25 est une capture d'écran de l'interface QClient. L'encadré 1 présente les informations relatives à la langue cible. Ici la langue cible est le français. L'encadré 2 présente les informations relatives à la langue source. Ici la langue source est l'anglais. La colonne 4 affiche les acceptions pour une forme lexicale recherchée dans la barre de recherche 8. Le syncon sélectionné s'affiche en blanc sur un fond bleu. Les syncons à fond vert sont des syncons mis en correspondance avec des syncons de langues sources. L'encadré 7 renseigne les informations relatives au syncon sélectionné comme son identifiant, ses attributs etc… La colonne 3 présente les pères du syncon sélectionné dans la colonne 4 selon la relation sémantique indiqué dans le menu déroulant 10. La colonne 5 présente les fils du syncon sélectionné dans le colonne 4 selon la relation sémantique indiquée dans le menu déroulant 10. La colonne 6 présente les syncons de la langue opposée mis en correspondance avec le syncon sélectionné dans la colonne 4. Le menu déroulant 9 permet de voir tous les lemmes Sensi où le mot renseigné dans la barre de recherche 8 est présent ainsi que les autres lemmes Sensi par ordre alphabétique (voir figure 27).

Dans cette figure 25, la forme lexicale recherchée est « fauve » dans la barre de recherche 8. Cette recherche propose comme lemme Sensi « fauve » dans le menu déroulant 9. Il est possible d'ouvrir le menu déroulant pour chercher un autre lemme Sensi contenant le mot « fauve » ou les lemmes Sensi proches.

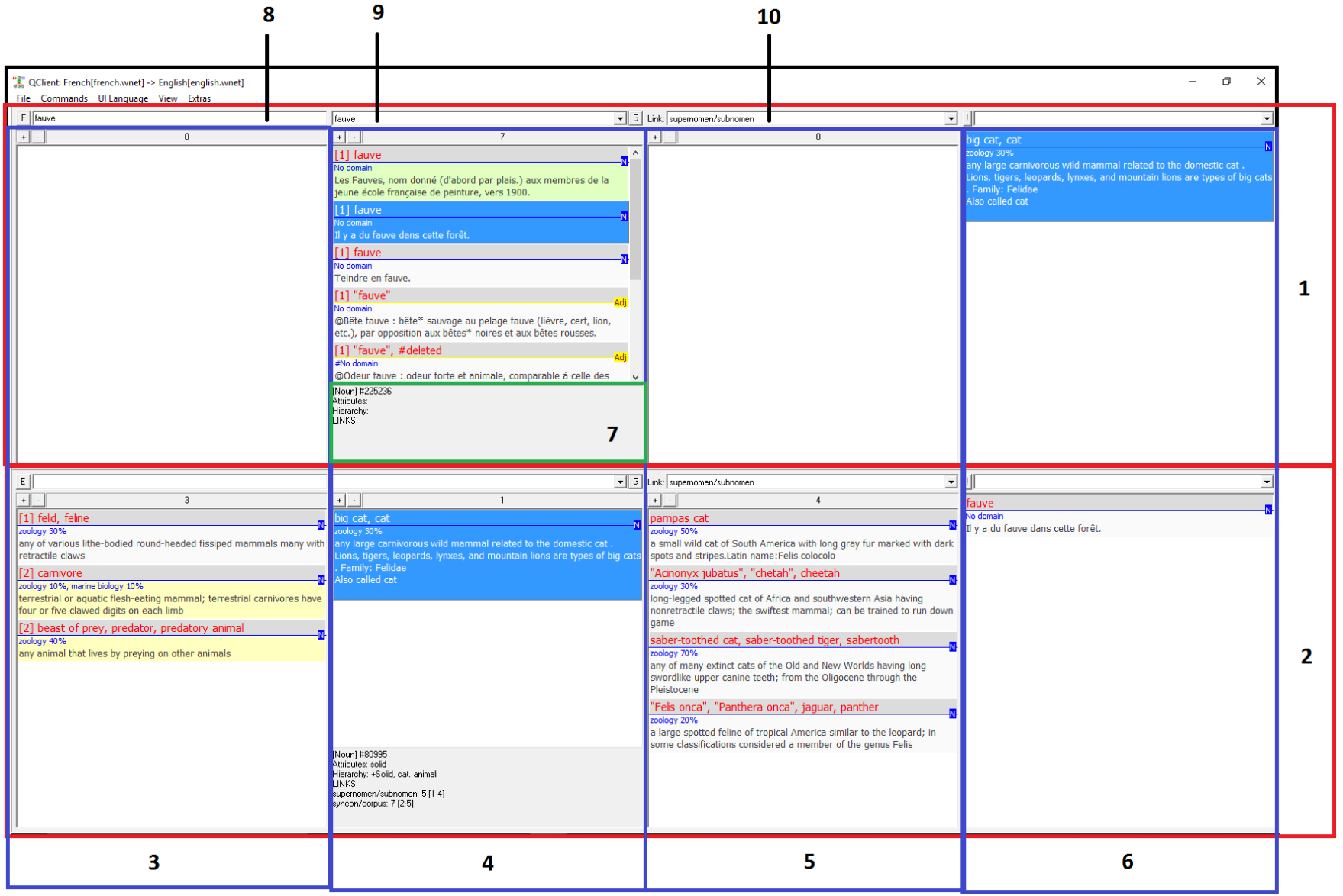


Figure 25 : Capture d'écran de l'interface QClient où le lemme Sensi recherché est « fauve ».

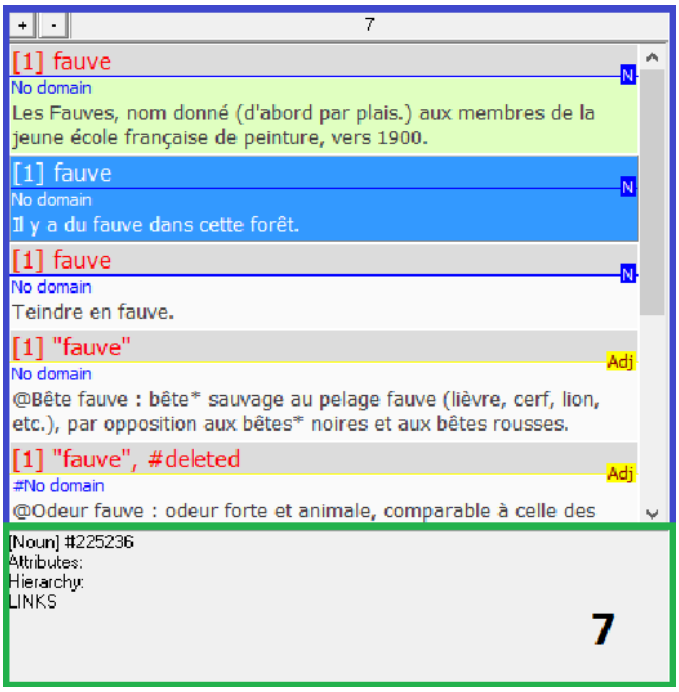


Figure 26 : Capture d'écran des syncons de noms associés au lemme Sensi « fauve » dans le QClient.

La figure 26 indique que le syncon sélectionné est le syncon avec le fond bleu dont l'identifiant est le numéro #2255236 figurant dans l'encadré 7.

Pour un Sensigrafo monolingue, dont les relations sémantiques sont établies entre les syncons, cette interface permet de naviguer d'un syncon à un autre dans la chaîne hiérarchique conceptuelle selon le type de relation sémantiques sélectionné. Dans la figure 25, la chaîne hiérarchique conceptuelle selon la relation sémantique d'hyperonymie/hyponyme (supernomen/subnomen) du syncon sélectionné en français n'est pas encore établie car aucun syncon n'est présent dans les colonnes 3 et 5 contrairement à l'anglais.

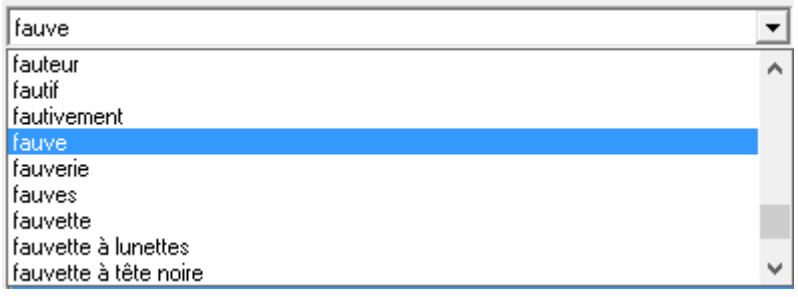


Figure 27 : Capture d'écran du menu déroulant du Qclient portant le numéro 9 dans la figure 25.

Le QClient est l'outil utilisé pour les tâche d'ajustement des fréquences et la mise en correspondance des syncons.

Les trois outils présentés précédemment sont des outils de développement, de l'entreprise Expert System, qui ont pour but de faciliter les tâches manuelles d'enrichissement et de tests. Les tâches effectuées lors de ce stage sont présentées dans la Partie 2 de ce document. L'utilisation pratique de ces outils y est détaillée. La figure 28 représente l'application des tâches effectuées lors du stage et indique également les informations générales relatives à chaque tâche. Par exemple : la tâche de « Test de performance » est la première étape du stage. C'est une tâche de test utilisant le Sensigrafo FR fixe et l'outil Cogito Desambiguator.

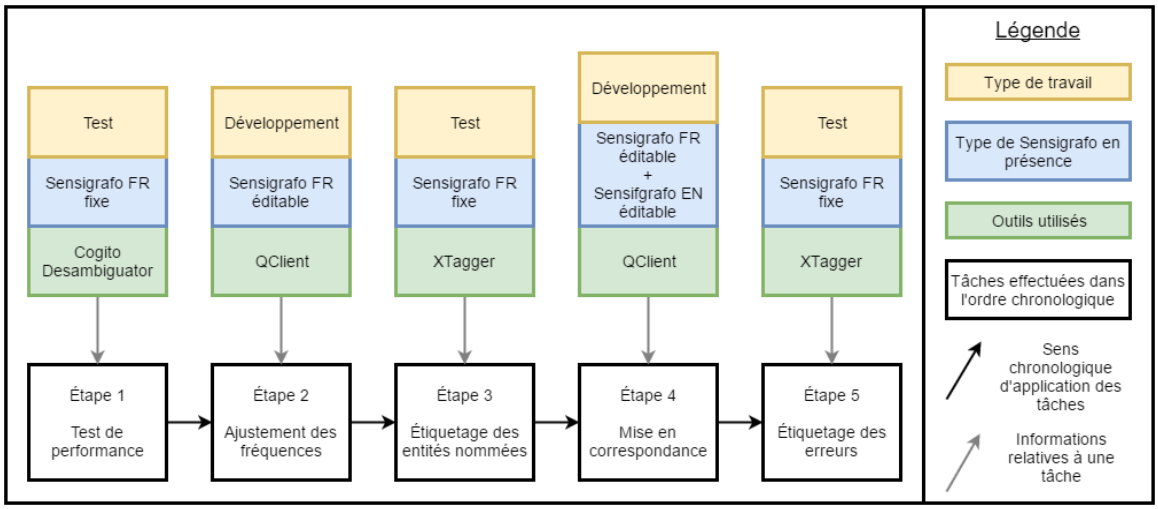


Figure 28 : Représentation schématique de la chronologie des tâches effectuées ainsi que les informations générales relatives à chaque tâche.

Partie 2

-

Tâches effectuées

Le stage s'est déroulé en plusieurs étapes. Il y a eu deux phases de développement et 3 phases de test de performance. Ces étapes s'inscrivent parfois dans le cadre du Sensigrafo FR fixe, parfois dans le Sensigrafo FR éditable et parfois même dans le Sensigrafo EN éditable. Pour comprendre correctement ces phases de travail il est impératif d'avoir compris les stratégies de développement et la structure du Sensigrafo étendu.

Chapitre 1 - Test de performance du Sensigrafo FR à partir du Cogito Desambiguator

Le travail sur le Sensigrafo FR a débuté environ 6 mois avant le début du stage. Ces 6 mois ont permis d'établir une base dictionnairique. À ce moment, la priorité est de connaître la qualité des résultats obtenus ainsi que ses points faibles pour appliquer les développements selon un ordre de priorité. Cette tâche s'inscrit dans le cadre d'observations effectuées sur une version encore peu travaillée du Sensigrafo FR fixe.

1. Présentation du travail

Cette tâche est une phase de test du moteur d'analyse sémantique du français. Une tâche de test a été effectuée à partir du Cogito Desambiguator pour observer les performances de désambiguïsation du moteur d'analyse sémantique et ainsi déterminer quels sont les points à travailler en priorité.

Avant tout, cette tâche nécessite un corpus qui a donc été constitué à partir d'articles de presses issus du web. Les sources sont : 20 minutes, L'Alsace, La Provence et Le Figaro. Ce corpus est composé de 5 articles répartis sur 2 domaines : culture et société. Ce corpus compte 1667 mots dont 807 mots pleins. Chaque article traite d'un sujet différent. Nous l'appellerons « CT1 » pour corpus de test 1.

Pour rédiger nos résultats, nous nous sommes fondé sur un tableau Excel existant, permettant ainsi d'entrer des données pertinentes par phrases analysées.

Colonne du premier tableau Excel

- Phrase (avec le mot cible en rouge)
- Syncon ID (ID complet du mot cible, nouvel identifiant et identifiant standard)
- Lemme (du mot cible)
- Présence d'une erreur de catégorie grammaticale (oui/non)
- Description de l'erreur de catégorie grammaticale
- Présence d'une erreur de groupe grammatical (oui/non)

- Description de l'erreur de groupe grammatical
- Présence d'une erreur sémantique (oui/non)
- Description de l'erreur sémantique
- Présence d'une erreur syntaxique (oui/non)
- Description de l'erreur syntaxique
- Suggestions de corrections
- Titre du document (de la phrase)
- Numéro de la phrase dans le document

Nous avons effectué quelques changements pour plus de lisibilité et de précision. Les erreurs de catégorie grammaticale et de groupe grammaticaux ont été fusionnées pour constituer une colonne "Erreurs grammaticales". Trois colonnes pour donner le type des erreurs ont été ajoutées. Nous avons formalisé les réponses possibles de ces colonnes, ce qui permet par la suite de travailler plus rapidement et de faire des statistiques de manière beaucoup plus précise et rapide.

A. Types d'erreurs de grammaire

- Segmentation Incorrecte

Une segmentation incorrecte signifie qu'un groupe grammatical ne comprend pas les éléments qui devraient l'être. Il existe deux types de segmentation incorrecte :

- Le groupe grammatical est réduit, il ne comprend pas tous les éléments qu'il devrait. La figure 29 est une copie d'écran d'une phrase désambiguïsée dans laquelle se trouve une erreur de segmentation. Le logiciel a désambiguïsé « *un* » « *peu* » et « *partout* » en trois groupes séparés respectivement GN, GV et GV ; il aurait dû regrouper les éléments « *un* », « *peu* » et « *partout* » ensemble sous la forme d'un groupe adverbial.
- Le groupe grammatical est élargi, il comprend des éléments qui devraient être dans des groupes séparés. La figure 30 est un exemple dans lequel se trouve cette erreur. « *On a toujours* » forme un groupe verbal or « *On* » devrait former un groupe nominal et « *a toujours* » un groupe verbal.

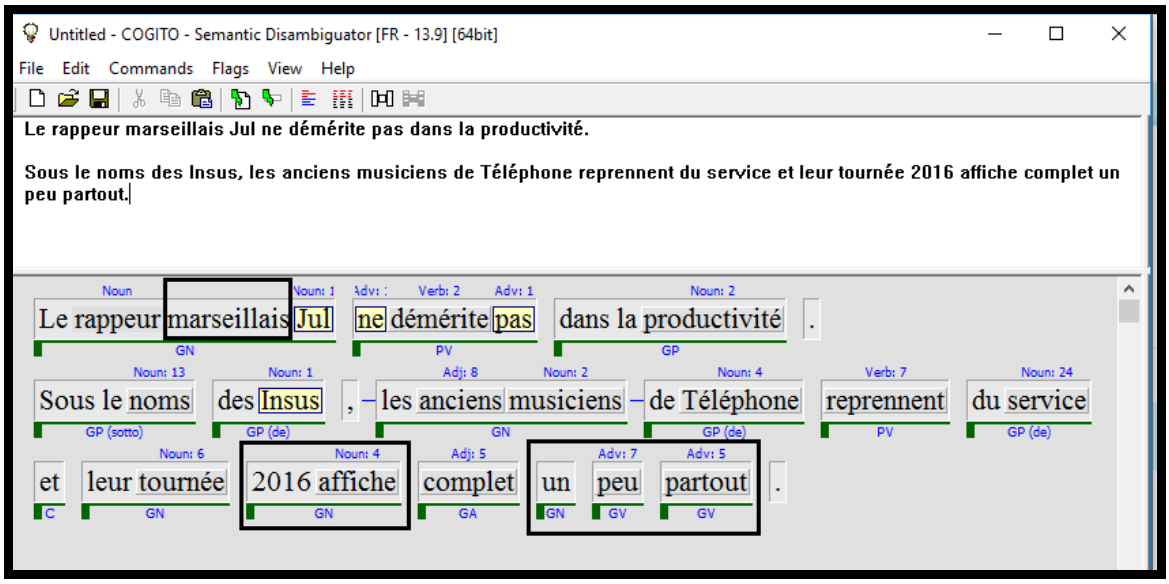


Figure 29 : Capture d'écran d'une phrase désambiguïsée par le Cogito Desambiguator. Les erreurs sont encadrées.

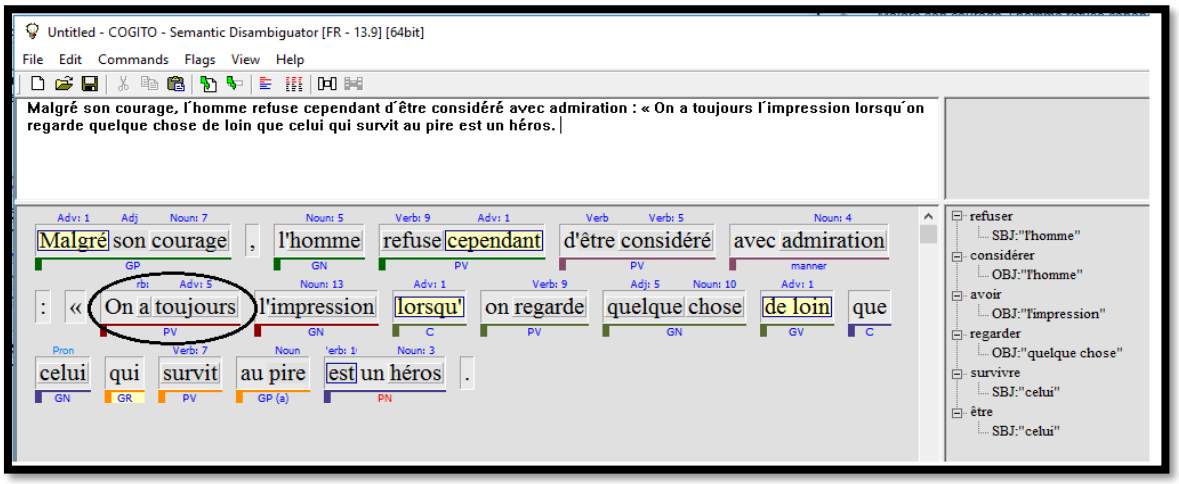


Figure 30 : Capture d'écran d'une phrase désambiguïsée par le Cogito Desambiguator.

- Catégorie grammaticale incorrecte

La catégorie grammaticale détectée est incorrecte. La désambiguïsation présentée dans la figure 29 montre que le mot « *affiche* » a été désambiguïsé en tant que nom au lieu de verbe.

- Catégorie grammaticale absente

La catégorie grammaticale n'est parfois pas reconnue. C'est aussi une erreur de grammaire. Par exemple dans la figure 29, l'adjectif « *marseillais* » n'a pas de catégorie grammaticale assignée.

- Groupe grammatical incorrect

C’est une erreur où le groupe grammatical reconnu est incorrect. Toujours dans la figure 29, le groupe « *2015 affiche* » est reconnu en tant que groupe nominal et non en tant que groupe verbal. L’erreur est liée à l’erreur de catégorie grammaticale.

- Groupe grammatical absent

Le Cogito Desambiguator fourni systématiquement une information sur le groupe grammatical. Il arrive que l’information fournie ne soit pas complète. La figure 31 indique un exemple où un groupe obtient uniquement l’information de « génitif » (traduction de : genitive case). La catégorie du groupe grammatical n’est pas renseignée.

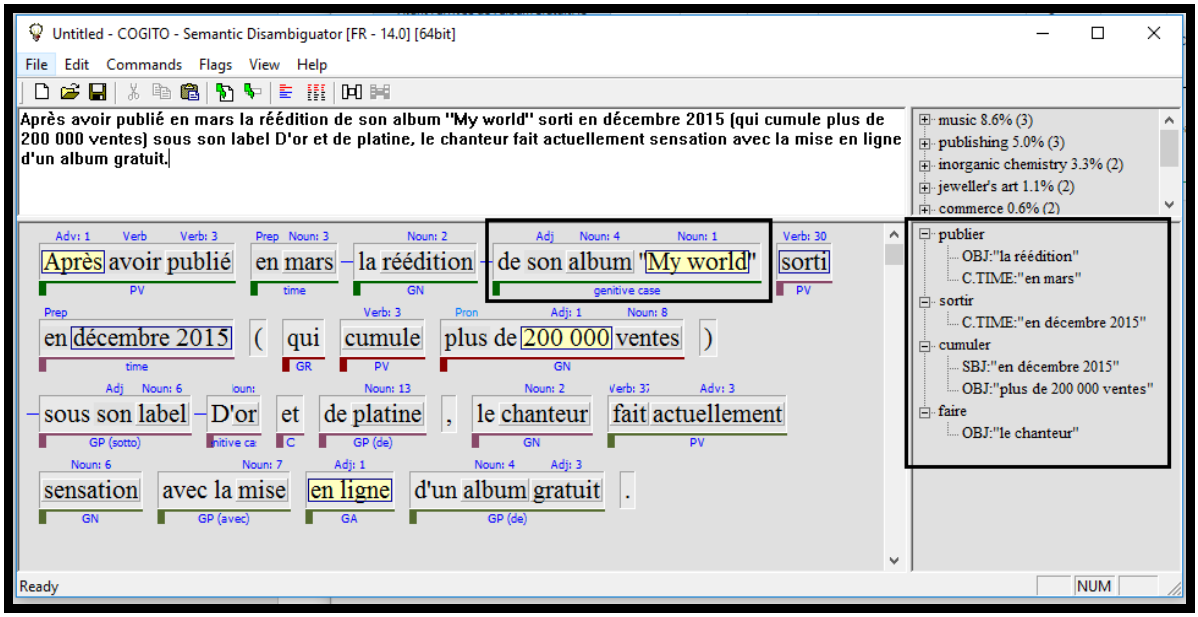


Figure 31 : Capture d’écran d’une phrase désambiguïsée dans le Cogito Desambiguator.

B. Types d’erreurs sémantiques

- Syncon Incorrect - Change

Une erreur de syncon incorrect avec l’appellation « change » apposée signifie qu’un syncon est associé à un mot lors de la désambiguïsation mais que ce n’est pas le syncon correspondant au sens en contexte. De plus le syncon correspondant existe dans le Sensigrafo FR. Par exemple, prenons la phrase :

> « Après avoir publié en mars la réédition de son album "My world" sorti en décembre 2015 (qui cumule plus de 200 000 ventes) sous son label D'or et de platine, le chanteur fait actuellement sensation avec la mise en ligne d'un album gratuit. »

Dans ce cas le mot « *label* » est désambiguïsé comme correspondant au syncon suivant :

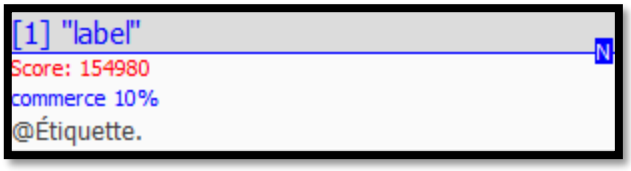


Figure 32 : Capture d'écran d'un syncon de « label » présent dans le Sensigrafo FR fixe.

Cependant, ce concept ne correspond pas à l'acception dans le contexte de la phrase. Le concept correspondant à l'acceptation dans le contexte de la phrase serait le syncon suivant :

Figure 33 : Capture d'écran d'un syncon de « label » présent dans le Sensigrafo FR fixe.

- Syncon Incorrect - Ajout

  Une erreur de syncon incorrect avec l'appellation « ajout » apposée signifie qu'un syncon est associé à un mot lors de la désambiguïsation mais que ce n'est pas le syncon correspondant au sens en contexte. L'erreur est la même que celle présentée précédemment dans le point « ***Types d'erreurs sémantiques*** » point « Syncon Incorrect – Change » seulement ici le syncon correspondant au concept en contexte du mot n'existe pas dans le Sensigrafo FR.

- Indéfini

  Une erreur de sémantique « indéfini » signifie qu'aucun syncon n'a été associé au mot traité. Dans le Cogito Desambiguator les mots associés à un syncon sont encadrés or dans la figure 34 les mots « *rappeur* » et « *marseillais* » n'ont pas de cadre comme les mots « *productivité* » ou « *noms* » par exemple. Lors de ce test aucuns syncons n'a été associé à ces mots.

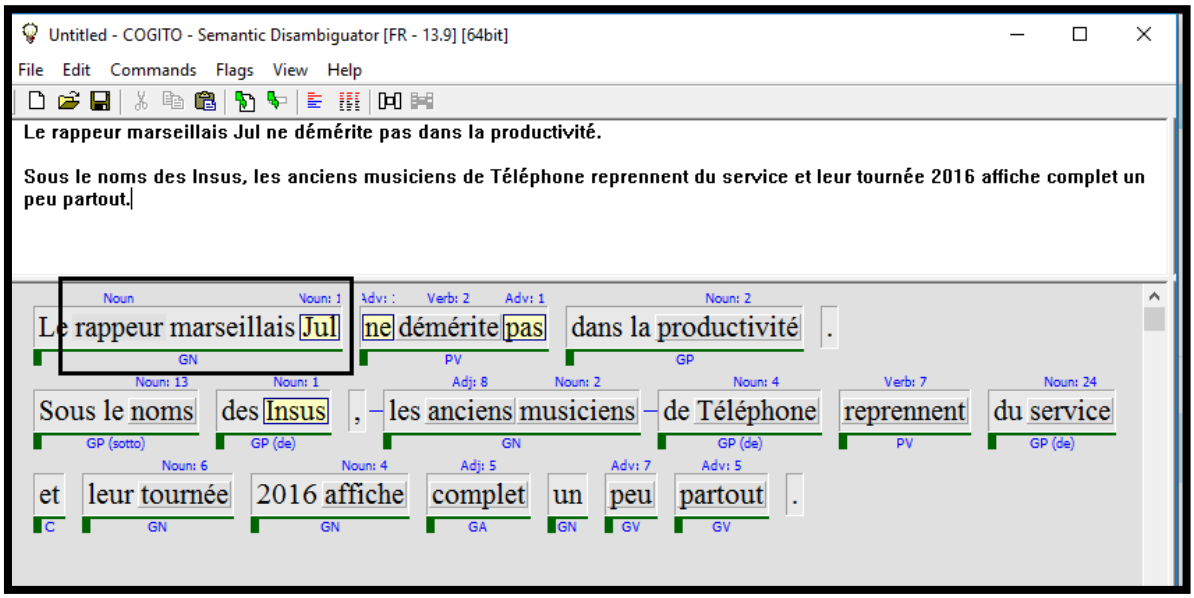


Figure 34 : Capture d'écran d'une phrase désambiguïsée via le Cogito Desambiguator.

- Autres

Il est possible que des erreurs rencontrées ne correspondent à aucunes des erreurs présentées ci-dessus. Elles ont été notées comme « Autres » car leurs nombres étaient trop faibles pour que de nouveaux types d'erreurs sémantiques soient créés. Ces erreurs peuvent provenir de différents phénomènes.

- Un syncon peut avoir un glossa imprécis qui mériterait d'être plus détaillé. La figure 35 montre le glossa du syncon associé au terme « *YouTube* ». La description du syncon pour décrire le concept d'un site web d'hébergement de vidéos se trouve dans l'encadré noir.

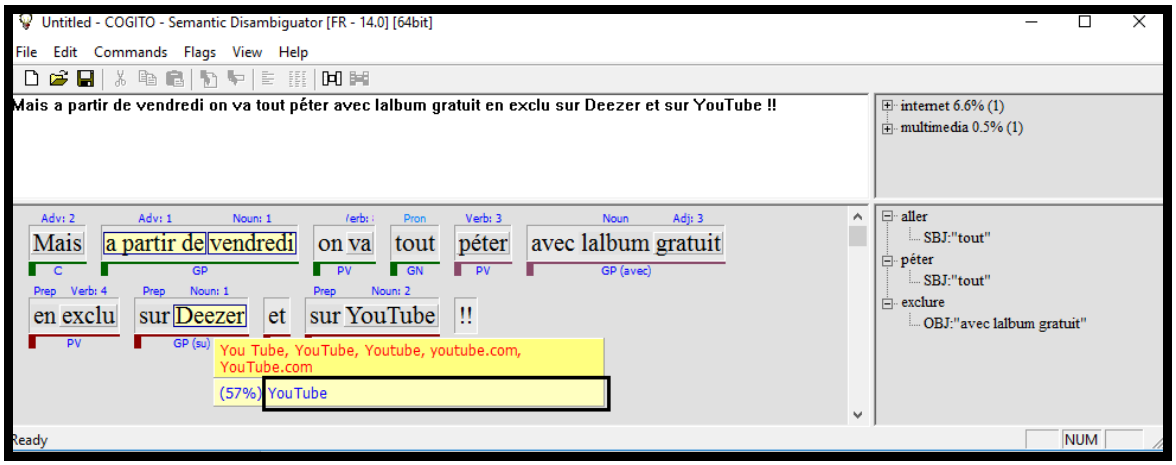


Figure 35 : Capture d'écran d'une phrase désambiguïsé via le Cogito Desambiguator.

- L'erreur peut être liée à une autre. Si un mot est désambiguïsé comme étant un nom au lieu d'un verbe c'est une première erreur grammaticale « Catégorie Grammaticale Incorrecte ». Dans ce cas, la désambiguïsation sémantique sera fausse car les syncons ne peuvent avoir qu'une seule catégorie grammaticale, le syncon associé sera obligatoirement un syncon ayant la catégorie grammaticale de « nom » au lieu de « verbe ». C'est une erreur de sémantique du type « Syncon Incorrect » due à l'erreur de grammaire « Catégorie grammaticale Incorrecte ».

C. Types d'erreurs de syntaxe

- Incorrect

Une erreur de syntaxe de type « Incorrect » se présente quand la relation syntaxique détectée est incorrecte c'est-à-dire que le sujet ou l'objet associé à un verbe n'est pas celui qui devrait l'être. Dans la figure 31 se trouve la phrase :

> « Après avoir publié en mars la réédition de son album "My world" sorti en décembre 2015 (qui cumule plus de 200 000 ventes) sous son label D'or et de platine, le chanteur fait actuellement sensation avec la mise en ligne d'un album gratuit. »

Ici le verbe « *cumuler* » est noté comme ayant pour sujet « *en décembre 2015* ». Pourtant cette information est fausse, « *réédition* » est le véritable sujet de « *cumuler* ».

- Indéfini

Une relation syntaxique absente représente une erreur que nous appelons « Indéfini ». Dans la figure 36, le sujet du groupe « *sorti* » est le groupe « *Son dernier album Shadow* ». Cependant, cette relation n'a pas été reconnue par le logiciel. C'est un exemple d'erreur de syntaxe du type « Indéfini ».

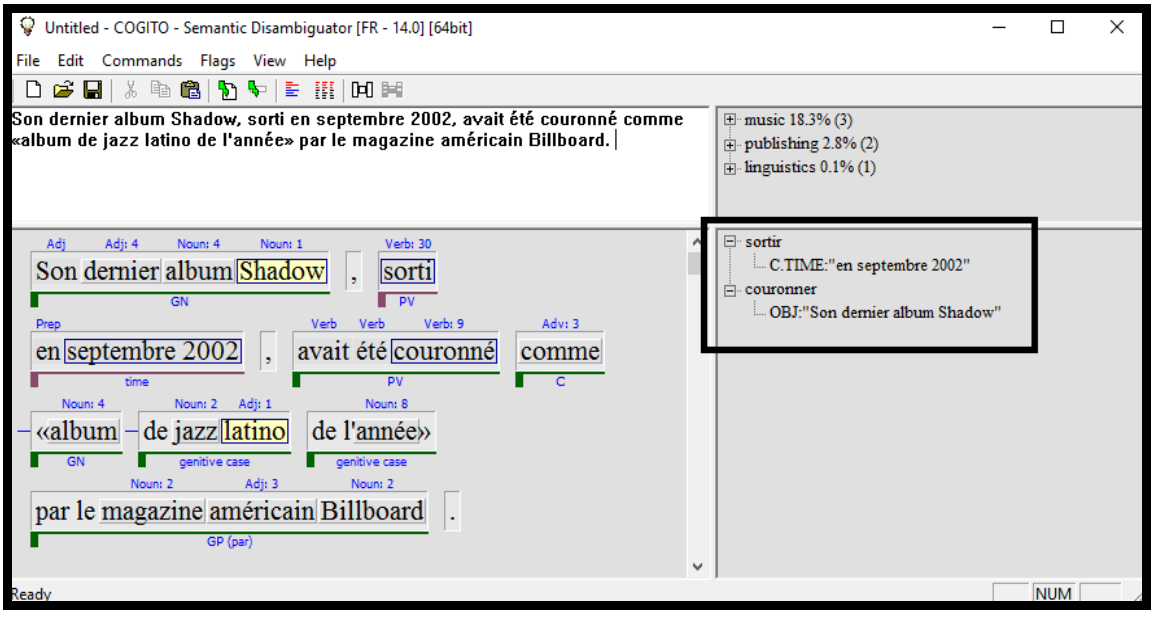


Figure 36 : Capture d'écran d'une phrase désambiguïsée via le Cogito Desambiguator.

Au début nous avons noté dans le document chaque verbe, nom, adjectif et adverbe même dans le cadre d'énoncés correctement désambiguïsés par le système. Puis nous avons décidé de noter uniquement les éléments porteurs d'une erreur. L'objectif de ce travail était d'observer les erreurs pour prioriser les développements à effectuer par la suite. Les éléments ne comportant pas d'erreur n'apportent aucune information supplémentaire dans ce but mais augmente le nombre de ligne de rédaction dans le document et prend du temps à rédiger. Donc nous avons gardé uniquement les éléments portant au minimum une erreur. En particulier les erreurs qui sont la cause d'autres erreurs. Nous les appelons les "erreurs principales". Par exemple si un mot est reconnu comme un adjectif au lieu d'un nom c'est une erreur grammaticale. Et cette erreur grammaticale implique systématiquement une erreur de sémantique. Ce changement dans notre analyse nous a permis de clarifier notre fichier et d'avoir moins de redondances. Le document est plus facile à lire ainsi.

Nous avons ajouté une colonne " gravité " des erreurs ce qui nous permet de donner des priorités aux tâches de correction. Cet indice de gravité est fondé sur une échelle de 1 (le plus sévère) à 5 (le moins sévère). Cette gravité est subjective car il n'y a pas de référentiel précis qui puisse être suivi. Le critère principal était d'imaginer la réaction d'un client potentiel face aux erreurs.

Pour avoir une idée nous nous sommes mis d'accord sur quelques critères :

- Un nom propre reconnu comme un nom commun aura une gravité de 5 car cette erreur n'implique pas d'autres erreurs.
- Une erreur grammaticale sur la catégorie aura une gravité de 1 car elle implique systématiquement une erreur de sémantique.
- Une erreur grammaticale sur le groupe aura une gravité de 1 ou 2 car elle implique la plupart du temps une erreur sémantique et/ou une erreur syntaxique.
- Une erreur sémantique qui n'est pas due à une autre erreur sera évaluée selon l'écart entre la définition reconnue et la définition réelle. Ainsi, dans le cas où nous sommes en présence du mot « *fils* » en tant que nom portant le concept « *Être humain du sexe masculin, considéré par rapport à son père et à sa mère ou à l'un des deux seulement* » est reconnu comme un « *boyau de chat* » (catgut en anglais), la gravité sera de 1 car les concepts sont très loin l'un de l'autre. Mais si nous sommes en présence du mot « *reprendre* » en tant que verbe portant le sens de « *commencer à nouveau* » (to start again en anglais) est qu'il est reconnu avec le

sens « *faire une tentative* » (to make an attempt en anglais) la gravité sera de 3 car les concepts sont relativement proches.

- Une erreur syntaxique qui n'est pas dû à une autre erreur serait de 1 car le système de désambiguïsation de relations syntaxiques est la cause de l'erreur.

Colonnes du tableau Excel retravaillé

1. Phrase (avec le mot cible en rouge)
2. ID du syncon reconnu
3. Lemme du mot cible
4. Présence d'une erreur grammaticale (oui/non)
5. Description de l'erreur grammaticale
6. Type de l'erreur grammaticale
7. Présence d'une erreur sémantique (oui/non)
8. Description de l'erreur sémantique
9. Type de l'erreur sémantique
10. Présence d'une erreur syntaxique(oui/non)
11. Description de l'erreur syntaxique
12. Type de l'erreur syntaxique
13. Suggestions de corrections
14. Titre du document source
15. Numéro de la phrase dans le document source
16. Gravité de l'erreur analysée

2. Résultats

Grâce à notre analyse nous avons identifié les principales faiblesses de la version française des désambiguïsations effectuées par le moteur d'analyse sémantique. Cette tâche a été effectuée sur seulement 5 articles et 345 mots pleins comprennent au moins une erreur sur les 807 mots analysés.

Avant tout il est nécessaire de faire la distinction entre le nombre d'erreurs et le nombre d'éléments comportant au moins une erreur. Certains éléments comportent une seule erreur et d'autres en comporteront 2 ou 3. C'est à ce titre que le nombre d'erreurs est plus important que le nombre d'éléments comportant au moins une erreur. Il y a 807 mots analysés. 345 d'en eux comportent au moins une erreur et 462 erreurs sont recensées. Il n'est

pas possible qu'un élément comporte plus de 3 erreurs car ce sont les erreurs générales (erreurs de grammaire, erreurs de sémantique, erreurs de syntaxe) dont nous parlons ici.

A. Résultats d'ensemble

Nous allons présenter les statistiques de nos résultats. Nos résultats ne prennent pas systématiquement en compte le fait que certaines erreurs sont dues à d'autres. Ces pourcentages sont donc des estimations.

Le tableau 4 comporte les totaux généraux de ce test. Le nombre de mots pleins et de mots grammaticaux représente le total de tous les éléments lexicaux présents dans le corpus. Le nombre de mots pleins représente le total de mot pour lesquels une désambiguïsation est effectuée. Le nombre d'erreurs représente le nombre de mots pleins comportant au minimum une erreur de désambiguïsation. Le pourcentage 1 a pour base 100 le total de mots pleins et de mots grammaticaux soit 1667. Le pourcentage 1 représente donc la répartition des mots pleins et des erreurs par rapport au nombre total d'éléments lexicaux du corpus. Ainsi, 48.41% du corpus dans son intégralité sont des mots pleins et 20.60% de ce corpus comporte au moins une erreur. Le pourcentage 2 a pour base 100 le total de mots pleins du corpus soit 807. Le pourcentage 2 représente donc la répartition des erreurs sur le total des mots analysés. 42.70% des mots analysés comportent au moins une erreur de désambiguïsation.

Aperçu global	Nombre de mots pleins et mots grammaticaux par articles	Nombre de mots pleins	Nombre de mots pleins comportant au minimum une erreurs
Total	1667	807	345
Pourcentage 1	100%	48.41%	20,60%
Pourcentage 2		100%	42,70%

Tableau 4 : Tableau des répartitions des mots pleins et des erreurs par rapport au corpus total et la répartition des erreurs par rapport aux mots pleins.

Le tableau 5 nous indique les pourcentages des erreurs générales. Le pourcentage 3 représente le pourcentage d'erreurs par rapport au nombre d'erreurs détectées. Le tableau 6 indique la répartition des combinaisons d'erreurs rencontrées par rapport au nombre de mots pleins (pourcentage 4) et par rapport au nombre de mots comportant au moins une erreur (pourcentage 5).

Aperçu global	Nombre	Pourcentage 3
Total d'erreurs	462	100%
Erreurs de grammaire	129	27,92%
Erreurs de sémantique	247	53,46%
Erreurs de syntaxe	86	18,61%

Tableau 5 : Tableau de répartition des erreurs par rapport au nombre d'erreurs comptabilisées.

Aperçu global	Nombre	Pourcentage 4	Pourcentage 5
Total de mots pleins	807	100%	
Total mots pleins comportant une erreur au moins	345	42.75%	100%
Erreurs de grammaire	48	5.94%	13.91%
Erreurs de sémantique	159	19.7%	46.08%
Erreurs de syntaxe	35	4.33%	10.14%
Erreurs de grammaire + Erreurs de sémantique	52	6.44%	15.07%
Erreurs de grammaire + Erreurs de syntaxe	15	1.8%	4.34%
Erreurs de syntaxe + Erreurs de sémantique	22	2.72%	6.37%
Erreurs de grammaire + Erreurs de sémantique + Erreurs de syntaxe	14	1.73%	4.05%

Tableau 6 : Tableau des pourcentages d'erreurs selon les types généraux.

B. Résultats par types d'erreurs et gravités

Les tableaux 7, 8 et 9 se focalisent respectivement sur les types d'erreurs grammaticales, sémantiques et syntaxiques. Les résultats indiquent qu'un grand nombre d'erreurs proviennent d'une erreur de sémantique de type « Change ». Ce qui veut dire que le Sensigrafo FR est plutôt complet car les définitions existent. Mais c'est au niveau de la désambiguïsation que des problèmes apparaissent. Ces résultats montrent aussi que le système est relativement robuste car nous avons presque dans tous les cas un étiquetage qui est effectué. Seulement 9,31 % n'ont pas de définition associée. De plus cette erreur peut être provoquée par d'autres. Il arrive qu'un adverbe soit reconnu comme un pronom. Or les pronoms n'apparaissent pas dans le Sensigrafo FR. Donc il n'y a pas de désambiguïsation

sémantique effectuée sur cet élément. L'élément n'aura pas de syncon associé et sera noté comme une erreur de sémantique « indéfini ».

Erreurs de grammaire par type	Nombre	Pourcentage 6
Total	129	100%
Segmentation incorrecte	15	11,62
Catégorie grammaticale absente	2	1,55
Catégorie grammaticale incorrecte	63	48,83
Groupe grammatical absent	1	0,77
Groupe grammatical incorrect	27	20,93
Catégorie grammaticale incorrecte + Groupe grammatical incorrect	14	10,85
Catégorie grammaticale incorrecte + Segmentation incorrecte	2	1,55
Catégorie grammaticale incorrecte + Segmentation incorrecte+ Groupe grammatical incorrect	1	0,77
Groupe grammatical incorrect + Catégorie grammaticale absente	1	0,77
Groupe grammatical absent + Catégorie grammaticale incorrecte	1	0,77
Segmentation incorrecte+ Catégorie grammaticale incorrecte	1	0,77

Tableau 7 : Tableau présentant la répartition des types spécifiques d'erreurs grammaticales par rapport au total d'erreurs grammaticales.

Erreurs de sémantique par type	Nombre	Pourcentage 7
Total	247	100%
type « Ajout »	38	15,38%
type « Change »	158	63,96%
type « Autres »	28	11,33%
type « Indéfini »	23	9,31%

Tableau 8 : Tableau présentant la répartition des types spécifiques d'erreurs sémantiques par rapport au total d'erreurs sémantiques.

Erreurs de syntaxe par type	Nombre	Pourcentage 8
Total	86	100%
type « Incorrect »	46	53,48%
type « Indéfini »	35	40,69%
type « Autres »	5	5,81%

Tableau 9 : Tableau présentant la répartition des types spécifiques d'erreurs syntaxiques par rapport au total d'erreurs syntaxiques.

Le tableau 10 donne les pourcentages d'indice de gravité et montre que les erreurs les plus graves représentent 26.66% des erreurs rencontrés contre 18,23% pour les moins graves. Cependant, le graphique de la figure 37 montre que la répartition n'est pas progressive de manière linéaire.

Indice de gravité	Nombre	Pourcentage 9
Total	345	100%
Gravité 1	92	26,66%
Gravité 2	63	18,26%
Gravité 3	84	24,34%
Gravité 4	47	13,62%
Gravité 5	59	17,10%

Tableau 10 : Ce tableau présente le pourcentage de mots pleins comportant au moins une erreur par indice de gravité.

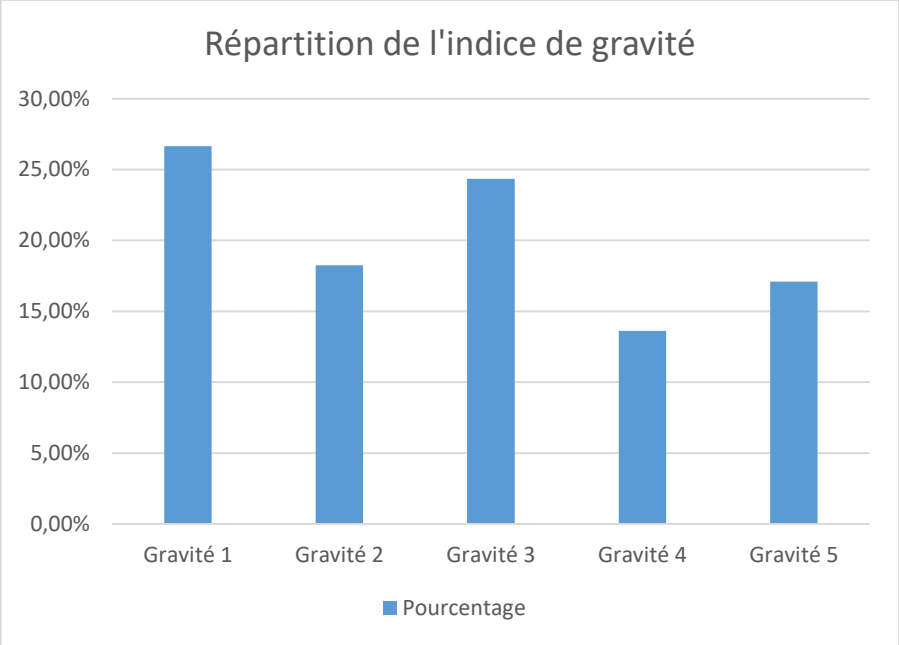


Figure 37 : Graphique de la répartition de l'indice de gravité entre les différents niveaux de gravité.

Les tableaux 11 à 14 détaillent les pourcentages de gravité par types d'erreurs générales. Le tableau 12 montre que 37,20% des erreurs de grammaire ont une gravité de 1. Le tableau 13 montre que 20% des erreurs de sémantique ont une gravité de 1 contre 54,6% pour les erreurs de syntaxe dans le tableau 14.

Erreurs générales	Gravité 1	Gravité 2	Gravité 3	Gravité 4	Gravité 5	Totaux par types d'erreurs générale
Erreurs de grammaire	48	16	12	11	42	129
	10,38%	3,46%	2,59%	2,38%	9,09%	27,92%
Erreurs de sémantique	49	58	70	38	32	247
	10,60%	12,55%	15,15%	8,22%	6,92%	53,46%
Erreurs de syntaxe	47	13	16	5	5	86
	10,17%	2,81%	3,46%	1,08%	1,08%	18,61%
Totaux par indice de gravité	144	87	98	54	79	462
	31,16%	18,83%	21,21%	11,68%	17,09%	100%

Tableau 11 : Tableau dans lequel se trouve la répartition des gravités par types d'erreurs générales par rapport au total d'erreurs.

Erreur de grammaire	Nombre	Pourcentage 10
Total	129	100%
Gravité 1	48	37,20%
Gravité 2	16	12,40%
Gravité 3	12	9,30%
Gravité 4	11	8,52%
Gravité 5	42	32,55%

Tableau 12 : Tableau indiquant le pourcentage d'erreurs de grammaire selon l'indice de gravité.

Erreurs de sémantique	Nombre	Pourcentage 11
Total	247	100%
Gravité 1	49	20%
Gravité 2	58	23,48%
Gravité 3	70	28,34%
Gravité 4	38	15,38%
Gravité 5	32	13%

Tableau 13 : Tableau indiquant le pourcentage d'erreurs de sémantique selon l'indice de gravité.

Erreurs de syntaxique	Nombre	Pourcentage 12
Total	86	100%
Gravité 1	47	54,6%
Gravité 2	13	15,1%
Gravité 3	16	18,6%
Gravité 4	5	5,8%
Gravité 5	5	5,8%

Tableau 14 : Tableau indiquant le pourcentage d'erreurs de syntaxe selon l'indice de gravité.

Le temps de travail passé sur l'ensemble de cette tâche est de 52h.

3. Problèmes rencontrés et réflexions

A. Généralités

La version du Sensigrafo FR fixe qui était utilisée dans le Cogito Desambiguator avait la particularité d'avoir les glossas des syncons écrits en anglais. Ces glossas comprenaient également un grand nombre d'erreurs de grammaire (comme le montre la figure 38). Parfois des prépositions en italien s'y trouvaient insérées. Dans certains cas les glossas étaient entièrement écrits en italien. Ce problème de traduction a affecté nos capacités à analyser les informations présentées. Il était parfois difficile de comprendre le sens exprimé.

Des aspects de l'ergonomie du logiciel pourraient être retravaillés pour rendre la manipulation de cet outil plus agréable en particulier le choix de la taille des fenêtres qui pour l'instant sont verrouillées.

B. Le génitif

La règle du génitif est très sensible : en français, la possession peut être exprimée par la préposition « de », comme dans exemple « Le vélo de mon frère ». Mais « de » est une préposition hautement grammaticalisée. Elle peut exprimer une grande variété de sens comme la composition, l'origine, la fonction… Par conséquent, la simple présence de la préposition « de » est insuffisante pour qualifier un groupe prépositionnel comme étant un Génitif. C'est pourtant une erreur récurrente.

C. Le participe passé utilisé en tant qu'adjectif

Les participes passés avec une fonction adjectivale sont considérés comme des prédicats verbaux au lieu de groupes adjectivaux. Ce phénomène mène à différentes erreurs

syntaxiques puisque le système semble essayer de construire des relations syntaxiques pour ces participes avec comme conséquence une reconnaissance erronée d'un groupe nominal en tant que sujet ou objet. Le plus souvent c'est un groupe proche.

D. Syntaxe

Le système a des difficultés à effectuer des analyses systématiques sur des phrases complexes, en particulier quand plusieurs propositions sont présentes ou quand le sujet est après le verbe. Les structures emphatiques (du type « c'est […] qui […] ») sont aussi traitées avec une certaine difficulté.

Aussi les formes narratives (du type « raconte-t-il », « poursuit-t-il » …) font que le système ne note pas d'importantes relations syntaxiques, car le sujet est à une position relativement inhabituelle dans cette structure.

Plusieurs modèles récurrents de conjugaison en français ne sont pas identifiés. Si le système était capable d'identifier des modèles comme « venir de + [v.inf] », « faire + [v.inf] » ou « aller + [v .inf] » et de les traiter en tant que prédicat verbal simple, l'analyse syntaxique devrait être beaucoup plus juste.

E. Sémantique

Les syncons reconnus pour la plupart des mots sont faux ou ne correspondent pas au contexte spécifique. Les erreurs de sémantiques représentent 53.46% des erreurs notées. De plus, une erreur de catégorie grammaticale induit systématiquement une erreur de sémantique puisque les syncons ne peuvent avoir deux catégories grammaticales.

Un certain nombre de syncons doit être ajouté dans le Sensigrafo FR, en particulier pour prendre en compte le sens le plus récent de certains lemmes. Par exemple : "chaîne » peut faire référence à une chaîne YouTube, pourtant ce sens n'existe pas encore dans le Sensigrafo FR. Plusieurs locutions et collocations ne sont pas reconnues et pourraient bénéficier de la création de syncons spécifiques comme « à partir de », « en ligne », « bande originale » « grand public », « à son actif », « faire le point », « en cours » ou encore « quitte à ».

Certaines erreurs pourraient être évitées par une meilleure reconnaissance des entités nommées. En effet, quand le nom, ou une partie du nom, d'une entité non-identifiée est un homonyme d'un nom commun, cela pourrait mener à une erreur de catégorie grammaticale par exemple.

F. Ressources dictionnairiques

Il est indispensable de préciser qu'à ce moment du stage, nous disposions uniquement de la première version du Sensigrafo FR. Les glossas des syncons étaient rédigés dans un anglais souvent incorrect (voir figure 38). Parfois des mots d'italien et parfois des glossas entièrement en italien étaient observés (voir figure 39). La version fournie par la suite présente les définitions en français correct.

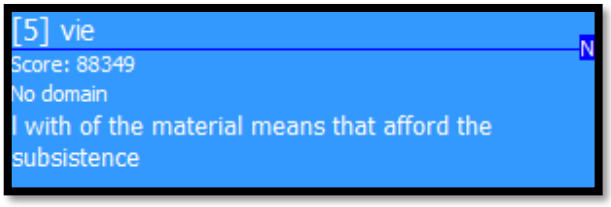


Figure 38 : Capture d'écran d'un syncon associé au lemme Sensi « vie » dans la première version du Sensigrafo FR fixe utilisée dans le Cogito Desambiguator.

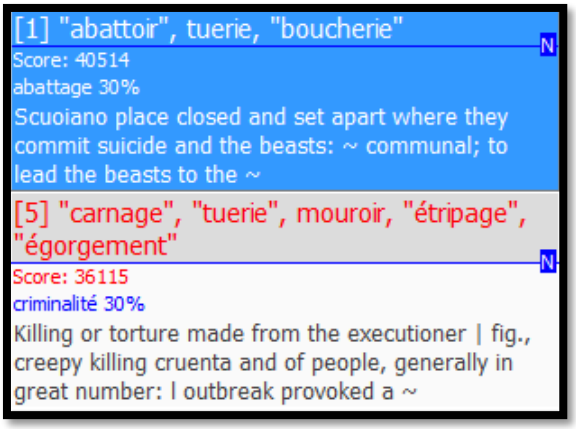


Figure 39 : Capture d'écran de syncons associés au lemme Sensi « tuerie » dans la première du Sensigrafo FR fixe utilisée dans le Cogito Desambiguator.

Il faut dire que cette traduction a été assez handicapante et a limitée nos capacités à évaluer la validité de chaque définition. Quelques-unes étaient difficilement compréhensibles, d'autres étaient ambiguës ne serait-ce que par la langue mais aussi par rapport à d'autres syncons. Elles n'étaient parfois pas assez explicites pour prendre une décision sur quel syncon définissait le mieux le mot en contexte parfois.

Les ressources dictionnairiques utilisées comprennent plusieurs centaines de milliers de sens. De plus ces ressources sont faites pour une utilisation de référence et non une utilisation informatique. Les structurations ne sont pas standardisées de manière stricte ce qui a conduit à des erreurs d'extraction récurrentes.

Ce sont des ressources particulièrement importantes en terme de quantité où se trouve un très grand nombre de termes de spécialités dans un large panel de domaines mais

aussi de nombreux termes de différentes époques. L'exemple du « hérisson » illustre parfaitement ce point avec ses 11 acceptions qui se trouve dans le tableau 1 page 19. Il n'est pas possible de quantifier les syncons de spécialisation ou historique. Cependant, aujourd'hui le nombre de lemmes Sensi est de plus de 247 000 pour environ 261 000 syncons.

Cette tâche a été effectuée en tout début de stage. Nous partions du principe que ce réseau serait utilisé pour traiter tous les textes en français de toutes les époques et domaines confondus. En suivant ce résonnement le choix des ressources dictionnairiques larges est pertinent. Mais cet objectif nécessite énormément de travail.

Cependant, si l'objectif n'est pas de traiter tous les textes peu importe le domaine et l'époque, le changement de ressources dictionnairiques peut être intéressant. Il est cependant nécessaire de calculer ce qui est le plus bénéfique entre traiter une très grande quantité d'éléments ou constituer une nouvelle base dictionnairique à partir de nouvelles ressources pour ainsi traiter un nombre moins important d'éléments.

Un autre point saillant attire notre attention : les erreurs de catégorie grammaticale. Le nombre d'erreurs de catégorie grammaticale peut être réduit grâce à un travail effectué sur les règles de désambiguïsations grammaticales. N'ayant pas eu accès à cette partie du moteur d'analyse sémantique nous ne détaillerons pas ce point.

Chapitre 2 - Ajustement des fréquences via le QClient

L'ajustement des fréquences est une phase de développement qui s'applique sur le Sensigrafo FR éditable. Le but est d'effectuer une première étape de nettoyage des erreurs issues de l'extraction des ressources dictionnairiques ainsi que d'ajouter des informations utilisées lors de la désambiguïsation sémantique. Cette tâche s'inscrit dans le cadre du Sensigrafo FR éditable.

1. Présentation du travail

Cette tâche est une tâche de développement du réseau Sensigrafo FR. La désambiguïsation sémantique consiste à sélectionner le concept correct associé à un terme dans le contexte. L'une des informations présentées dans la première partie de ce document est la fréquence. La fréquence est une échelle de 1 à 100 qu'on applique aux différentes acceptions d'un terme. Une seule acception par lemme peut avoir une fréquence de 1. Cette

fréquence de 1 signifie que c'est l'acception la plus courante dans la langue française standard. La suite est échelonnée sur une grille flexible.

À ce moment, une liste de noms et de verbes (voir annexe 6) automatiquement extraite m'a été confiée. Les termes les plus courants avaient été traités automatiquement.

Pour ce travail, un terme est entré dans la barre de recherche du QClient (encadré 1 de la figure 40). Une liste de syncons ayant ce lemme Sensi associé sont affichés dans la colonne du centre. Dans la figure 40 ces syncons sont affichés dans les encadrés 2, 3 et 4. Les syncons dont le fond est vert ont été mis en correspondance avec un syncon d'une langue source. Ce sont ces syncons qu'il faut traiter. Nous rappelons que cette tâche s'effectue toujours entre les acceptions, d'un terme, ayant la même catégorie grammaticale. Cette tâche s'effectue en deux étapes. Premièrement la vérification des syncons puis la distribution des fréquences.

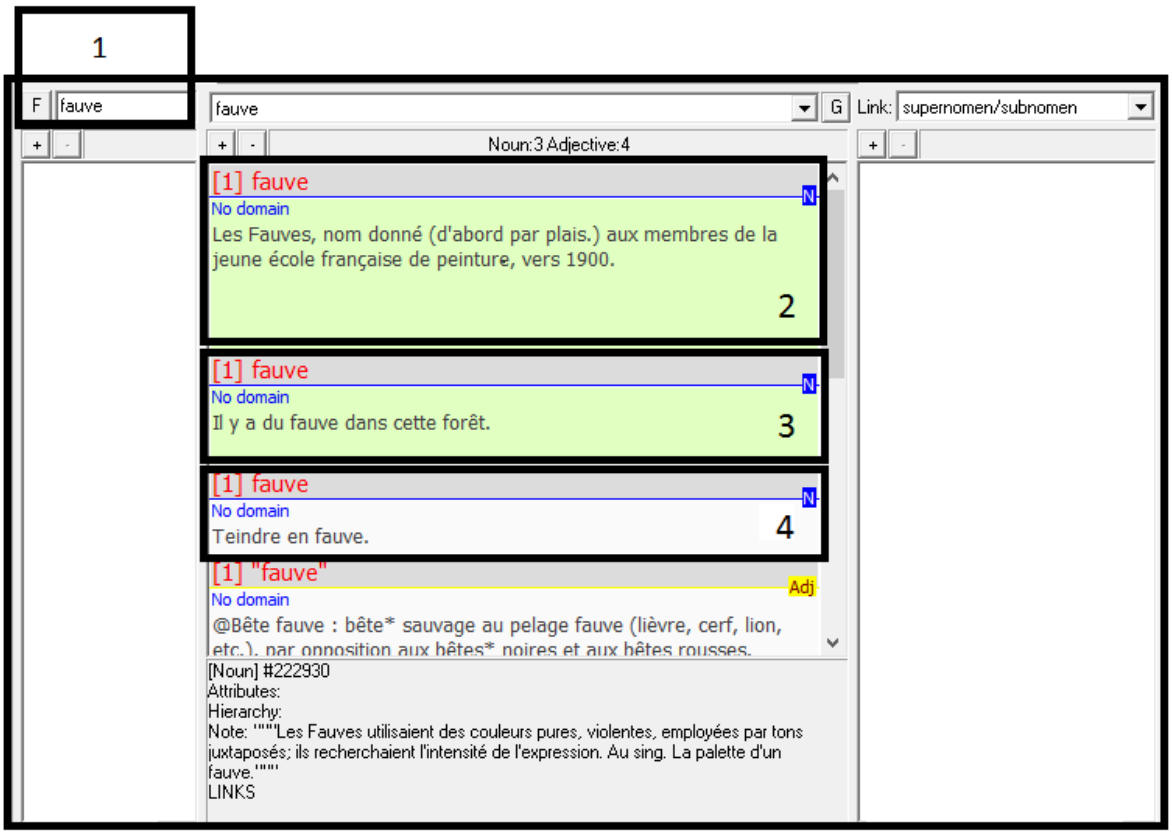


Figure 40 : Capture d'écran de l'interface QClient où le lemme recherché est « fauve ».

A. Vérification des syncons

L'extraction des ressources dictionnairiques a été exécutée automatiquement. Parfois des définitions ont été extraites alors qu'elles ne correspondaient qu'à un exemple comme dans la figure 40 dans l'encadré 3. Le glossa correspond en réalité à une phrase d'exemple. Une personne francophone maternelle comprend que le terme « fauve », dans ce

contexte, exprime le concept d'un type de mammifère carnivore ayant le pelage couleur fauve comme les lions ou les lynx.

D'autres fois des définitions n'ont pas été extraites alors qu'elles sont nécessaires. Dans ces cas-là elles sont notées pour être rajoutées par la suite. Ce phénomène est dû aux problèmes relatifs à l'extraction des ressources dictionnairiques. N'ayant pas eu accès à ces informations nous ne détaillons pas plus le sujet.

Nous avons aussi noté des erreurs de catégories grammaticales, par exemple : une définition pour un adjectif extrait en tant que nom. Dans ces cas aussi, les syncons concernés ont été notés de côté pour que leur catégorie grammaticale soit corrigée par la suite.

C'est pour ces raisons qu'il est nécessaire de vérifier les syncons associés à un lemme Sensi. Cependant les cas les plus importants lors de cette étape sont les syncons en double ou les syncons erronés. Il arrive qu'un syncon existe mais que le glossa soit faux, ou non explicite et ne permette pas de d'effectuer une désambiguïsation correcte. La figure 41 présente un cas où les glossas ne peuvent être désambiguïsés l'un par rapport à l'autre.

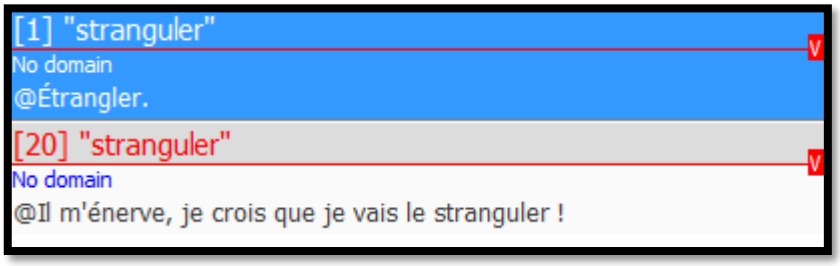


Figure 41 : Capture d'écran de syncons associés au lemme Sensi « stranguler ».

Ainsi, il est possible de noter qu'un syncon doit être supprimé par la balise #deleted# ou #review#.

B. Appliquer les fréquences

La fréquence est une information qui est donnée en fonction de la relation entre un lemme et un syncon (voir Partie 1 ; Chapitre 2 ; section 2. ; C. ; 3.).

Cette étape consiste à renseigner pour chaque syncon une fréquence d'utilisation dans la langue française standard par rapport aux autres acceptions pour le même lemme Sensi donné. Cette « distribution » des fréquences est relativement flexible et s'applique à l'aide d'une grille d'échelonnage des fréquences présentée dans la Partie 1 ; Chapitre 2 ; section 2. ; C. ; 3.

2. Objectifs

Une liste de 4407 termes à traiter nous a été fournie. En moyenne, entre 150 et 250 lemmes sont traités par jour et par une personne en fonction du nombre de syncons associés et de la difficulté d'analyse. Nous avons effectué cette tâche en 346h de travail.

3. Problèmes rencontrés et réflexions

Les erreurs d'extractions issues de l'extraction des ressources dictionnairiques forment un problème qui se répercute sur plusieurs niveaux de travail. Ce problème a été cité lors de la tâche de test de comparaison sur le Sensigrafo FR avant l'ajustement des fréquences. Le nombre de syncons incorrects allonge le temps de travail d'ajustement des fréquences car il est nécessaire de vérifier si le syncon est bel et bien une erreur d'extraction avant de le supprimer.

La finesse des descriptions des syncons est également un point qui nous a posé problème. Même en étant un locuteur natif de la langue française il est parfois difficile de distinguer des descriptions entre elles, par exemple dans le cas des phrases suivantes pour le lemme Sensi « repasser ».

« *Passer, faire traverser, transporter* de nouveau ou en arrière.* »

« *Franchir, traverser* de nouveau ou en retournant.* »

« *[retourner, revenir] Passer en arrière, retourner à l'endroit d'où l'on est venu.* »

Le passif d'une personne influence ses connaissances et sa vision de la norme. Ainsi, il est possible que pour une personne venant du sud-ouest, le concept du lemme Sensi « poche » en tant que sac plastique soit une acception parfaitement répandue alors que cette acception est inconnue pour un grand nombre de personnes dans le reste de la France.

La gestion du féminin et du masculin est un point qui a parfois posé problème. Il est apparu que des lemmes Sensi étaient en réalité des formes féminines de lemme Sensi existant. En effet, dans le Sensigrafo FR éditable il existe le lemme Sensi « chancelière » ainsi que le lemme Sensi « chancelier ». Les syncons associés à « chancelière » sont des acceptions ne pouvant apparaître que sous la forme féminine. Cependant dans la logique de structuration du réseau, le lemme Sensi « chancelière » devrait être supprimé. Toutefois, cette forme doit être associée au lemme Sensi « chancelier » dans le dictionnaire de flexions. Les syncons ne pouvant apparaître qu'à la forme féminine devront porter cette indication dans leurs informations comme décrit dans la Partie 1 de ce document. Cette stratégie de

structuration à l'inconvénient qu'il n'est pas possible de rechercher un lemme Sensi à la forme féminine. Ainsi, si le syncon décrivant la femme d'un chancelier portant la forme chancelière doit être consulté, la recherche s'effectue à partir du lemme Sensi « chancelier ». Le concept de la cigarette de la marque Gauloises, instancié uniquement par les formes féminines du singulier et du pluriel dans la langue française, est associé au lemme Sensi « gaulois » par exemple.

Les locutions sont des éléments à part. En effet, les locutions sont des regroupements de mots portant un sens unique à ce regroupement. Il est difficile d'associer une locution à l'un des lemmes Sensi présents dans ce regroupement. La question du choix du lemme sélectionné pour l'association ne comporte pas de solution évidente. Par exemple pour la locution « voir midi à sa porte », « voir », « midi » et « porte » pourraient être candidat pour porter l'association, seulement aucun ne fait sens car le regroupement en lui-même fait sens seulement.

La question du temps nécessaire pour effectuer ce travail se pose ici car c'est un travail qui est extrêmement long car la grande majorité des personnes ne connaît pas toutes les acceptions dans tous les domaines et les époques possibles pour chaque mot. Il est donc nécessaire de vérifier l'existence d'une acception dans un grand nombre de cas. De là, le questionnement sur la reconstruction de la base dictionnairique intervient une nouvelle fois. Lors de ce stage nous avons effectué ce travail sur 4400 lemmes Sensi en environs 340h. Le Sensigrafo FR compte à ce jour 149 000 syncons de noms, verbes, adjectifs et adverbes. Nous savons qu'un certain nombre a été traité de manière automatique pour l'ajustement des fréquences mais plusieurs milliers restent à faire.

Chapitre 3 - Étiquetage d'entités nommées avec le XTagger

L'étiquetage des entités nommées avec le XTagger correspond à une phase de test du système de reconnaissance de ces données sémantiques.

1. Présentation du travail

Cette tâche est une phase de test de l'étiquetage automatique des entités nommées. Les entités nommées sont des unités textuelles saillantes, sémantiquement parlant, contenues dans les textes. Elles expriment des informations pertinentes en partie pour certaines tâches comme l'extraction d'informations, la catégorisation, la veille de données documentaires etc… Elles peuvent exprimer différentes informations comme les lieux, les adresses, les

personnes, les dates… Il est important de définir les types d'entités nommées qui seront utiles en fonction du travail effectué. Il a été décidé, ici, de travailler sur une liste réduite de types d'entités énumérées dans le tableau 15. Trois catégories d'entités ont été choisies pour les décrire. Il y a les entités les plus importantes ; ce sont celles qui sont les plus attendues et qui représenteront la plus grande part des entités. Ces entités sont représentées en orange dans le tableau 15. D'autres entités du même ordre mais d'importance et de fréquence moindres sont représentées en vert dans le tableau 15. Les entités de structures correspondent à des entités permettant de reconnaître la structure du texte. Ces entités structurelles sont représentées en bleu dans le tableau 15. Toutes les descriptions sont fournies à l'annexe 5.

Cette tâche est un test effectué sur l'extracteur d'entités nommées. L'interface du XTagger permet de désambiguïser un texte et d'en extraire et annoter les entités nommées présentes dans ce dernier. Une phase de vérification des annotations est importante car elle permet de connaître le pourcentage d'entités qui n'ont pas été annotées correctement. Pour cela la première étape consiste à vérifier si les entités annotées le sont correctement. Si l'une d'entre elles n'est pas correctement annotée il est possible de corriger l'erreur en changeant le type ou la forme. La deuxième étape consiste à rechercher les entités qui n'ont pas été annotées. Si une entité n'a pas été annotée il est possible de le faire manuellement. Les textes traités produisent en sortie un fichier au format RTF et un fichier au format log. Le premier fichier contient le texte originel avec la liste des étiquettes d'entités présentes dans le texte. Le second fichier contient des informations non pertinentes pour ce travail. Il est important de dire que l'objectif de la reconnaissance des entités nommées, ici, n'est pas de reconnaître toutes les apparitions d'une entité, une seule occurrence suffit.

A. Le corpus

L'objectif est de posséder un corpus de 1000 textes en français pour lesquels les entités sont étiquetées. Nous avons constitué un corpus de 400 articles pour compléter le corpus existant. L'association de ces deux corpus nous permet d'atteindre l'objectif des 1000 articles. Nous avons traité l'intégralité de notre corpus de 400 articles en 223h de travail.

Ce corpus est un corpus d'articles de presse issus du web et ne traitent jamais le même sujet. Les sources sont : 20 minutes, Auto Moto, Corse-Matin, Cuisine et Vins, Direct-Matin, Femme Actuelle, Le Figaro, Le Figaro Madame, GEO, Le Huffington Post, L'Humanité, Le Parisien, La Montagne, La Provence, L'Alsace, Le Dauphiné Libéré, L'Équipe, L'Étudiant, Ouest France, Libération, Le Monde, La Voix Du Nord, Métro,

Phosphore, Sciences et Avenir ainsi que Télérama. Le but est d'avoir un maximum d'entités pour établir un corpus de test hétérogène. Nous appellerons ce corpus « RTF1 ».

Types les plus importants		Autres		Structures	
NPH	Personne/humains	PRD	Produits	MEA	Mesures
ORG	Organisations / Institutions	DEV	Appareils	MON	Monnaies
COM	Compagnies / Sociétés / Entreprises	VCL	Véhicules	DIG	Numéros
MMD	Médias de Masses / Réseaux sociaux	FDD	Nourriture et Brevages	PHO	Numéros de téléphone
GEO	Zones Géographiques Administratives	BLD	Bâtiments	SSN	Numéros de sécurité sociale
GEA	Zones Géographiques Naturelles	WRK	Ouvrages produits par des intelligences humaines	HOU	Heures
GEX	Zones Géographiques étendues	DOC	Documents textuels	SCO	Scores
		EVN	Évènements	EMO	Émoticons
		HOL	Fêtes	FOL	Dossiers
		LEN	Entités légales	WEB	URL
		PPH	Phénomène physique	MAI	Adresses e-mail
		ANM	Animaux	ADR	Adresses postales
				PCT	Pourcentages
				DAT	Dates

Tableau 15 : Tableau des types d'entités nommées utilisés avec leur abréviations.

B. Vérification et correction

La première étape de cette tâche est de vérifier si les entités étiquetées l'ont été correctement. La figure 42 est un exemple de texte désambiguïsé dans le XTagger sur l'onglet des entités nommées. Le texte traité se trouve dans l'encadré (1). Les entités reconnues par l'outil sont surlignées selon les couleurs associées au type reconnu. Ces informations sont indiquées dans l'encadré (3). D'ailleurs ici « Eduardo Paes » est une personne. Cette entité a été reconnue mais n'a pas été enregistrée car le type n'a pas été reconnu. Il est donc nécessaire de sélectionner « Eduardo Paes » et de l'enregistrer avec le bouton « NPH » dans l'encadré (2). « Cool Japan » étant un concept il ne sera pas étiqueté. Cela dit « Maracana » est reconnu comme une personne (figure 42 (1)) alors que ce terme fait référence, ici, à un bâtiment et non à une personne. « Super Marion » est noté en tant qu'entreprise (figure 42 (2)) alors qu'il s'agit d'une personne.

Dans la figure 42 (3) à droite d'une entité étiquetée se trouve un identifiant de syncon. Lorsque le syncon d'une entité existe dans le Sensigrafo FR fixe il doit être renseigné.

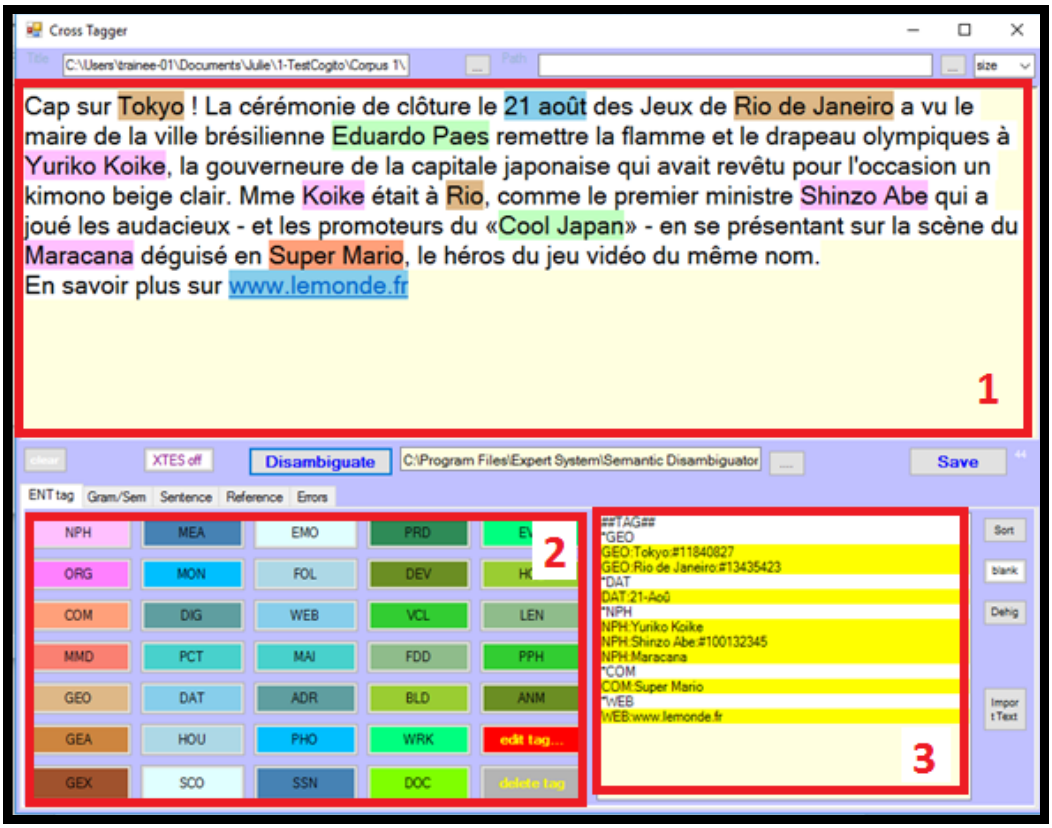


Figure 42 : Capture d'écran d'une désambiguïsation des entités nommées sur un texte dans le XTagger (version5).

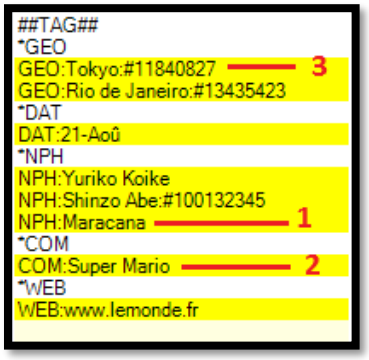


Figure 43 : Capture d'écran de la liste des étiquettes d'entités nommées de l'exemple donné dans la figure 42.

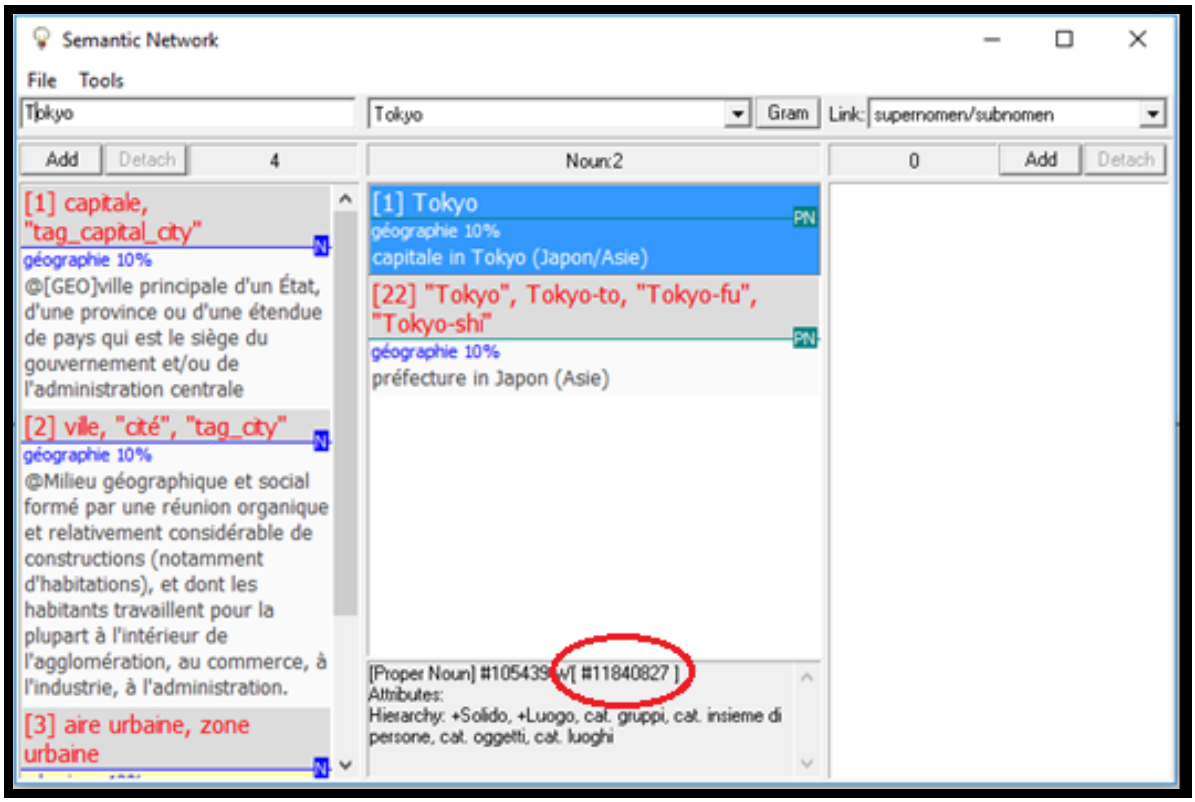


Figure 44 : Capture d'écran des syncons associés au lemme Sensi « Tokyo ».

C. Résultats

Nous avons effectué des calculs sur la précision[15] et le rappel[16] de la désambiguïsation des entités nommées. Le Cogito Desambiguator propose une fonction d'extraction des informations issues des fichiers au format RTF. L'application de cette fonctionnalité permet d'obtenir le nombre d'entités présentes dans le texte. Ce résultat

[15] Précision : La précision est un calcul donnant le nombre d'éléments correctement reconnus par rapport au nombre d'éléments reconnus.
[16] Rappel : Le rappel est un calcul donnant le nombre d'éléments correctement reconnus par rapport au nombre d'éléments qui devrait être reconnus.

correspond au nombre d'étiquettes associées à un texte après la tâche d'étiquetage manuel présenté dans ce chapitre. Le nombre d'entités qui ont été ajoutées après la désambiguïsation du logiciel représente le silence, ce sont les entités non reconnues par le XTagger. Le nombre d'entités qui ont été reconnues mais de manière incorrecte représente les reconnaissances incorrectes de désambiguïsation. Le nombre d'entités qui ont été reconnues comme des entités, alors qu'elles n'en sont pas, représente le bruit.

Nombre d'entités correctement reconnues	Nombre d'entités non reconnues	Nombre d'entités reconnues incorrectement
6986	4083	1652

Tableau 16 : Totaux des entités correctement reconnues, des entités non reconnues et des entités reconnues de manière incorrecte.

1. Précision

Précision : nombre d'entités correctement reconnues / nombre d'entités reconnues
Nombre d'entités reconnues = nombre d'entités correctement reconnues + nombre d'entités reconnues incorrectement.
Précision = (6986/(6986+1652))*100 = 80.88%

2. Rappel

Rappel : nombre d'entités correctement reconnues / nombre réel d'entités
Nombre réel d'entités = Nombre d'entités correctement reconnues + Nombre d'entités non reconnues.
Rappel = (6986/(6986+4083))*100 = 63.11%

2. Problèmes rencontrés et réflexions

Nous avons rencontré régulièrement des entités dont le type était ambiguë. Les cas suivants sont des cas que nous avons rencontrés pour lesquels nous avons dû discuter pour nous mettre d'accord sur le type à attribuer.

1. Métonymie entre concepts de territoire et d'organisation politique

Certains éléments sont parfois ambigus. Dans un contexte ils appartiendront à une catégorie alors que dans une autre situation elles appartiendront à une catégorie complètement différente. Par exemple, un Reich est un terme qui à l'origine désigne un territoire sur lequel s'applique le pouvoir d'un prince, roi, empereur ou État. Il sera étiqueté

à ce moment-là en tant qu'entité de géographie administrative (GEO). Cependant, la république de Weimar et le régime Nazi ont également la dénomination de Reich alors qu'ils représentent des régimes. Dans ce cas, ces éléments seront étiquetés en tant qu'organisations (ORG). De plus, le concept du régime comprend également une notion de territoire sur lesquelles s'applique ce régime. Nous avons décidé d'appliquer l'étiquette d'organisation car cette entité comprend la notion de territoire.

2. Métonymie[17] entre territoire et organisation sportive

Une autre difficulté rencontrée est celle des équipes sportives représentant une nation. En effet, il est courant d'entendre parler d'un match « France-Italie » ou d'une rencontre « Allemagne-Brésil ». Cependant la question que nous nous sommes posée est de savoir si l'utilisation des termes France, Italie, Allemagne et Brésil représentent un pays (GEO) ou bien une équipe (ORG). Nous avons décidé d'accorder l'étiquette ORG aux équipes sportives dans la logique sémantique qu'une organisation peut être sujet de verbe auquel un élément géographique ne peut être sujet.

3. Métonymie entre concepts de bâtiment et d'organisation politique

La personnification d'élément inanimé est un phénomène courant comme avec des institutions portant le nom de leur bâtiment. Prenons les exemples de « Bercy » et de « Élysée ». Ce sont à l'origine des bâtiments (BLD) dans lesquelles se trouvent des institutions (ORG). Ainsi, les exemples précédents peuvent être rencontrés en tant que bâtiment :

> « Michel Sapin remplace Emmanuel Macron et devient le premier super ministre à Bercy depuis 2007. »
>
> « L'actuel hôte de **l'Élysée** est François Hollande, président de la République française depuis le 15 mai 2012. »

Les phrases suivantes présente les éléments « Bercy » et « Élysée » en tant qu'organisation cette fois :

> « Résultat, plus de 18 millions de contribuables ont déclaré leurs revenus de façon dématérialisée, selon un bilan dévoilé ce lundi par **Bercy**. »

[17] Métonymie : La métonymie est une figure de style où l'utilisation d'un concept est remplacée par un autre concept avec lequel il a un lien logique direct.

« Les Etats-Unis ont bien espionné **l'Elysée** en 2012, selon l'ex-directeur technique de la DGSE. »

4. Nouveautés

Les hashtags, appelés mot-dièse ou mot-clic parfois, sont des marqueurs de métadonnées utilisés sur internet. Ils permettent d'annoter un contenu avec un mot clé. Il s'agit d'un élément particulièrement utilisé sur les réseaux sociaux. Cet élément ne correspond pas à une de nos catégories. Cet élément n'a donc pas été pris en compte lors de ce stage, mais il correspond à une donnée intéressante qu'il faudra probablement prendre en compte dans un futur proche. En effet elle apparaît maintenant très souvent et pourrait être utilisée au sein de données de type verbatim[18].

5. Ambiguïtés d'états

Nous nous sommes également retrouvé face à des entités archéologiques de différents types et la question de la catégorie d'entité s'est posée. Dans le cas où nous sommes en présence d'un site archéologique avec des ruines encore débout nous sommes en présence d'une entité zone géographique administrative (GEO) ou d'un bâtiment (BLD). Le cas des arènes de Nîmes sera considéré en tant que bâtiment (BLD), mais le cas d'une zone comprenant plusieurs bâtiments comme le site archéologique du Michu Pichu est considéré en tant que zone géographique naturelle (GEA). Si nous sommes en présence d'un site archéologique rasé jusqu'au sol nous sommes en présence d'une zone géographique naturelle (GEA). Si un musée ou un bâtiment de travail se trouve sur le même site comme le site archéologique de Saint-Romain-en-Gal qui fait partie du musée gallo romain à Vienne (France) alors c'est un bâtiment (BLD).

Nous l'avons vu les utilisations pas métonymie sont courantes est ce phénomène est un phénomène qui est compliqué à désambiguïser. L'ajout de syncons différents entre les types d'entités pour les noms propres pourrait permettre de résoudre en partie cette difficulté grâce aux informations attribuées qui seront différentes.

[18]Verbatim : Les verbatim sont des comptes rendus fidèles, des reproductions intégrales de propos prononcés par quelqu'un. Un verbatim peut être la reproduction d'un propos prononcé lors d'une enquête téléphonique par exemple. Ce terme est également largement utilisé pour exprimer un commentaire ou message laissé par un utilisateur comme par exemple une assurance qui va recevoir questions ou les réclamations de ses clients.

Chapitre 4 – Mise en correspondance de syncons via le QClient

La mise en correspondance des syncons est une phase de développement qui s'applique sur le Sensigrafo FR éditable mais également sur le Sensigrafo EN éditable. Le but ici est de faire des liaisons entre les deux réseaux ce qui facilitera l'héritage d'informations d'un réseau à l'autre. C'est une phase de mise en place clé qui permet d'effectuer l'étape charnière de la dévirtualisation.

1. Présentation du travail

Cette tâche est une tâche de développement du réseau Sensigrafo FR. La mise en correspondance de syncons entre deux Sensigrafos monolingues est une étape très importante dans le développement d'un Sensigrafo monolingue. Cette étape de développement est présentée dans la Partie 1 de ce document au Chapitre 2 point 4.

Lors du stage, une liste de plusieurs milliers de lemmes Sensi nous a été fournie (voir annexe 7). Le travail consiste à rentrer un lemme Sensi dans le QClient et pour chaque acception trouver un syncon issu du Sensigrafo EN correspondant de manière conceptuelle.

La tâche n'est pas toujours simple, en particulier pour les termes très spécifiques ce qui n'est pas rare dans la base dictionnairique du français. Prenons un cas rencontré : un « croisé » est un terme technique en escrime désignant un mouvement de l'épée pour faire sauter l'épée de la main de l'adversaire.

Plusieurs stratégies sont possibles pour trouver un syncon de correspondance conceptuelle. L'approche classique de mise en correspondance est de chercher la traduction la plus proche que nous utiliserions naturellement. Pour le lemme Sensi FR « chat » avec l'acception de l'animal nous allons rentrer le lemme Sensi EN « cat ». Pour le lemme Sensi FR « chat » avec l'acception du forum de discussions nous allons rentrer le lemme Sensi EN « chat ». Cependant, parfois l'approche classique ne fonctionne pas. Dans ce cas il est possible de chercher à partir d'un père ou d'un fils sémantique. Prenons le cas d'une couleur dont nous ne connaissons pas la traduction. Il est possible de rechercher la catégorie des couleurs dans le Sensigrafo EN et de faire défiler les fils dans la colonne de droite. Il est possible que le Sensigrafo EN n'ai simplement pas de syncon équivalent. Dans ce cas il est possible de mettre en correspondance un père ou un fils sémantique avec l'information que l'un des deux concepts (en précisant lequel) est un générique de l'autre concept.

Ces stratégies sont parfois utilisées pour résoudre des problèmes de chaînes hiérarchiques conceptuelles incomplètes dans la phase de dévirtualisation.

Une fois que nous avons mis en correspondance des syncons entre eux nous pratiquons une étape de revue croisée. C'est-à-dire que nous échangeons nos listes entre personnes travaillant sur cette tâche. Cette revue croisée permet d'avoir plusieurs avis sur une mise en correspondance afin de standardiser la méthodologie à employer pour résoudre les différents cas.

2. Problèmes rencontrés et réflexions

Nous nous sommes retrouvé face à un problème pour lequel nous n'avons pas trouvé de solution. Nous nous sommes rendu compte qu'il serait pertinent, parfois, de pouvoir chercher un équivalent dans une autre langue par la définition plus que par un terme d'instanciation car nous avons un lemme Sensi à disposition mais parfois il n'aide pas à la recherche. Le second élément que nous avons à disposition est le glossa lui-même. Nous avons cherché des outils sur internet pour faire une recherche de traduction par mots-clés. Nous avons trouvé le site onelook.com qui propose cette fonctionnalité mais les résultats ne sont pas concluant.

Nous avons parlé des termes rares ou obsolètes précédemment dans les stratégies de recherche. Ce problème est, à nouveau, lié au problème de la base dictionnairique qui comporte un grand nombre de termes de spécialité.

Nous savons que la mise en correspondance permet d'appliquer l'héritage des informations, d'un syncon du Sensigrafo source vers le syncon d'un Sensigrafo cible, lors de la dévirtualisation. Alors, il est important de mettre en correspondance des syncons ne pouvant comporter que les mêmes informations exactement. Certaines informations grammaticales risquent d'être erronées. En effet, la position par rapport à d'autres éléments dans la phrase est une information des adjectif par exemple. Et nous savons que les adjectifs en français et en anglais se positionnent différemment. Prenons l'exemple des couleurs tout simplement. Pour la phrase « Mon père a une voiture **rouge**. » la traduction est « My dad has a **red** car. ». Les syncons français et anglais du concept de la couleur rouge risque d'être, logiquement, mis en correspondance. Et lors de l'héritage le syncon français recevra l'information de positionnement avant un nom ce qui est faux. Le fait de prendre en compte toutes les informations possibles est une gymnastique de l'esprit qui est loin d'être simple et nous sommes attentifs à la répercussion de certaines informations.

Chapitre 5 - Étiquetage d'erreurs avec le XTagger

Le travail sur le Sensigrafo FR a débuté environ 6 mois avant le début du stage et cette phase de test se situe entre 4 et 6 mois après le début du stage ce qui permet d'avoir une version du Sensigrafo FR fixe retravaillé par rapport à la tâche présentée dans le Chapitre 1 de la Partie 2. Le but est d'obtenir les nouveaux résultats des désambiguïsations du moteur d'analyse sémantique.

1. Présentation du travail

Cette tâche est une phase de test du moteur d'analyse sémantique du français. Ce travail nécessitait un nouveau corpus. Il a été demandé que ce corpus soit plus varié que les précédents avec des textes de différents types :

- Actualités françaises et francophones
    - Domaines : politique, économie, culture et social
- Publicités
- Textes scientifiques
- Littératures
- Réseaux sociaux et forums
- E-mails

Les textes n'ont jamais le même sujet. L'objectif était d'avoir environ 70 000 concepts (noms, verbes, adverbes et adjectifs), ce qui correspond à environ 140 000 mots.

A l'aide du XTagger (version 4 ou 5), il est possible d'étiqueter les mots en leur donnant un type d'erreur :

- Concept absent = Le concept n'existe pas dans le Sensigrafo FR.
- Reconnaissance incorrecte = Le lemme reconnu est incorrect.
- Désambiguïsation incorrecte = Le syncon reconnu est incorrect.
- Grammaire incorrecte = La classe grammaticale associée est incorrecte.
- Entité absente = Le terme devrait être identifié en tant qu'entité mais ce n'est pas le cas.
- Type d'entité incorrect = Le type d'entité identifié est incorrect.
- Forme d'entité incorrecte = La segmentation de l'entité n'est pas correcte.

Il est également possible d'étiqueter l'erreur avec un indice de gravité allant de 1 à 3.

La gravité 1 correspond à une erreur légère, qui pourrait être acceptée. Par exemple, quand il est question de former une équipe pour un jeu ou une activité quelconque la définition : « *Groupe de personnes unies dans une tâche commune.* » paraît plus adaptée que la définition « *Groupe de joueurs pratiquant un même sport et associés en nombre déterminé pour disputer des compétitions, des matches, des championnats.* ». Cependant, la seconde définition pourrait être acceptée quand même.

La gravité 2 correspond aux erreurs dites lourdes. Ce sont les erreurs qu'il est impératif de corriger. Par exemple, prenons un contexte où le terme « univers » désigne l'univers d'une personne ou d'une histoire comme l'univers du « Disque-monde » de Terry Pratchet. Si la définition associée à « univers » dans ce contexte est « *Ensemble des étoiles et des planètes.* » alors c'est une erreur de gravité 2 car l'analyseur aurait dû sélectionner le syncon « *Monde, milieu réel, matériel ou mental* ».

La gravité 3 correspond aux erreurs très graves. Ce sont les erreurs ne pouvant être présentées à des clients. Par exemple le fait qu'un concept ne soit pas présent dans le Sensigrafo FR fixe est une erreur très grave. Nous sommes plus tolérants lorsqu'un syncon n'existe pas si le concept est un concept très spécifique comme le nom d'un produit récent ou peu connu comme par exemple « Gamescape » qui est le nom d'une petite entreprise d'activités de groupe. Les erreurs de catégorie grammaticale sont dans la grande majorité notée avec la gravité 3.

Source	Nombres de mots pleins	Nombres d'erreurs	Pourcentages d'erreurs par rapport aux nombres de mots pleins	
Article 1	129	56	43,41%	
Article 2	114	42	36,84%	
Article 3	221	131	59,28%	
Article 4	429	196	45,69%	
Article 5	185	90	48,65%	
Article 6	124	49	39,52%	
Article 7	401	187	46,63%	
Article 8	310	120	38,71%	
Article 9	86	46	53,49%	
Article 10	122	50	40,98%	
Totaux	2121	967	45%	Moyenne

Tableau 17 : Répartition des erreurs d'un corpus de 10 articles.

	Gravité 1	Gravité 2	Gravité 3	Totaux	%
Syncon manquant	1	8	154	163	16,87%
Entité manquante	17	17	18	52	5,38%
Désambiguïsation incorrecte	255	272	45	572	59,21%
Forme de l'entité incorrecte	7	17	30	54	5,59%
Type de l'entité incorrect	14	2	7	23	2,38%
Catégorie grammaticale incorrecte	2	6	56	64	6,63%
Reconnaissance incorrecte	2	6	30	38	3,93%
Totaux	298	328	340	966	966=100%
%	30,85%	33,95%	35,20%	966=100%	

Tableau 18 : Répartition des erreurs du corpus selon l'indice de gravité et les types d'erreurs.

Prenons un exemple concret. Le texte suivant a été annoté des erreurs effectuées lors des désambiguïsations dans le XTagger. Les éléments surlignés en jaune représentent les erreurs de gravité 1, les éléments surlignés en vert sont les erreurs de gravité 2 et en rouge ce sont les erreurs de gravité 3. Le tableau 19 est une reproduction partielle des informations sur les erreurs fournies par le XTagger à la suite de la sauvegarde de l'étiquetage des erreurs d'un texte.

Source : Article 1

Dans un live escape game, vous êtes les héros ! Gamescape vous propose trois live escape game au cœur de Paris. Dans des décors inspirés des univers de Méliès, Jules Verne, Tolkien et J.K. Rowling, venez vous mesurer à nos live escape game ! Gamescape vous fait voyager dans le temps pour vivre les légendes de la ville lumière comme si vous y étiez. Votre cohésion fera votre victoire ! Enfermés dans une pièce mystérieuse 60 minutes pour vous échapper Par équipe de 3 à 5 joueurs. Le principe du "Live escape game" est simple. Vous et votre équipe (3 à 5 joueurs) serez enfermés dans une pièce étrange appelée escape room et il faudra vous en échapper. Comment, me direz-vous ? Dans un live escape game il faut être astucieux, débrouillard et... rapide ! Vous n'avez que 60 minutes pour triompher. Chaque seconde compte, vous et votre équipe devrez faire preuve d'une grande cohésion si vous voulez réussir à vous évader. Gamescape propose trois jeux d'évasion "live escape game" au cœur de Paris. Chez Gamescape, trois salles de live escape game différentes vous attendent avec chacune son ambiance, ses énigmes, sa difficulté. Que vous soyez novices ou experts dans l'art de l'évasion, vous trouverez un challenge à votre niveau chez Gamescape. Toutes nos escape rooms ont pour thème des histoires emblématiques de la ville de Paris. Toutes regorgent de surprises et de mystères. Vous ne saurez jamais vraiment quand le jeu démarre et quand il s'arrête.

Légende des colonnes du tableau 19 :

1 – Type de l'erreur

2 – Indice de Gravité

3 – Catégorie grammaticale reconnue par le XTagger

4 – Élément concerné

5 – Identifiant standard du syncon reconnu

6 – Position de l'élément dans le texte (la ponctuation est comprise)

1	2	3	4	5	6
Entité manquante	1	NOU	game	128825	5
Entité manquante	0	NPR	Gamescape	-1	12
Syncon manquant	2	VER	propose	85412	14
Entité manquante	1	NOU	game	128825	18
Désambiguïsation incorrecte	1	NOU	décors	28203	26
Désambiguïsation incorrecte	1	NOU	univers	22033	29
Désambiguïsation incorrecte	1	VER	mesurer	82148	41
Syncon manquant	1	NOU	game	128825	46
Entité manquante	0	NPR	Gamescape	-1	48
Désambiguïsation incorrecte	1	VER	fait	70822	50
Désambiguïsation incorrecte	1	VER	voyager	91387	51
Désambiguïsation incorrecte	0	NOU	légendes	31218	58
Reconnaissance incorrecte	2	NOU	lumière	866	62
Catégorie grammaticale incorrecte	2	NOU	etiez	-1	66
Syncon manquant	2	NOU	cohésion	22750	69
Désambiguïsation incorrecte	1	VER	Enfermés	72796	74
Syncon manquant	2	NOU	pièce	29498	77
Désambiguïsation incorrecte	0	ADJ	mystérieuse	103492	78
Désambiguïsation incorrecte	0	NOU	équipe	1124	84
Désambiguïsation incorrecte	1	NOU	principe	23820	92
Syncon manquant	1	NOU	game	128825	97
Désambiguïsation incorrecte	0	NOU	équipe	1124	104
Désambiguïsation incorrecte	1	VER	enfermés	72796	112
Syncon manquant	2	NOU	pièce	29498	115
Syncon manquant	2	NOU	escape	30393	118
Désambiguïsation incorrecte	1	VER	faudra	79628	122
Désambiguïsation incorrecte	0	VER	échapper	73595	125
Désambiguïsation incorrecte	0	VER	échapper	74957	82
Syncon manquant	1	NOU	game	128825	138
Désambiguïsation incorrecte	1	VER	faut	79531	140
Désambiguïsation incorrecte	1	ADJ	rapide	97055	147
Désambiguïsation incorrecte	2	NOU	seconde	59311	158
Syncon manquant	2	VER	compte	86524	159
Désambiguïsation incorrecte	0	NOU	équipe	1124	164
Syncon manquant	2	VER	faire	74611	166
Désambiguïsation incorrecte	1	ADJ	grande	6414	170
Syncon manquant	2	NOU	cohésion	22747	171
Désambiguïsation incorrecte	1	VER	évader	73595	178
Entité manquante	0	NPR	Gamescape	-1	180
Syncon manquant	2	VER	propose	85412	181
Désambiguïsation incorrecte	1	NOU	jeux	1354	183
Syncon manquant	1	NOU	game	128825	189
Entité manquante	0	NPR	Gamescape	-1	196
Syncon manquant	2	NOU	salles	1E+08	199
Syncon manquant	1	NOU	game	128825	203
Désambiguïsation incorrecte	1	ADJ	différentes	94403	204
Désambiguïsation incorrecte	2	NOU	difficulté	21142	216
Désambiguïsation incorrecte	1	NOU	novices	46508	221
Désambiguïsation incorrecte	1	NOU	challenge	20871	234
Désambiguïsation incorrecte	1	NOU	niveau	36398	237
Entité manquante	0	NPR	Gamescape	-1	239
Syncon manquant	2	NOU	escape	30393	243
Syncon manquant	2	NOU	thème	23306	247
Désambiguïsation incorrecte	0	NOU	surprises	188158	260
Désambiguïsation incorrecte	1	NOU	mystères	45384	263
Désambiguïsation incorrecte	1	VER	arrête	81507	278

Tableau 19 : Tableau des erreurs du texte en exemple page 81.

	Gravité 1	Gravité 2	Gravité 3	Totaux
Syncon manquant	0	5	12	17
	0%	8.9%	21.4%	30.3%
Entité manquante	5	2	0	7
	8.9%	3.6%	0%	12.5%
Désambiguïsation incorrecte	8	20	2	30
	14.3%	35.7%	3.6%	53.6%
Forme de l'entité incorrecte	0	0	0	0
	0%	0%	0%	0%
Type de l'entité incorrecte	0	0	0	0
	0%	0%	0%	0%
Catégorie grammaticale incorrecte	0	0	1	1
	0%	0%	1.8%	1.8%
Reconnaissance incorrecte	0	0	1	1
	0%	0%	1.8%	1.8%
Totaux	13	27	16	56
	23.2%	48.2%	28.6%	100%

Tableau 20 : Tableau de la répartition des erreurs selon l'indice de gravité et les types d'erreurs disponibles lors de la tâche de l'étiquetage des erreurs pour le texte de l'article 1 présent dans l'exemple précédent.

2. Problèmes rencontrés et réflexions

1. Les auxiliaires et les verbes de modalités

La désambiguïsation des auxiliaires et des verbes de modalités n'est pas toujours évidente et s'en trouve parfois erronée. En théorie, ces éléments, tout comme les mots grammaticaux ne sont pas désambiguïsés. Cependant, il arrive que l'un d'entre eux le soit. Cette désambiguïsation peut être à l'origine d'autres erreurs de désambiguïsation.

2. Accentuation

En français les caractères écrits en majuscule peuvent être dépourvus de leur accent. Or le système est très strict et ne reconnaît pas un terme s'il n'a pas son accent. Par exemple pour le mot « sélection », la forme « SELECTION » se sera pas reconnu alors qu'il n'y a pas d'ambiguïté.

Toutefois, les adjectifs possessifs comme « notre » ou « nos » ont dans le Sensigrafo FR fixe un syncon associé avec un seul lemme Sensi qui est « nôtre ». Et pourtant la reconnaissance est systématique alors que l'accent du lemme Sensi est absent dans le texte. Une explication possible serait que le lemme « nôtre » ait, dans le dictionnaire des flexions, les formes « notre » et « nos » associées. Ceci expliquerait pourquoi « selection » n'est pas reconnu puisqu'il ne s'agirait pas d'une flexion de « sélection ».

3. Erreurs de frappe et d'orthographe

Les textes comprennent parfois des erreurs d'orthographe ou de frappe. La question de la tolérance et du niveau de correction possible se pose. Il est possible d'imaginer que les fautes minimes pourraient être prises en compte et désambiguïsées comme dans l'un des outils de l'entreprise qui n'a pas été abordé dans ce document. Mais il y a forcément un moment où ce n'est plus possible. S'il manque trop de caractères ou si les substitutions sont trop nombreuses. Cela dit un travail pourrait être amorcé sur ce sujet dans le contexte du projet du Sensigrafo FR.

4. Lemmes et formes doubles

Il peut arriver qu'on trouve deux lemmes identiques ou deux formes d'un même lemme utilisé comme attribut dans le Sensigrafo FR. Ce phénomène est problématique car certaines définitions sont associées uniquement avec un lemme alors que d'autres le sont uniquement avec l'autre lemme. On rajoute ainsi une étape lors de la désambiguïsation puisqu'une forme trouvée dans un texte devra être désambiguïsée parmi les lemmes le comprenant comme une forme. Plus il y aura de lemmes possibles plus il y aura de choix et donc plus de difficultés à parvenir à une désambiguïsation correcte.

En théorie les lemmes sont répertoriés dans le dictionnaire des flexions. Ainsi, le lemme « gaulois » aura pour flexions : « gaulois, gauloise, gauloises ». Le syncon désignant une personne originaire de la Gaule pourra apparaître sous les trois formes en fonction du genre et du nombre mais le syncon désignant la cigarette ne pourra apparaître qu'avec les formes féminines du singulier et du pluriel. On trouve l'exemple des lemmes « ambiguïté » et « ambigüité » qui forment également un doublon de lemmes.

Le travail d'étiquetage des erreurs effectué en fin de stage correspond à la première tâche de test de performance du Sensigrafo FR à partir de l'outil Cogito Desambiguator. Il est nécessaire d'indiquer que la fonctionnalité de l'étiquetage des erreurs du XTagger n'avait pas encore été développée au début du stage. Ces deux tâches ont été effectuées avec des outils différents ainsi que des versions différentes du Sensigrafo FR. Nous avions également des critères d'erreurs différents. Lors de l'étiquetage des erreurs, via le Xtagger, trois indices de gravité étaient disponibles contre cinq lors du premier test. Cependant, la description des erreurs est plus détaillée dans le premier test. La comparaison des résultats n'est donc pas

exacte. Cela dit, sans comparer les chiffres il est possible de comparer les tendances. Dans les deux cas les résultats montrent que les erreurs de sémantique représentent largement la majorité des erreurs. Le développement du Sensigrafo FR est donc le travail à effectuer en priorité absolue.

.

Partie 3

-

Test proposé

Chapitre 1 - Test de comparaison des différentes versions du Sensigrafo FR via le Cogito Desambiguator

Pour observer l'évolution concrète des désambiguïsations sémantiques obtenues via le Cogito Desambiguator nous avons décidé d'effectuer la tâche d'ajustement des fréquences sur tous les mots d'un texte. C'est-à-dire supprimer les syncons en trop ainsi qu'appliquer l'échelonnage des fréquences sur tous les syncons (voir Partie 2 ; Chapitre 2).

Notre travail durant le stage consistait à appliquer cette tâche d'ajustement des fréquences sur une liste de noms et de verbes (voir annexe 7) nous ayant été fournie. Une fois que nous avons effectué l'ajustement des fréquences sur une liste de mots nous avons voulu savoir si les changements apportés étaient efficaces lors des désambiguïsations. Alors nous avons choisi 19 mots présents dans la liste que nous avions traitée (voir annexe 8). Pour chaque acception d'un mot nous avons écrit une phrase où il apparaît. Par exemple, dans le Sensigrafo FR il existe cinq acceptions du nom « jésus » donc nous avons écrit cinq phrases où le mot « jésus » apparaît en tant que nom avec à chaque fois une définition différente. Nous avions donc au total 77 phrases (le corpus complet est disponible à l'annexe 8). Nous avons attendu que la compilation soit effectuée pour avoir un nouveau Cogito Desambiguator avec une nouvelle version du Sensigrafo FR fixe qui comprenait nos changements effectués. Nous avons effectué ce test deux fois. La première fois nous l'avons fait avant la compilation et la seconde fois après la compilation.

Le premier test a été effectué avec la seconde version du Cogito Desambiguator. Le Sensigrafo FR fixe était à ce moment en français. Nous avons effectué une passe de désambiguïsation par phrase et une phase de désambiguïsation avec toutes les phrases test pour un lemme Sensi. Les résultats sont différents. Pour la première passe le pourcentage de désambiguïsations sémantiques correctes est de 30.7%. Pour la seconde passe le pourcentage de désambiguïsations sémantiques correctes est de 39.7%.

Le second test a été effectué avec la troisième version du Cogito Desambiguator qui comprenait une nouvelle version compilée du Sensigrafo FR fixe. Nous avons effectué les mêmes passes : une où les phrases sont désambiguïsées seule et une ou toutes les phrases test d'un lemme Sensi sont désambiguïsées ensemble. La première passe à un résultat de 38.4% de désambiguïsations sémantiques correctes. La seconde passe présente 32.1% de désambiguïsations sémantiques correctes.

Nous avons pu constater qu'un grand nombre d'erreurs étaient lié à d'autres erreurs dans le texte. Par exemple pour la phrase « Un jésus est un gros saucisson des Vosges. » le terme « jésus » a été désambiguïsé en tant que « enfant, un chérubin » et le terme « saucisson » a été désambiguïsé avec « une personne courte, petite et grosse ».

Nous avons donc décidé de pousser l'expérimentation jusqu'au bout et d'observer la désambiguïsation sur un texte comprenant uniquement des mots pleins se trouvant dans le Sensigrafo FR fixe.

Malheureusement, le temps nous a manqué pour effectuer cette seconde étape car nous avions d'autres objectifs attendus. De plus, il aurait été difficile d'obtenir un texte dans lequel tous les mots présents existent dans le Sensigrafo FR avec l'ajustement des fréquences appliqué et effectif dans le Sensigrafo FR fixe.

Partie 4

-

Conclusion et Retours personnels

Chapitre 1 – Conclusion sur le Sensigrafo FR

Le Sensigrafo FR est en cours de développement. Lors de ce stage nous avons pu effectuer une analyse de l'état du Sensigrafo FR et de son niveau de développement. Nous avons observé des points sur trois niveaux qu'il pourrait être intéressant de retravailler.

1. Les désambiguïsations

Les désambiguïsations sont effectuées par le moteur d'analyse sémantique. Nous nous intéressons ici aux désambiguïsations grammaticales, en particulier celles qui concernent les catégories grammaticales. Il a été mentionné lors du stage que l'analyseur en C++ est un code unique appliqué pour toutes les langues ayant un Sensigrafo monolingue. Nous savons que les désambiguïsations sont effectuées par cet analyseur. Cependant, nous n'avons pas plus d'information sur ce sujet. Cela dit nous nous sommes rendu compte que des erreurs de désambiguïsation de catégorie grammaticale pouvaient survenir. Dans la Partie 2 ; Chapitre 1 ; page 59 le tableau 7, indiquant la répartition des erreurs grammaticales par type, montre que les erreurs de catégorie grammaticale incorrecte représentent plus de 48% des erreurs grammaticales. Les erreurs grammaticales représentent plus de 27% de l'ensemble des erreurs observées. Il serait intéressant de voir si les règles de désambiguïsations des catégories grammaticales, définies dans l'analyseur, peuvent être retravaillées pour améliorer la désambiguïsation des catégories grammaticales. Améliorer cette désambiguïsation permettrait de diminuer le taux d'erreurs de catégorie grammaticale incorrecte qui est le type d'erreurs de grammaire le plus important sur l'ensemble des erreurs grammaticales.

2. La compilation

Nous n'avons pas beaucoup d'informations sur la compilation effectuée. Cela dit nous avons remarqué qu'un certain nombre de syncons sont fusionnés à partir de la logique des regroupements de synonymes. Cependant, la mise en correspondance étant en partie automatique, il arrive que des éléments soient fusionnés alors qu'ils ne sont pas synonymes. Ainsi, on rencontre des fusions étranges mais surtout erronées qui induisent des erreurs de désambiguïsations.

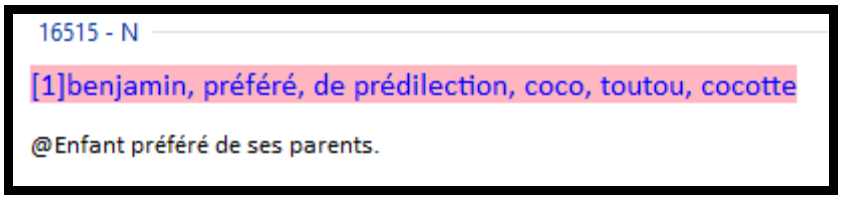


Figure 45 : Capture d'écran d'un regroupement de synonymes automatique incorrect autour du concept du nom « préféré ».

Figure 46 : Capture d'écran d'un regroupement de synonymes automatique incorrect autour du concept du nom « difficulté ».

3. Le Sensigrafo FR

Le Sensigrafo présente un grand nombre de syncons et tous ne sont pas nécessaires aux besoins de l'entreprise. Les définitions sont particulièrement précises ce qui est une difficulté supplémentaire nécessitant un temps de réflexion plus long. La ressource sémantique reste la grande priorité de développement du moteur d'analyse sémantique.

Le travail que nous avons effectué sur ces différents axes nous permet de faire un bilan de l'état du Sensigrafo FR. Dans l'ensemble, nous savons que le développement du français jusqu'ici a été effectué en 6 mois et qu'il nécessitera encore un certain temps de travail.

Pour effectuer tout le travail d'ajustement des fréquences il faudra, selon moi, plusieurs années-hommes. Cette estimation ne comprend pas le travail de mise en correspondance qui risque d'être aussi important que celui de l'ajustement des fréquences. Selon mon point de vue, le choix des ressources sémantiques ralenti le développement par rapport à l'estimation du temps de développement applicable aux autres langues. L'étape primordiale qui suit est de prendre une décision sur la stratégie de développement.

La première stratégie pourrait être que l'entreprise Expert System décide de conserver le Sensigrafo FR avec la base dictionnairique en l'état. Il est nécessaire, dans ce cas, d'effectuer un grand nettoyage sur cette base avant d'appliquer la mise en correspondance. Un travail sur les locutions est aussi à prendre en compte. Ce travail doit être associé à des mises à jour régulières du vocabulaire technique.

La seconde stratégie consisterait à constituer une nouvelle base dictionnairique à partir de ressources sémantiques différentes. Il serait essentiel, dans ce cas-là, d'opter pour une ressource moins détaillée.

Ces deux stratégies permettent d'arriver au résultat souhaité. Cependant pour prendre une décision entre ces deux stratégies il est nécessaire de savoir laquelle s'avèrerait être la plus rapide à réaliser en termes de temps.

Dans les deux cas le travail nécessaire à l'élaboration complète du Sensigrafo FR reste très important.

Chapitre 2 - Retours d'expérience

1. Le travail

Le projet « Sensigrafo FR » a débuté en Italie. Il a été mené pendant plusieurs mois à Modène où les bases du réseau ont été développées. Par la suite le projet a été confié à une équipe française sous l'autorité du responsable du projet « Sensigrafo étendu ».

Il faut garder à l'esprit qu'une fusion entre deux entreprises de nationalités différentes a eu lieu il y a peu de temps. Les habitudes de travail ne sont pas les mêmes ce qui parfois crée des obstacles.

Les tâches étaient manuelles, fastidieuses et difficiles à appréhender. Nous étions la première équipe à faire ce travail donc nous n'avons pas eu de retour d'expérience d'autres personnes sur ce réseau. De plus, nous n'avions pas d'expert sur le Sensigrafo Fr à qui nous pouvions facilement poser des questions. Toutefois notre tutrice a été très attentive à notre position et notre cadre de travail.

Le fait d'avoir traité différentes tâches entre tests et développements durant ces 6 mois de stage nous a permis de diversifier notre vision du réseau. Travailler à différents niveaux (réseau éditable, réseau fixe, réseau français, réseau anglais etc..) nous a permis de mieux comprendre le fonctionnement « interconnecté » du réseau Sensigrafo étendu.

Le fait que nous n'ayons pas eu de formation sur le fonctionnement des Sensigrafos étendu et monolingues autres que les formations sur les outils nécessaires aux tâches et les tâches en elles-mêmes a été regretté. Une formation plus approfondie aurait pris du temps sur le temps de travail consacré aux tests et aux développements mais elle nous aurait permis de mieux comprendre et ce plus rapidement le fonctionnement du réseau. Nous aurions eu une meilleure compréhension de l'utilisation du réseau dans les outils de l'entreprise. Ainsi,

nous aurions pu prendre des décisions en connaissance de cause, donc plus rapidement et à meilleur escient.

Il aurait été intéressant de travailler en parallèle avec des personnes de l'entreprise utilisant des outils appelant le moteur d'analyse sémantique du français pour avoir leur ressenti et leurs commentaires sur la ressource sémantique.

2. L'équipe

Travailler chez Expert System a été une expérience enrichissante. Ce stage a été une première expérience personnelle dans le monde de l'entreprise en particulier dans le domaine de mes études.

Le cadre de travail au bureau de Paris est très agréable. Il y a une réelle liberté d'expression et des personnes en face attentives et disponibles. La quantité et la qualité des échanges avec l'équipe en France sont appréciables. J'ai pu exprimer mes doutes et mes besoins. Cela dit les échanges avec l'équipe basée en Italie n'ont pas été aussi conséquents. Des échanges plus réguliers auraient été bienvenus. Ma mission a été interprétée de différentes manières entre l'équipe en France et l'équipe en Italie. Pour la France notre mission était d'établir une analyse sur l'état du Sensigrafo FR et pour l'Italie, principalement d'appliquer les tâches demandées. J'ai cherché à être, le plus possible, une force de proposition dans la stratégie de développement du Sensigrafo FR.

Ce stage a été une réelle formation au travail en entreprise et à la gestion du travail dans une équipe en particulier dans une équipe de travail répartie sur deux pays. Il est nécessaire de savoir trouver sa place au sein de l'équipe mais également au sein de l'entreprise.

Le travail n'a pas été facile tous les jours mais j'ai réellement apprécié travailler dans cette entreprise. Le projet Sensigrafo est un projet intéressant qui a piqué ma curiosité et m'a donné envie de l'enrichir, de le manipuler et d'observer les résultats et comportements qui en résultent.

Bibliographie - Sitographie

https://fr.wiktionary.org/wiki/Wiktionnaire:Statistiques (consultation 23/08/2016)

https://fr.wikipedia.org/wiki/Dictionnaires_Le_Robert (consultation 23/08/2016)

http://www.geonames.org/about.html (consultation 24/08/2016)

http://www.larousse.fr/dictionnaires/francais/d%C3%A9paysement/23728?q=d%C3%A9paysement#23607 (consultation 30/08/2016)

https://fr.wikipedia.org/wiki/D%C3%A9paysement (consultation 30/08/2016)

Glossaire

Acceptions : Une acception est une des définitions, des utilisations associées à une forme lexicale. Une acception est le sens dans lequel un mot est utilisé.

Catégories sémantique : Les catégories sémantiques dans ce document représentent les catégories conceptuelles permettant de regrouper des concepts.

Chaîne hiérarchique conceptuelle : La chaîne hiérarchique conceptuelle représente les liens entre plusieurs syncons selon une relation hiérarchique conceptuelle.

Collocation : Une collocation est une cooccurrence de mots privilégiée sans être fixe pour autant. Le sens des différents mots peut être plus ou moins interpréter séparément dans la construction d'une collocation.

Fils sémantique : Un fil sémantique est un concept moins général qu'un autre selon une relation sémantique donnée.

Fréquence : La fréquence représente la fréquence d'utilisation d'une acception pour un lemme Sensi donné dans la langue courante.

Genre : Le genre informe si le terme et son acception en présence sont au féminin ou au masculin.

Hyperonymie : L'hyperonymie est une relation sémantique qui signifie « est un type de ». Ainsi, il est possible de dire que si l'élément A est un type de B alors B est un hyperonyme de A.

Interne : Un logiciel interne désigne un logiciel existant et utilisé uniquement au sein d'une entreprise.

Langue cible : La langue cible est la langue pour laquelle le Sensigrafo est en cours de développement.

Langue source : La langue source est la langue pour laquelle le développement du Sensigrafo est terminé.

Lemme : En linguistique, un lemme est la forme de base d'un mot. Par exemple, pour les verbes, les infinitifs sont les lemmes.

Lemme Sensi : Un lemme Sensi est le terme utilisé dans le projet Sensigrafo FR pour désigner la forme lexicale de base permettant d'exprimer un concept.

Local : Un logiciel en local désigne un logiciel fonctionnement de manière autonome sur un poste informatique par opposition à un logiciel en réseau qui est connecté aux serveurs.

Locution : Une locution est une cooccurrence de mots relativement figée. C'est-à-dire qu'il n'est pas forcement possible d'interpréter les sens des mots séparément dans la construction. La construction dans son intégralité portera le sens.

Mapping : Le mapping est un terme technique propre à l'entreprise Expert System équivalent à « mise en correspondance ».

Méronyme : Un méronyme est un concept qu'on peut définir comme étant une « partie de » d'un autre concept. La relation sémantique associée s'appelle la méronymie.

Mot : Un mot est une suite de caractères graphiques ou de sons formant une unité sémantique.

Mots grammaticaux : Les mots grammaticaux sont les mots des catégories grammaticales des articles et déterminants, pronoms, prépositions, conjonctions (de coordination et de subordination) et des interjections.

Mots pleins : Les mots pleins sont les mots des catégories grammaticales : nom, verbe, adverbe et adjectif ainsi que nom propre par opposition aux mots grammaticaux.

Nombre : Le nombre informe si le terme et son acception en présence sont au singulier ou au pluriel.

Noms communs : Les noms communs sont les termes qui désignent en général, une personne, un animal ou une chose par opposition au noms propres qui eux désigne des personnes, animaux ou choses en particulier.

Ontologie : Une ontologie est un ensemble structuré des termes, concepts et liens existant, entre les informations, d'un domaine de connaissance donné. Une ontologie peut décrire un domaine très large comme une langue, ce qui est le cas dans le projet Sensigrafo, ou bien un domaine plus restreint comme Linked Jazz qui représente les relations dans la communauté du Jazz.

Père sémantique : Un père sémantique est un concept général par rapport à un autre selon une relation sémantique donnée.

Précision : La précision est un calcul donnant le nombre d'éléments correctement reconnus par rapport au nombre d'éléments reconnus.

Rappel : Le rappel est un calcul donnant le nombre d'éléments correctement reconnus par rapport au nombre d'éléments qui devrait être reconnus.

Registre : Le registre est un terme désignant un mode d'expression relatif à une situation d'expression particulière. Il est possible de choisir certains termes ou structures grammaticales adaptés selon la situation en présence.

Registre Sensi : Le terme « Registre Sensi » est un terme utilisé dans le projet Sensigrafo FR pour désigner un type d'information relative aux niveaux de langue associés aux syncons.

Relation hiérarchique conceptuelle : Une relation hiérarchique conceptuelle est une relation entre deux syncons selon une relation sémantique donnée.

(en) Réseau : Une interface en réseau signifie que les modifications sont prises en compte directement sur le serveur par opposition au modèle en local où un logiciel peut être autonome sur un poste informatique.

Sensigrafo cible : Le Sensigrafo cible est un réseau sémantique monolingue Sensigrafo dont le développement est en cours.

Sensigrafo source : Le Sensigrafo source est un réseau sémantique monolingue dont le développement est achevé.

Syncon : Syncon est un terme utilisé dans les réseaux sémantiques d'Expert System pour désigner des entités conceptuelles regroupant toutes les informations sur un concept dans une langue.

Text Mining : Le Text Mining est un domaine dans le traitement automatique des langues également appelé « Fouille de textes » qui rassemble des techniques de linguistique, du langage, de la sémantique, des statistiques ainsi que de l'informatique.

Lexique des abréviations

GN pour Groupe Nominal ;

PV pour Prédicat Verbal ;

PN pour Prédicat Nominal (forme : être + nom ou adjectif) ;

GP pour Groupe Prépositionnel ;

GA : Groupe Adjectival ;

GV : Groupe Adverbal ;

GR : Groupe Relatif ;

CN : Conjonction ;

PNT : Groupe Indéfini ;

CL : Fin de phrase ;

CR : Fin de paragraphe ;

Table des illustrations

Table des tableaux

Table des annexes

Annexe 1
Liste des relations sémantiques possibles dans un Sensigrafo monolingue.

ITALIEN	ANGLAIS	SENS
supernomen/subnomen	**supernomen/subnomen**	Le second syncon est "**UN TYPE DE**" du premier syncon.
superverbum/subverbum	**superverbum/subverbum**	Le second syncon est "**UN TYPE DE**" du premier syncon.
omninomen/parsnomen	**omninomen/parsnomen**	Le second syncon est "**UNE PARTIE DE**" du premier syncon.
omninomen/parsnomen (m)	omninomen/parsnomen (m)	(Non utilisé)
master/syncon	master/syncon	Le premier syncon indique le **PRODUCTEUR** du second syncon qui est le **PRODUIT**.
syncon/bias	syncon/bias	(Relation de condition entre le syncon et le domaine du document; Non utilisé).
syncon/corpus	syncon/corpus	Les deux syncons se trouvent dans un même corpus avec une distance maximale de 8 positions.
verbo/soggetto	verb/subject	Le second syncon est **SUJET** du premier syncon.
verbo/c.oggetto	verb/object	Le second syncon est **OBJET DIRECT** du premier syncon.
conflitto/syncon	conflict/syncon	Le premier syncon est un lemme qui entre en conflit avec le second élément qui est une locution lors de la désambiguïsation.
aggettivo/classe	adjective/class	Le second syncon est **DETERMINÉ** par le premier syncon.
syncon/geografia	**syncon/geography**	Le second syncon est une **DIVISION ADMINISTRATIVE** du premier syncon.
struttura/geografia	structures/geography	(Inverse de la relation geography/structures)
storia/geografia	history/geography	(Inverse de la relation geography/history)
geografia/struttura	geography/structures	Le second syncon est une **PARTIE STRUCTURELLE** du premier syncon.
geografia/storia	geography/history	Le second syncon est une **RÉFÉRENCE HISTORIQUE** du premier syncon.
aggettivo/geografia	adjective/geography	Le premier syncon est l'adjectif indiquant une **APPARTENANCE GÉOGRAPHIQUE** du second syncon.
verbo/sostantivo	verb/noun	Les deux syncons sont un verbe et un nom avec la même base lexicale.
avverbio/verbo	adverb/verb	Les deux syncons peuvent être rencontré dans cet ordre.
avverbio/sostantivo	adverb/noun	Les deux syncons peuvent être rencontré dans cet ordre.
contenitore/contenuto	container/content	(Le second syncon est quelque chose contenu dans le premier syncon.)
nome proprio/nazionalità	proper noun/nationality	Le premier syncon est le **GENTILÉ** correspondant au second nom géographique.
verbo/categoria	verb/category	(Le second syncon est un regroupement de syncons similaires au premier syncon.)
avverbio/avverbio	adverb/adverb	Les deux syncons peuvent être rencontré dans cet ordre.
avverbio/aggettivo	adverb/adjective	Les deux syncons peuvent être rencontré dans cet ordre.
syncon/sinonimo	syncon/synonim	(Les deux syncons sont synonymes.)
qualificante/classe	qualifying/class	Le second syncon est fortement modifié par le premier syncon.
aggettivo/contrario	adjective/antonym	Les deux syncons indiquent des qualifications opposées.
syncon/proprietà	syncon/property	(Le second syncon indiquent a propriété possédé par le premier syncon.)
syncon/accorpato	syncon/unification	Les deux syncons sont interchangeables.
appartenente/popolazione	member/population	(Le premier syncon indique un member du second groupe ethnique.)
verbo/copula	verb/copula	Le second syncon indiquent un prédicat lié au premier syncon.
syncon/implicazione	syncon/implication	Le premier syncon induit le second syncon.
syncon/causa	syncon/cause	Le second syncon est la **CAUSE** du premier syncon.
aggettivo/sostantivo	adjective/noun	Les deux syncons sont un adjective et un nom avec la même base lexicale.
verbo/sostantivo-persona	verb/noun-person	Les deux syncons sont un verbe et un nom avec la même base lexicale.
synset/accorpato	synset/unification	(Pour ce qui est des synsets wordnet.)
conflitto/synset	conflict/synset	(Pour ce qui est des synsets wordnet.)
synset/geografia	synset/geography	(Pour ce qui est des synsets wordnet.)
synset/causa	synset/cause	(Pour ce qui est des synsets wordnet.)
synset/implicazione	synset/implication	(Pour ce qui est des synsets wordnet.)
sostantivo/contrario	noun/antonym	Les deux syncons indiquent des concepts opposés.
verbo/contrario	verb/antonym	Les deux syncons indiquent des relations opposées.
avverbio/legame-sostantivo	adverb/tie-noun	(Les deux syncons sont un adverbe et un nom avec la même base lexicale.)
avverbio/legame-aggettivo	adverb/tie-adjective	(Les deux syncons sont un adverbe et un adjective avec la même base lexicale.)
aggettivo (prep)/verbo	adjective (prep)/verb	(Les deux syncons peuvent être rencontré dans cet ordre avec ou sans préposition.)
verbo/gruppo nominale	verb/pn	(Verbe intransitive suivi par un élément nominal.)
apposizione/sostantivo	apposition/noun	Le premier syncon est une APPOSITION du second syncon.
participio passato/verbo	part. pass/verb	(Le second syncon est la forme de base du premier syncon.)
sostantivo/sostantivo pn	noun/noun pn	Le premier syncon est sujet d'un prédicat nominal avec un nom qui est le second syncon.
omninomen/materiale	omninomen/material	Le premier syncon est réalisé à partir du second syncon.
verbo (prep)/verbo	verb (prep)/verb	Les deux syncons sont deux verbes à la suite, avec ou sans, préposition entre les deux verbes..
verbo+a/sostantivo	verb+to/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "to" entre les deux syncons.
sostantivo+di/sostantivo	noun+of/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "of" entre les deux syncons.
sostantivo+a/sostantivo	noun+to/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "to" entre les deux syncons.
sostantivo+da/sostantivo	noun+from/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "from" entre les deux syncons.

sostantivo+in/sostantivo	noun+in/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "in" entre les deux syncons.
verbo+per/sostantivo	verb+for/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "for" entre les deux syncons.
verbo+da/sostantivo	verb+from/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "from" entre les deux syncons.
sostantivo+per/sostantivo	noun+for/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "for" entre les deux syncons.
sostantivo+con/sostantivo	noun+with/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "with" entre les deux syncons.
sostantivo+su/sopra/sostantivo	noun+on/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "on" entre les deux syncons.
sostantivo+su/sostantivo	noun+on/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "on" entre les deux.
sostantivo+tra/fra/sostantivo	noun+between/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "between" entre les deux syncons.
sostantivo+sotto/sostantivo	noun+under/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "under" entre les deux syncons.
sostantivo+senza/sostantivo	noun+without/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "of" entre les deux.
verbo+senza/sostantivo	verb+without/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "without" entre les deux syncons.
aggettivo+di/sostantivo	adjective+of/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "of" entre les deux syncons.
aggettivo+a/sostantivo	adjective+to/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "to" entre les deux syncons.
aggettivo+da/sostantivo	adjective+from/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "from" entre les deux syncons.
aggettivo+in/sostantivo	adjective+in/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "in" entre les deux syncons.
verbo+di/sostantivo	verb+of/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "of" entre les deux syncons.
verbo+in/sostantivo	verb+in/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "in" entre les deux syncons.
verbo+con/sostantivo	verb+with/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "with" entre les deux syncons.
verbo+su/sopra/sostantivo	verb+on/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "on" entre les deux syncons.
aggettivo+per/sostantivo	adjective+for/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "for" entre les deux syncons.
sostantivo+contro/sostantivo	noun+against/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "against" entre les deux syncons.
sostantivo+per/verbo	noun+for/verb	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "for" entre les deux.
sostantivo+di/verbo	noun+of/verb	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "of" entre les deux.
aggettivo+su/sopra/sostantivo	adjective+on/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "on" entre les deux syncons.
aggettivo+con/sostantivo	adjective+with/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "with" entre les deux syncons.
aggettivo+tra/fra/sostantivo	adjective+between/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "between" entre les deux syncons.
aggettivo+senza/sostantivo	adjective+without/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "without" entre les deux syncons.
verbo+tra/fra/sostantivo	verb+between/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "between" entre les deux syncons.
sostantivo+da/verbo	noun+from/verb	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "from" entre les deux.
sostantivo+a/verbo	noun+to/verb	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "to" entre les deux syncons.
rule sostantivo+di/sostantivo	rule noun+of/noun	N'importe quels syncons des chaînes respectives peuvent être rencontré dans cet ordre avec la préposition "of" entre les deux syncons.
rule sostantivo+a/sostantivo	rule noun+to/noun	N'importe quels syncons des chaînes respectives peuvent être rencontré dans cet ordre avec la préposition "to" entre les deux syncons.
rule sostantivo+da/sostantivo	rule noun+from/noun	N'importe quels syncons des chaînes respectives peuvent être rencontré dans cet ordre avec la préposition "from" entre les deux syncons.
rule sostantivo+in/sostantivo	rule noun+in/noun	N'importe quels syncons des chaînes respectives peuvent être rencontré dans cet ordre avec la préposition "in" entre les deux syncons.
rule sostantivo+con/sostantivo	rule noun+with/noun	N'importe quels syncons des chaînes respectives peuvent être rencontré dans cet ordre avec la préposition "with" entre les deux syncons.
rule sostantivo+su/sopra/sostantivo	rule noun+on/noun	N'importe quels syncons des chaînes respectives peuvent être rencontré dans cet ordre avec la préposition "on" entre les deux syncons.
rule sostantivo+per/sostantivo	rule noun+for/noun	N'importe quels syncons des chaînes respectives peuvent être rencontré dans cet ordre avec la préposition "for" entre les deux syncons.
rule sostantivo+tra/fra/sostantivo	rule noun+between/noun	N'importe quels syncons des chaînes respectives peuvent être rencontré dans cet ordre avec la préposition "between" entre les deux syncons.
rule verbo+di/sostantivo	rule verb+of/noun	N'importe quels syncons des chaînes respectives peuvent être rencontré dans cet ordre avec la préposition "of" entre les deux syncons.

rule verbo+a/sostantivo	rule verb+to/noun	N'importe quels syncons des chaînes respectives peuvent être rencontré dans cet ordre avec la préposition "to" entre les deux syncons.
rule verbo+da/sostantivo	rule verb+from/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "from" entre les deux syncons.
rule verbo+in/sostantivo	rule verb+in/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "in" entre les deux syncons.
rule verbo+con/sostantivo	rule verb+with/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "with" entre les deux syncons.
rule verbo+su/sopra/sostantivo	rule verb+on/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "on" entre les deux syncons.
rule verbo+per/sostantivo	rule verb+for/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "for" entre les deux syncons.
rule verbo+tra/fra/sostantivo	rule verb+between/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "between" entre les deux syncons.
verbo+sotto/sostantivo	verb+under/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "under" entre les deux syncons.
verbo+contro/sostantivo	verb+against/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "against" entre les deux syncons.
sostantivo+tramite/sostantivo	noun+by/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "by" entre les deux syncons.
sostantivo+come/sostantivo	noun+as/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "as" entre les deux syncons.
sostantivo+davanti/sostantivo	noun+before/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "before" entre les deux syncons.
sostantivo+at/sostantivo	noun+at/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "at" entre les deux syncons.
aggettivo+at/sostantivo	adjective+at/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "at" entre les deux syncons.
verbo+at/sostantivo	verb+at/noun	Les deux syncons peuvent être rencontré dans cet ordre avec la préposition "at" entre les deux syncons.
event/syncon	event/syncon	Le premier syncon est un fait/un évènement lié au second syncon.
Les liens en gras sont des chaînes hiérarchiques.		
Les liens en rouge sont les liens les plus importants.		
Les liens avec le fond jaune fonctionnent pour les sentiments.		
Les liens avec le fond bleu sont les liens qui ne sont pas présents dans cette langue.		
"XX" se traduit par n'importe quelle POS.		
"LL" se traduit par lien		

Annexe 2
Liste des registres possibles pour les syncons d'un Sensigrafo monolingue.

Anglais **(termes donnés dans l'interface de travail)**	**Italien** **(termes donnés dans l'interface de travail)**	**Français** **(termes issus de la traduction et l'interprétation personnelle)**
Figurative	Figurato	Figuré
Dialectal	Dialettal	Dialectale
Slang	Gergale	Argot
Informal	Colloquiale	Informel
Foreign	Stranier	Étranger
Humorous	Ironico	Ironique
Hyperbolic meaning	Uso iperbolico	Hyperbole
British English	Solo lingua propria	Anglicisme
Old word	Antico	Vieux
Local word	Regionale	Régional
Vulgar	Volgare	Vulgaire
Familiar	Familiare	Familier
Imported word	Adattamento	Adaptation
Improper	Scorretto	Incorrect
Abbr./Simbolo	Abbr ./Simbolo	Abréviation/Symbole
Literacy	Letterario	Littéraire
Popular	Popolare	Populaire
Latinism	Latinismo	Latinisme
Uncommon	Non comune	Rare
Euphemism	Eufemismo	Euphémisme

Annexe 3
Liste des domaines possibles dans un Sensigrafo monolingue.

Français	Anglais
Abattage	slaughtering
Accessoire	accessory
Acoustique	Acoustics
activités touristiques	touristic facilities
administration publique	public authority
Aéronautique	aeronautics
aéronautique militaire	air force
Aérotechnique	aerotechnics
Agriculture	agriculture
Alpinisme	mountaineering
Ameublement	interior decoration
Anatomie	Anatomy
Anthropologie	anthropology
Antiquité	Antiques
Apnée	Apnea
arboriculture fruitière	fruit farming
Arc	Archery
Archéologie	archaeology
Architecture	architecture
Archivage	Archiving
Arithmétique	arithmetic
Armes	weaponry
Art	Art
Artillerie	Artillery
Artisanat	Craftwork
artisanat d'art	artistic handicraft
arts martiaux	martial arts
Assurance	Insurance
Astrologie	Astrology
Astronautique	astronautics
Astronomie	astronomy
Astrophysique	astrophysics
Athlétisme	Athletics
Automation	automation
Automobilisme	motor racing
Aviation	Aviation
Aviron	Rowing
Balistique	Ballistics
Ballet	Ballet
bande dessinée	Comics
Banque	Bank
base de données	Database
base-ball	Baseball
basket-ball	basketball
Bible	Bible
Bibliothèque	libraries
Bijouterie	jewellery
Billard	billiards
Biochimie	biochemistry
Biologie	biology

biologie marine	marine biology
Biophysique	biophysics
Biotechnologie	biotechnology
Bobsleigh	bobsled
Botanique	botany
Bouddhisme	Buddhism
Bourse	stock exchange
Bowling	bowling
Boxe	boxing
Bricolage	do it yourself
Bridge	bridge
Broderie	lacework
Cad	CAD (computer aided design)
Cardiologie	cardiology
cartes graphiques	graphic's cards
Cartographie	cartography
Céramique	pottery
Chant	singing
Charbonnage	coal mine
Chasse	hunting
Chaussure	footwear
chemin de fer	railway
Chimie	chemistry
chimie inorganique	inorganic chemistry
chimie organique	organic chemistry
Chirurgie	surgery
Christianisme	Christianity
Chronique	news
Cinéma	cinema
Cirque	circus
Collection	collecting
Commerce	commerce
Communication	mass communication
composants électroniques	electronic components
Composition	composition
Comptabilité	accounting
computer art	computer art
Construction	building industry
Cosmétologie	cosmetics
Cosmographie	cosmography
Cricket	cricket
Criminalité	crime
Crochet	crochet work
Cyclisme	cycling
cyclisme sur piste	track racing
Cytologie	cytology
Dactylographie	typewriting
Dames	checkers
Découpage	decoupage
Dermatologie	dermatology
Dessin	drawing
dessin industriel	industrial design
Didactique	didactics
Diététique	dietetics
Diplomatie	diplomacy
Disques	disks
Droit	law
droit civil	civil law
droit commercial	commercial law
droit international	international law

droit pénal	criminal law
droit privé	private law
droit public et administratif	public and administrative law
Échecs	chess
École	school
Écologie	ecology
e-commerce	e-commerce
économie	economics
Écran	monitor
Édition	publishing
électricité	electricity
électronique	electronics
électrotechnique	electrotechnics
Élevage	breeding
e-mail	e-mail
embryologie	embryology
Énergie	energy
enseignement	education
entomologie	entomology
entreprise	company
épigraphie	epigraphy
équitation	equitation
Escrime	fencing
état providence	welfare
Éthique	ethics
ethnologie	ethnology
étude de monnaie	study of coins
événements télévisés	broadcasting events
Évolution	evolution
Finance	finance
finances privées	private finance
finances publiques	public finance
Fisc	revenue
Folklore	folklore
Football	football
football à cinq	five-a-side football
football américain	American football
free climbing	free climbing
Galerie	gallery
gastronomie	gastronomy
généalogie	genealogy
génétique	genetics
Génie	engineering
génie aérospatial	aerospace engineering
géographie	geography
Géologie	geology
géométrie	geometry
géophysique	geophysics
go-kart	go kart
Golf	golf
Graffiti	graffito
grammaire	grammar
graphique	graphic arts
Gravure	engraving
gymnastique	gymnastics
gymnastique rythmique	callisthenics
gymnastique sportive	artistic gymnastics
haltérophilie	heavy athletics
Handball	handball
haute couture	tailoring

héraldique	heraldry
hindouisme	Hinduism
Hippisme	horse racing
Histoire	history
histoire ancienne	ancient history
histoire contemporaine	contemporary history
histoire des temps modernes	modern history
histoire du moyen âge	medieval history
Hockey	hockey
hockey sur glace	ice hockey
Homéopathie	homeopathy
Hôpital	hospital
Horlogerie	watch making
Hydraulique	hydraulics
Hydrographie	hydrography
Ichtyologie	ichthyology
Immunologie	immunology
Impression	printing
Imprimantes	printers
Industrie	industry
industrie aéronautique	aeronautics industry
industrie alimentaire	food industry
industrie automobile	car industry
industrie céramique	ceramic industry
industrie chimique	chemical industry
industrie de chemin de fer	railway industry
industrie de guerre	war industry
industrie de la chaussure	footwear industry
industrie de la cosmétique	cosmetics industry
industrie de motos	motorcycle industry
industrie de verre	glass manufacturing
industrie des meubles	furniture industry
industrie extractive	mining industry
industrie naval	shipping industry
industrie textiles	textile industry
Infographie	graphic
Informatique	computer science
Institutions	institutes
instruments de musique	musical instruments
Internet	internet
Intranet	intranet
Islam	Islam
Jardinage	gardening
Jeu	game
jeu de boules	bowls
jeux de cartes	card games
jeux de hasard	gambling game
jeux de table	board games
jeux d'esprit	enigmatography
jeux vidéo	video games
Jouets	toys
Journalisme	journalism
Journaux	newspapers
Judaïsme	Judaism
Judo	judo
Karaté	karate
Kayak	kayak
Législation	legislation
Linguistique	linguistics
Lithographie	lithography

Littérature	literature
Liturgie	liturgy
Logiciel	software
Loterie	lottery
Luge	sled
Maçonnerie	masonry
Manufacture	manufacturing
Marine	fleet
marine militaire	navy
Marketing	marketing
Matériel informatique	hardware
Mathématiques	mathematics
Mécanique	mechanics
Médecine	medicine
médecine douce	alternative medicine
médecine vétérinaire	veterinary science
Mémoire	memory
Menuiserie	carpentry
Métallurgie	metallurgy
Météorologie	meteorology
Métrique	metrics
Meubles	furniture
Microbiologie	microbiology
Militaire	military
Minéralogie	mineralogy
Mines	mines
Ministères	Ministries
Missile	rocketry
Mode	fashion
Motocyclisme	motorcycling
Multimédia	multimedia
Musées	museums
Musique	music
musique classique	classical music
Mythologie	mythology
Nage	swimming
Nourriture	food
Numismatique	numismatics
Occultisme	occultism
Odontologie	orthodontics
Œnologie	enology
Oncologie	oncology
Opéra	opera
Ophtalmologie	ophthalmology
Optique	optics
Orfèvrerie	jeweller's art
Ornithologie	ornithology
Orthopédie	orthopedics
ouvrage d'aiguilles	needlework
Paléographie	paleography
Paléontologie	paleontology
Paléozoologie	paleozoology
Papyrologie	papyrology
Parachutisme	parachuting
Parapsychologie	parapsychology
Parfumerie	perfumery
Parlement	Parliament
Patinage	skating
Pêche	fishing
pêche sous-marine	underwater fishing

pêche sportive	sport fishing
Pédagogie	pedagogy
Peinture	painting
peinture sur verre	glass painting
Pétrochimie	petrochemistry
Pétrole	oil
Pharmacie	pharmacy
Pharmacologie	pharmacology
Philatélie	stamp collecting
Philologie	philology
Philosophie	philosophy
Phonétique	phonetics
Phonologie	phonology
Photographie	photography
photographie artistique	artistic photography
Physiologie	physiology
Physique	physics
physique atomique	atomic physics
Phytopathologie	phytopathology
ping-pong	table tennis
planche à voile	windsurf
Plongée	scuba diver
Plongeon	diving
Poésie	poetry
Poker	poker
Police	police
Politique	politics
Polo	polo
Pornographie	pornography
Poste	post
Préhistoire	prehistory
Processeurs	processors
Programmation	programming
programme de bureau	business programs
Psychanalyse	psychoanalysis
Psychiatrie	psychiatry
Psychologie	psychology
Publicité	advertising
Radio	radio
Radiologie	radiology
relations publiques	public relations
Religion	worship
Réseaux	nets
Restauration	catering
Rhétorique	rhetoric
Robotique	robotics
Route	highway
Rugby	rugby
Santé	health board
Scénographie	scenography
science pure	exact science
science-fiction	science fiction
sciences sociales	social science
Sculpture	sculpture
Sectes	sects
sécurité sociale	social services
services sociaux	social service
Sexologie	sexology
Sidérurgie	iron and steel industry
Sismologie	seismology

Ski	skiing
ski alpin	downhill skiing
ski de fond	cross-country skiing
ski nautique	water skiing
Snowboard	snowboard
Sociologie	sociology
Softball	softball
Spectacle	show
Sport	sport
sports motorisés	design and engineering
Squash	squash
Statistique	statistics
Sténographie	shorthand
Surf	surf
Sylviculture	silviculture
Syndicats	unions
système d'exploitation	operating systems
Tannerie	tanning
Technicisme	mechanical drawing
Technologie	technology
Télécommunication	telecommunications
Télégraphe	telegraph
Téléphonie	telephony
Télévision	television
Tennis	tennis
termes bureaucratiques	bureaucratic terms
termes culturels	cultural terms
termes scientifiques	scientific terms
termes techniques	technical terms
Théâtre	theatre
Théologie	theology
Thermodynamique	thermodynamics
Thermohydraulique	thermohydraulics
tir à la cible	target-shooting
tir au pigeon	clay-pigeon shooting
Tissus	fabric
Topographie	topography
Tourisme	tourism
Traînage	sleddog
Transports	transports
Travail	work
Trekking	trekking
Tricot	knitting
Université	university
Urbanisme	city planning
ustensiles de cuisine	kitchen utensils
véhicule automobile	motor vehicle
Vente	auction
Verrerie	glassware
Vêtements	clothing
Viticulture	vine-growing
Voile	sailing
Volcanologie	volcanology
volley-ball	volleyball
water-polo	water polo
Zoologie	zoology
Zootechnie	zootechnics

Annexe 4
Liste des attributs de verbe possibles pour les verbes dans le Sensigrafo FR.

Anglais (termes donnés dans l'interface de travail)	**Italien** (termes donnés dans l'interface de travail)	**Français** (termes issus de la traduction et l'interprétation personnelle)
Transitive	Transitivo	Transitif
Intransitive	Intransitivo	Intransitif
Impersonal	Impersonale	Impersonnel
Reflexive	Riflessivo	Réflexif
Copulative	Copulativo	Copule
It take the subj	Regge il cong.	Attribut non utilisé et non explicité.
It take the to + ve	Regge to+infinito	Attribut non utilisé et non explicité.
It take the -	Regge -ing	Attribut non utilisé et non explicité.
Pronominal transitive	Transitivo pron.	Transitif pronominal
Pronominal intransitive	Intransitivo pron.	Intransitif pronominal
Modals/auxiliary	Modale/Servile	Verbe de modalité / Auxiliaire
Reflexive recip	Rifl. Reciproco	Verbe réfléxif réciproque
Predicative sbj	Pred. Del Sogg.	Prédicat du sujet
Predicative obj	Pred. Dell'Ogg.	Prédicat de l'objet

Annexe 5
Liste des catégories d'entités nommées utilisées lors de la tâche 3.

Abbréviations	Types	Descriptions
NPH	Personnes/humains	Cette catégorie prend en compte les personnes par les prénoms, les noms de famille, les noms complets, les surnoms et tous termes similaires. Ces personnes sont réelles ou imaginaires comme Rowan Atkinon est réel et Dora l'exploratrice est imaginaire.
ORG	Organisations / Institutions	Cette catégorie prend en compte les groupes de personnes au sens large. En théorie, la taille minimum d'un groupe est de deux personnes, comme « frères Cohen ». Cela comprend également les organisations et les institutions comme les ministères, les gouvernements etc.
COM	Compagnies / Sociétés / Entreprises	Cette catégorie prend en compte les entreprises contrairement à la catégorie précédente ORG. Par exemple : Siemens est une entité de la catégorie COM contrairement au FBI qui est une entité de type ORG.
MMD	Médias de Masses / Réseaux sociaux	Cette catégorie prend en compte les journaux, les émetteurs ou toutes organisations sources d'informations ayant comme principale tâche de publier des nouvelles. Les réseaux sociaux sont également compris dans cette catégorie.
GEO	Zones Géographiques Administratives	Cette catégorie prend en compte toutes les zones de la planète terre qui représente des ensembles d'humains et sont, d'une certaine manière, reconnues par un gouvernement respectif. Ce sont des zones organisées où sont instaurés des systèmes de taxes et d'élections par exemple. Ce sont également des zones peuplées. Cette catégorie comprend les villes, les pays, les régions, les départements etc.
GEA	Zones Géographiques Naturelles	Cette catégorie prend en compte les zones de la planète terre reconnues. Elles peuvent être des lieux aujourd'hui perdues. Dans cette catégorie se trouve les montagnes, rivières, fleuves, mers, océans, forêts etc…Cette catégorie comprend les lieux qui étaient peuplés et qui aujourd'hui sont abandonnés.
GEX	Zones Géographiques étendues	Cette catégorie prend en compte les entités de type géographique au-delà de la planète Terre et autre lieux géographiques imaginaires comme « Ironforge » qui est la capitale des Nains du clan des Barbe-de-bronze dans le monde du jeux vidéo « World of Warcraft » ou « Vénus » qui est une planète du système solaire de la planète Terre.
PRD	Produits	Cette catégorie prend en compte les entités de types articles, en particulier les articles commercialisés. Ce sont les entités qui ne sont pas comprises dans d'autres catégories en générale. Par exemple un « oscar » (le prix remis lors de la cérémonie du même nom) est un PRD selon nous.
DEV	Appareils	Cette catégorie prend en compte les entités de modèle comme le nom indiquant le modèle de votre smartphone ; par exemple : le modèle « Galaxy S4 mini » de la marque Samsung.
VCL	Véhicules	Cette catégorie prend en compte les entités de modèle d'automobile comme « Clio », « BMW MIII » ou la « Ford Mustang ».
FDD	Nourritures et Brevages	Cette catégorie prend en compte les entités comme les plats, les aliments, les brevages avec un nom propre comme le « Nutella », « Schweppes » ou « Malabar ».
BLD	Batîments	Cette catégorie prend en compte les entités de bâtiment comme « Empire State Building », « Palais de l'Élysée »,
WRK	Ouvrages produits par des intelligences humaines	Cette catégorie prend en compte les entités de type œuvres d'art ou de littérature. Les jeux vidéo sont considérés comme des œuvres du travail de l'intelligence humaine

DOC	Documents textuels	Cette catégorie prend en compte principalement les entités commerciales ou légales comme les traités, les lois, les accords financiers…
EVN	Évènements	Cette catégorie prend en compte les évènements comme les festivals, les rencontres sportives etc…
HOL	Fêtes	Cette catégorie prend en compte les entités relatives aux fêtes, dates ou évènements fériés par exemple comme les fêtes religieuses, les dates de commémorations. Par exemple : « Noël », « fête des couleur », « 14 juillet », « D-Day »…
LEN	Entités légales	Cette catégorie prend en compte les entités légales en particulier les taxes.
PPH	Phénomènes physiques	Cette catégorie prend en compte les phénomènes physiques et naturels comme les catastrophes météorologiques, par exemple : l'ouragan « Katrina » ou les épidémies, par exemple : la « Peste noire ».
ANM	Animaux	Cette catégorie prend en compte les noms d'animaux comme « Cecil » le lion à crinière noir du parc Hwange au Zimbabwe ou « Garfield » le nom du chat de la bande dessinée du même nom.
MEA	Mesures	Cette catégorie prend en compte les mesures comme un nombre de jours, de mois, d'années, ou une quantité comme des kilogrammes, des grammes, des millilitres, des litres ou tout autres unités de mesures.
MON	Monnaies	Cette catégorie prend en compte les entités monétaires comme une somme d'euros, de livres, de dollars ou de yuans par exemple.
DIG	Numéros	Cette catégorie prend en compte les entités numériques qui ne rentrent pas dans d'autres catégories cependant cette catégorie est rarement utilisée et doit être utilisée le moins possible.
PCT	Pourcentages	Cette catégorie prend en compte les pourcentages comme « 42% ».
DAT	Dates	Cette catégorie prend en compte les dates comme le « 4 novembre 2015 »
HOU	Heures	Cette catégorie prend en compte les heures comme « 19h », « midi » etc..
SCO	Scores	Cette catégorie prend en compte les scores dans le cadre du sport. Par exemple « Le score du match France-Angleterre est de 55-0 ». « 55-0 » sera étiqueté SCO.
EMO	Émoticons	Cette catégorie prend en compte les émoticons comme « ^^ », « -_-' », « ;) », « ☺ », « XD ».
FOL	Dossiers	Cette catégorie prend en compte les entités de type dossiers comme un dossier juridique.
WEB	URL	Cette catégorie prend en compte les adresses web du type URL.
MAI	Adresses e-mail	Cette catégorie prend en compte les adresses e-mail.
ADR	Adresses postales	Cette catégorie prend en compte les entités de type adresse postale. Cela dit l'intégralité de l'adresse n'est pas nécessaire. On trouvera, par exemple, « 213 chemin du bois 38000 Grenoble », « 5ème arrondissement de Marseille » ou « 5ème » si une ellipse est effectuée. Il est possible de rencontrer « Nation » dans le sens de la station de métro.
PHO	Numéros de téléphone	Cette catégorie prend en compte les numéros de téléphone.
SSN	Numéros de sécurité sociale	Cette catégorie prend en compte les numéros de sécurité sociale.

Annexe 6
Liste des noms et verbes traité en fréquences.

Verbes

abonder	abouler	abrutir	absenter	abuser	acagnarder	acclimater	accoter
accoutumer	accréditer	accroire	acérer	actionner	actualiser	adhérer	adjurer
adorner	adosser	adouber	adoucir	advenir	affabuler	affaiblir	affaler
affermer	affouiller	affranchir	affrioler	Affûter	agonir	agréer	agripper
aguerrir	ahaner	ahurir	aiguillonner	ajuster	alambiquer	alarmer	alcooliser
allaiter	allécher	alléger	alourdir	Altérer	aluminer	amaigrir	amatir
ambitionner	amender	ameublir	ameuter	amocher	amortir	amplifier	amurer
ancrer	anéantir	animaliser	annuler	ânonner	aplanir	aplatir	apostasier
aposter	appareiller	appesantir	appesantir	applaudir	apurer	arabiser	araser
arbitrer	argenter	arguer	arraisonner	arrérager	arrimer	arsouiller	asperger
assagir	assainir	asservir	assombrir	assortir	atomiser	attiédir	attiger
attitrer	attrouper	ausculter	avaliser	Aviner	avoisiner	bâcher	badigeonner
badiner	bagarrer	bâiller	bâillonner	balbutier	baliser	ballonner	barboter
barguigner	barricader	basaner	bastonner	bâtonner	bauger	baver	béatifier
bécoter	biffer	bifurquer	bigarrer	Bigler	biner	biser	bisser
bivouaquer	blâmer	blanchir	blêmir	blinder	bluffer	bombarder	bonifier
boudiner	bouffonner	boulanger	bouquiner	bourlinguer	bousiller	boxer	brailler
branler	brasser	bredouiller	brider	Briffer	briguer	brillantiner	brinquebaler
brocanter	brocarder	brocher	broncher	brouillasser	brouillonner	broyer	bruiner
busquer	câbler	cabotiner	cabrioler	cafarder	caillebotter	cailler	cajoler
calfeutrer	calibrer	câliner	camper	cancaner	candir	canonner	canoter
cantiner	capoter	capter	capturer	caquer	caracoler	caraméliser	carboniser
caréner	carillonner	cartographier	cartonner	caserner	castrer	catalyser	catapulter
catéchiser	catégoriser	cautériser	censurer	chahuter	chamailler	chamarrer	chambarder
chantonner	chapeauter	chaperonner	charrier	chatoyer	chevaucher	chicoter	chinoiser
chiper	choir	chômer	chorégraphier	choyer	chuinter	cicatriser	cingler
circoncire	cisailler	clabauder	claironner	clapoter	clapper	claustrer	cligner
cliquer	cloisonner	cloner	coasser	cocufier	cogiter	cohabiter	collaborer
collationner	collectiviser	colmater	colporter	coltiner	commercialiser	commissionner	commotionner
commuer	commuter	compasser	compatir	complanter	comploter	composter	comptabiliser
compulser	concasser	concéder	concerter	concilier	confluer	confronter	congratuler
conjuguer	connecter	connoter	conspirer	constiper	consumer	contacter	contingenter
contorsionner	contrebalancer	contredire	contrefaire	contresigner	convoyer	coopter	corroborer
coter	couiner	coulisser	courbaturer	courroucer	courtiser	couturer	crachoter
cramponner	cranter	crapahuter	crawler	crayonner	crétiniser	criailler	croasser
crocher	croquer	croustiller	crypter	cuirasser	cuver	daguer	débagouler
débander	débaptiser	débarbouiller	débarder	débarrer	débaucher	débiner	débiter
déblatérer	débonder	déboucler	débourrer	débrailler	débrancher	débrayer	débrider
débusquer	décalotter	décamper	décaniller	décanter	décaper	décapuchonner	décarcasser
décaver	décentrer	décérébrer	déchristianiser	déclasser	décliner	décloisonner	décoincer
décolérer	décommander	décompresser	déconcerter	décongestionner	déconnecter	déconseiller	décontracter
décorner	décortiquer	découcher	découpler	découronner	décrépiter	décréter	décrier
décrisper	décrotter	déculotter	dédaigner	dédommager	dédorer	dédouaner	défier
défleurir	défoncer	déganter	dégarnir	dégauchir	dégazer	dégeler	dégeuler
dégorger	dégrader	déguerpir	déhancher	délaisser	délester	délibérer	délimiter
démailler	démailloter	démanger	démanteler	démâter	dématérialiser	démener	démettre
démissionner	démolir	démonétiser	démoraliser	déniaiser	dénier	dénouer	densifier
dépanner	déparier	départager	dépecer	dépersonnaliser	dépeupler	déphaser	dépister
dépiter	déplier	déporter	dépoussiérer	déprécier	dérailler	dérider	déroger
dérouter	désactiver	désaltérer	désarçonner	désargenter	désarmer	désavantager	désemplir
désenchanter	désengager	désennuyer	désensibiliser	déshabituer	désincarner	désobéir	désodoriser
désopiler	désorganiser	désosser	dessouder	désunir	déterger	détromper	détrôner
détrousser	dévaler	dévaliser	déverser	dévitaliser	dévoiler	dévouer	dévoyer
diagnostiquer	dialyser	difformer	digérer	Diguer	dilapider	dilater	diligenter
disconvenir	discréditer	disgracier	disqualifier	disséquer	disserter	dissocier	distancier
distordre	diviniser	divulguer	documenter	dodeliner	dogmatiser	doper	dorloter
dramatiser	draper	dropper	drosser	Duper	dynamiter	ébarber	éborgner
ébouler	échancrer	écharper	éclisser	éconduire	écoper	écosser	écourter
écrabouiller	écrémer	écrêter	écussonner	Editer	édulcorer	effeuiller	égratigner

électrocuter	élider	élimer	émacier	émanciper	émasculer	emballer	embarrer
embastiller	emberlificoter	embêter	embosser	emboucher	embourber	embouteiller	embrayer
embrigader	embrocher	embrouiller	émécher	émigrer	émincer	emmailloter	emmancher
emménager	emmurer	émoustiller	empailler	empaqueter	empâter	emperler	empeser
empester	empiffrer	empiler	empoisonner	empourprer	empresser	emprisonner	émulsionner
encaisser	encapsuler	encarter	enchâsser	enchérir	enclencher	enclore	encorder
encorner	endeuiller	endiabler	endiguer	endoctriner	endolorir	enduire	endurcir
enflammer	enfler	enfreindre	englober	engorger	engouer	engouffrer	engourdir
engranger	enguirlander	enkyster	enlacer	enlaidir	enneiger	ennoblir	enorgueillir
enrager	enrayer	enrégimenter	enrhumer	enrubanner	ensemencer	ensevelir	ensoleiller
ensorceler	ensuivre	enténébrer	entériner	entrelarder	envenimer	épancher	épandre
éparpiller	épater	épier	épiler	époumoner	éprendre	érailler	ériger
érotiser	éructer	escarmoucher	escorter	esquinter	essaimer	essorer	essouffler
estampiller	estiver	estomper	estoquer	estourbir	estrapader	estropier	étamer
étêter	étioler	étourdir	étrécir	étrenner	évacuer	évader	évertuer
évincer	exacerber	excentrer	exciper	excommunier	exécrer	exfolier	exhiber
exorciser	expectorer	expérimenter	expier	expliciter	expulser	exsuder	extrapoler
extravaguer	façonner	fainéanter	falsifier	faner	farcir	farfouiller	faucher
faufiler	féminiser	ferrailler	fertiliser	fieffer	filocher	fissurer	flageoler
flairer	flasher	fléchir	flinguer	flirter	fluer	flûter	focaliser
foisonner	fonctionnariser	fortifier	fouailler	foudroyer	fouetter	fourgonner	fragiliser
fraîchir	frayer	frictionner	fringuer	fritter	froncer	fructifier	fulminer
fureter	fusiller	gâcher	gager	galonner	gambader	gargariser	gauler
gausser	gésir	gigoter	gîter	gloser	goberger	goinfrer	gominer
gommer	gouailler	goupiller	gracier	grailler	gravir	grêler	grenouiller
gribouiller	grillager	grincer	grognonner	grossir	grouiller	gruger	guigner
haleter	haranguer	harmoniser	harnacher	helléniser	hennir	hériter	houspiller
hucher	hydrofuger	idéaliser	idolâtrer	immiscer	immortaliser	immuniser	impartir
impatienter	improviser	inactiver	inaugurer	inciser	incomber	incruster	incuber
induire	infantiliser	infatuer	infecter	infester	infirmer	infliger	infuser
ingérer	inhiber	innerver	innocenter	inoculer	inonder	insensibiliser	insinuer
insonoriser	inspecter	instaurer	instiller	insuffler	interligner	internationaliser	interpeller
interposer	intimider	introniser	invertir	invétérer	ironiser	itérer	jacasser
japper	jargonner	jouter	jouxter	juponner	klaxonner	lacer	lainer
lanciner	lanterner	laper	laquer	larder	lésiner	lessiver	libeller
licher	liciter	lifter	limoger	lisérer	lotir	loucher	louper
louvoyer	lubrifier	mâcher	macler	magnifier	malaxer	manager	manigancer
manipuler	maquer	maquiller	marauder	marmotter	maroufler	martyriser	matcher
materner	maximiser	mécontenter	mégoter	mémoriser	mendier	mignoter	millésimer
mimer	minauder	minéraliser	mitonner	mobiliser	moduler	mollir	momifier
moquetter	morfler	morfondre	motiver	mouvementer	muer	mugir	mûrir
muter	mutiler	nationaliser	naturaliser	naviguer	négocier	niaiser	nieller
noircir	normaliser	oblitérer	obnubiler	obséder	obtempérer	occidentaliser	occire
occlure	odorer	offenser	oindre	ondoyer	onduler	ornementer	orthographier
osciller	ossifier	outiller	oxyder	pacifier	pagayer	paillarder	pâlir
palper	palpiter	pâmer	panner	papilloter	parader	paraphraser	parfumer
parier	patrouiller	pauser	pavaner	peinturlurer	pénaliser	percuter	perdurer
perforer	perfuser	péricliter	périr	pérorer	perpétuer	personnaliser	pervertir
pester	péter	pétrifier	pétrir	phagocyter	piaffer	pianoter	picorer
piéter	piétiner	pieuter	piqueter	placarder	plaider	plâtrer	plébisciter
plomber	poêler	poétiser	poignarder	poinçonner	poivrer	polissonner	politiser
polluer	pomponner	pondérer	pontifier	populariser	portraiturer	poudroyer	pouponner
pourprer	praliner	prêcher	préconiser	prédestiner	prédisposer	prélever	prémunir
prévaloir	procréer	prodiguer	profaner	prophétiser	putréfier	pyramider	quémander
questionner	quintessencier	rabibocher	râbler	raboter	raccommoder	raccourcir	racheter
racler	racornir	radoter	radoucir	raffermir	raffiner	rafistoler	rafraîchir
ragaillardir	raidir	rainer	râler	rallier	rallonger	ramager	rameuter
ramollir	ramoner	rancarder	ranimer	rapetasser	rapiner	rapprendre	rasseoir
ratiboiser	ratiociner	rauquer	ravigoter	réanimer	réassortir	rebâtir	rebattre
rebeller	rebiffer	rebiquer	reborder	rebouter	recéper	rechausser	recoiffer
recoller	réconcilier	reconstruire	reconvertir	recracher	récréer	récrier	recroqueviller
rectifier	récurer	redire	redistribuer	redonner	redorer	redoubler	réédifier
rééditer	rééquilibrer	refendre	refondre	refouler	réfracter	réfrigérer	régenter
régionaliser	regrouper	réinsérer	réinstaller	rejaillir	rejouer	relativiser	relaver
remâcher	remarcher	rembourrer	rembourser	rembrunir	remédier	remodeler	remorquer
remplumer	remporter	renâcler	renfler	rengorger	renier	renquiller	réordonner
repenser	repentir	répercuter	repiquer	replacer	réprimer	répugner	resserrer
resservir	retendre	retordre	retracer	retrancher	retravailler	rétrécir	réunifier
revaloir	revendiquer	réverbérer	reverser	revirer	revivifier	révolutionner	révulser
rifler	riposter	roder	rôder	ronéotyper	ronfler	rosir	rôtir
roucouler	rouir	rouscailler	rupiner	saboter	sabouler	saillir	salir
sanctifier	sanctionner	sangler	saper	sarcler	satiner	sauvegarder	savourer

scalper	sceller	scintiller	scotcher	scruter	sécréter	séculariser	sécuriser
segmenter	semoncer	sensibiliser	sextupler	sidérer	singulariser	slalomer	socialiser
solenniser	solidifier	somatiser	sonoriser	sophistiquer	soufrer	souiller	soûler
souquer	sourciller	soustraire	spatialiser	spéculer	squatter	standardiser	stipendier
stopper	stranguler	stresser	submerger	subodorer	subsumer	subtiliser	succomber
sucer	sulfurer	surabonder	suralimenter	surcharger	suréquiper	surestimer	surexciter
surfer	surimposer	surir	surmener	surmultiplier	surtaxer	survolter	suspecter
sustenter	symboliser	sympathiser	synchroniser	systématiser	talonner	tapiner	tapisser
tapoter	tarauder	tarifer	tartiner	tempérer	tétaniser	tiédir	tiquer
tisser	tituber	tomer	tonifier	tonner	toper	torchonner	touer
touiller	tournoyer	tracter	traficoter	traînailler	tranquilliser	transférer	transfuser
transgresser	transhumer	transir	transmuer	transplanter	transposer	transvaser	traquer
traumatiser	trébucher	trémousser	trémuler	tressaillir	tresser	tricoter	trier
triller	trimballer	trimer	tripatouiller	tronquer	truander	truster	turbiner
ululer	uniformiser	universaliser	urbaniser	usurper	vacciner	vaciller	vadrouiller
valser	vanner	vaporiser	vaquer	varloper	ventiler	verbaliser	verdoyer
versifier	vexer	vicier	viner	virevolter	viriliser	viroler	vocaliser
vociférer	voguer	volatiliser	vulgariser				

Noms

abattement	abcès	abdomen	abjection	aboutissement	absinthe	abstinence	abstraction
accablement	accolade	accompagnateur	accompagnement	accomplissement	accouplement	accoutrement	accrochage
acolyte	acquiescement	acquisition	acrobatie	adoration	adversité	aération	affichage
affliction	agrafe	agriculture	aigrette	aiguillon	aine	albâtre	alchimie
allégeance	almanach	alternance	alvéole	amazone	ambigüité	ambiguïté	ambre
aménagement	amont	analogie	analyste	ananas	anarchie	anéantissement	anémone
angine	animosité	annulation	anomalie	anonymat	antéchrist	anticipation	antidote
antipode	appellation	appendice	apport	âpreté	arceau	archange	archet
argot	aristocratie	armateur	armature	arrachement	arrivage	arrosoir	aspérité
avènement	aviron	avorton	babiole	bacchante	bacon	baignade	bajoue
baldaquin	balise	ballade	ballast	ballerine	ban	bandoulière	baraquement
barbiche	barda	baril	baromètre	barrette	bastille	bâtardise	bâtonnet
baudrier	bave	bavure	belette	bercail	berline	biceps	bidoche
bidule	bienséance	bijouterie	billot	biologie	biquet	bison	bistro
blanchisserie	blason	bleuet	blocage	boa	bocage	bolide	bonbonne
bonus	boom	bordée	boucan	bouillonnement	bouledogue	bourgeon	brame
brassard	brasse	brassée	bravoure	breuvage	briefing	brimade	bronzage
broutille	bruine	buffle	bureaucrate	burin	buvette	cabanon	cabas
cabriolet	cactus	caddie	cadenas	caillasse	caillot	calot	calvitie
cambrousse	caméléon	camelot	cancan	candélabre	candeur	canicule	canif
canotier	cantonnement	canular	capteur	carat	carence	carlingue	carne
carrousel	cartel	casaque	casbah	casemate	cataclysme	catacombe	cataracte
cavalcade	cavité	cécité	censeur	cerceau	chahut	chaînette	chaland
chaloupe	chambrée	chancellerie	chandelier	chanvre	charabia	chardon	charognard
charpente	charpie	charrue	charte	châssis	châtaigne	châtaignier	chaudière
chaux	cheminot	chérubin	chevalerie	chevron	chimpanzé	chrome	chroniqueur
chronomètre	circoncision	civilité	clapier	clapotis	claquette	clarinette	classeur
clavecin	clébard	clémence	clergé	clique	clone	cloporte	coccinelle
coche	cogne	cohérence	cohésion	coiffe	collation	collectivité	colloque
colonnade	colt	commandeur	commentateur	commodité	commotion	compas	compère
complainte	complexité	conciliabule	condescendance	condisciple	confessionnal	confins	confiserie
confrontation	congestion	conjonction	conjoncture	constance	consternation	contagion	contamination
conteneur	continuité	contraction	contrainte	contravention	contrebande	contrefort	contretemps
continuité	contraction	contrainte	contravention	contrebande	contrefort	contretemps	controverse
contraction	contrainte	contravention	contrebande	contrefort	contretemps	controverse	conversion
copeau	coqueluche	coran	corbillard	cordelette	cordialité	corneille	cornette
corolle	corporation	couette	couffin	couleuvre	coupon	courbette	courbure
courge	courroux	couscous	crac	crémier	crevasse	crieur	criquet
crispation	crottin	crucifix	crypte	cuirassier	cuite	culasse	cultivateur
cyclope	cyprès	dallage	damier	datte	déboire	débouché	décadence
décalage	déchaînement	déchirement	décomposition	découpage	déficit	déformation	dégradation
délectation	délibération	délinquance	démangeaison	démarreur	démence	démenti	démolition
denier	dénonciation	dépaysement	déplaisir	déportation	dépouillement	dérapage	dérobade
descendance	déséquilibre	désertion	déshonneur	désillusion	destinataire	destroyer	détriment
détritus	devin	diabète	diadème	diaphragme	diffusion	dilemme	discorde
dissimulation	divergence	doctorat	documentation	dogme	dosage	drachme	draperie
drôlerie	drugstore	dynastie	éboulement	échafaud	échancrure	échappatoire	échappée
échappement	écharde	échassier	écheveau	échiquier	échoppe	éclaircie	éclaircissement

écluse	écrasement	écriteau	écrou	écuelle	éden	effacement	effigie
ego	éjaculation	électrophone	élimination	élixir	éloquence	émanation	embarcadère
embranchement	embrun	embûche	emplette	empoisonneur	enchaînement	enchevêtrement	enclos
enclume	encombrement	énervement	engeance	engourdissement	énormité	entaille	entrave
envolée	épaulette	épiderme	épluchure	épopée	épouvantail	équinoxe	équipée
érable	éraflure	érotisme	éruption	escarpin	espionnage	essor	estafette
estuaire	étain	étal	étrangeté	étrave	étrier	étron	euphorie
évêché	exagération	examinateur	excrément	exhibition	exorcisme	expansion	expertise
exportation	extermination	extraction	fanal	fanatisme	fanion	faucille	feinte
fève	fiente	figuration	filature	filière	filigrane	filin	filleul
financement	fixation	fixité	flambée	fléchette	fleuret	fleurette	flirt
fourbi	fourmilier	fournaise	fournée	fourrage	foutoir	frac	frégate
friction	fringale	frise	frivolité	frôlement	frondaison	fronde	frottement
fusain	fuseau	gabardine	gaine	galanterie	gale	galopade	gangrène
garniture	garrot	gaule	gazette	germe	gibet	giclée	gigolo
girouette	glas	glisse	gloussement	glu	goal	godillot	gong
gosier	gouine	goujat	gourbi	gourmette	gourou	gouttelette	gradin
gratin	gravats	gredin	grès	griotte	grive	groom	grosseur
gruyère	guérilla	gui	guimauve	guinguette	hâle	hampe	hanneton
hardiesse	harmonica	harnais	harpon	havre	hébétude	hémisphère	hennissement
hérédité	hérisson	hernie	héron	herse	hippopotame	hobby	hochement
hochet	hotte	idéalisme	illumination	imbécillité	immigration	immondice	implication
importation	imprécation	impresario	imprimeur	inauguration	incision	inclinaison	inclination
incohérence	incompétence	incrédulité	indigestion	indignité	indolence	ineptie	infamie
infidélité	infiltration	infinité	infusion	ingratitude	ingrédient	initiation	injonction
innovation	insinuation	insouciance	intérim	intermède	invective	investigation	iode
islam	jactance	jalon	jérémiade	jésus	jeté	jogging	joute
juriste	kamikaze	kermesse	kleenex	lama	lanceur	lansquenet	lapon
laque	larcin	laverie	lavette	légalité	légation	léthargie	lévrier
lichen	licorne	lie	liesse	limbe	limon	lingot	liquidation
litron	loggia	losange	louche	loueur	lutin	luxure	luzerne
macadam	machination	machinerie	maçonnerie	madeleine	magma	magnolia	mandibule
manette	mangue	manipulateur	manucure	maquignon	marchandage	marelle	marmelade
marmot	martèlement	martinet	martini	mascotte	masseur	massue	matador
matrone	mélopée	ménagerie	méprise	météore	miaulement	mica	micheton
migration	milord	minuterie	mire	mite	mitraille	mitron	mobilisation
mocassin	modération	moissonneur	molécule	monarchie	mondanité	monotonie	monstruosité
montagnard	monteur	montreur	morpion	morse	morve	motivation	mouflet
moulinet	mugissement	multiplication	muse	mutilé	myriade	myrte	nacre
nageoire	naphtaline	narration	nationalisme	navet	nervure	névrose	nickel
noirceur	nonchalance	normalien	notice	notoriété	nougat	nullité	obstruction
octave	officine	ogive	ombrage	ombrelle	omission	opposant	optique
opulence	oracle	orée	orifice	oriflamme	originalité	orme	osselet
ouate	oubliette	ourlet	ourson	outillage	outrance	ovation	paddock
pagode	palabre	palissade	pamplemousse	papeterie	paquetage	paravent	paria
parodie	paroxysme	passoire	pastis	patine	patineur	patriarche	pécule
pelisse	pelure	pénitencier	pépère	pépiement	perceur	percussion	perdrix
périple	perplexité	perquisition	perruche	perturbation	pesée	phobie	pie
pierraille	piété	piéton	pif	pigeonnier	pillard	pincement	pipeau
piqueur	pissotière	pitance	pitre	plaidoirie	plaignant	plain	planteur
planton	pleureur	plomberie	plumage	plumeau	podium	pointillé	poivron
poli	polisson	pollution	polo	ponction	popote	popularité	porcherie
pornographie	postérité	postulant	potin	pouf	poulailler	poulie	poupon
pourtour	praticien	préambule	précepte	prédilection	préface	préjudice	préméditation
présomption	pressing	prestidigitateur	prévenance	prévention	primate	privation	probabilité
procuration	projectile	promontoire	prospection	protéine	prothèse	prototype	prunier
psaume	psychanalyse	psychiatrie	pudding	pulpe	pulsion	purge	qualification
quéquette	questionnaire	quiétude	quolibet	rachat	racket	raclure	radis
raffut	rafraîchissement	raillerie	rainure	rallonge	rambarde	rameur	rancard
rangement	rapace	râpe	rapt	rareté	raseur	rate	râteau
râtelier	réacteur	receveur	rechute	récif	récitation	réclamation	récrimination
recrutement	recueillement	récupération	reddition	rédemption	redingote	regain	réglisse
réincarnation	relâchement	remboursement	réminiscence	renfoncement	renne	renommée	renouvellement
repassage	réprobation	répulsion	requiem	réquisition	réquisitoire	résonance	rétablissement
rétine	retombée	retouche	retourne	retournement	revenant	revendeur	revendication
rhétorique	rhinocéros	ribambelle	ricochet	rieur	robert	rogne	ronron
rossignol	rot	rotation	rotule	roule	rubis	rudesse	rugby
rush	rustre	sabbat	saignée	saladier	sanatorium	sarcophage	sautoir
sauvegarde	savane	sbire	scalpel	scellé	scepticisme	sceptre	scierie
sénéchal	septième	séquelle	serge	serrurier	shoot	silex	silo
singerie	singularité	siphon	sire	skieur	sole	solidité	sommité
somnolence	souffleur	souillure	souk	soulèvement	soupente	soupirail	souteneur

spéculation	sphinx	spray	sprint	stabilité	stèle	stéréo	stigmate
stratège	substitution	subvention	suc	sucrier	sudiste	suivi	suppression
surenchère	survêtement	suspense	suspicion	synthèse	tabatière	tabernacle	taie
tata	té	teinture	télescope	tempo	ténor	terreau	terrine
texture	thérapeute	thorax	timbale	timing	tintamarre	tirelire	torsade
torsion	tortionnaire	tourne	tourniquet	tracas	tragédien	traîtrise	transaction
transistor	transplantation	traquenard	travée	treillis	trépied	treuil	tribut
trille	tringle	trip	tripot	troc	trogne	tromperie	tropique
trouée	truelle	tuberculose	tuerie	tulle	turbulence	turpitude	tutelle
tuyauterie	ultimatum	unisson	urne	vacherie	vadrouille	vagabondage	valoche
valve	vanille	vantail	varech	véhémence	velléité	vénération	vengeur
ventilateur	ventilation	vernissage	versement	vessie	veuvage	vicaire	vidange
vignoble	villégiature	virtuose	viscère	visée	viseur	vogue	voilette
volière	vomissement	vraisemblance	vrille	yoga	zigzag	zoom	zouave

Annexe 7
Liste des mots traités lors de la mise en correspondance.

abat-jour	absolu	adultère	aimant	alentour	ambulancier
antécédent	apéritif	appelé	avant-guerre	avant-poste	avoir
barbare	barbu	belle-fille	biais	bohémien	bossu
bouffon	branleur	breton	brillant	brouillon	brûlé
caille	câlin	canaille	casseur	charcutier	charpentier
chauve-souris	chic	chiffonnier	chouette	cinquième	cocher
commode	communiant	comprimé	concubin	condoléance	conquérant
contre	cornichon	coucher	courtisan	crasse	croisé
croiseur	croque-mort	cube	débile	découvert	dedans
dément	demeuré	déporté	dormeur	drogué	eau-de-vie
éducateur	effectif	élastique	émigré	environ	éthique
étrange	explosif	exposé	extra	extraordinaire	extrême
fainéant	familier	fanatique	fantastique	farceur	faste
fauve	feignant	fétiche	fifi	flash	fleuriste
foutre	galant	gaulois	gourde	guérisseur	hold-up
homicide	homo	hot-dog	inconscient	initié	intermédiaire
interne	intestin	inverse	itinéraire	lointain	loup-garou
maigre	ménager	mercenaire	métis	meuf	muet
music-hall	mutant	mystique	naufragé	observateur	oranger
pique-nique	porte-monnaie	potager	procès-verbal	protecteur	prussien
raciste	rapide	réduit	résistant	rigolo	ringard
routier	scanner	solitaire	sous-bois	sous-lieutenant	souterrain
superbe	terre-plein	tranchant			

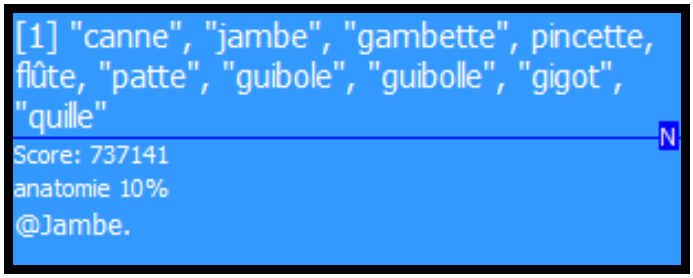
Annexe 8
Liste des phrases utilisées pour le test de comparaison des versions du Sensigrafo FR via le Cogito Desambiguator

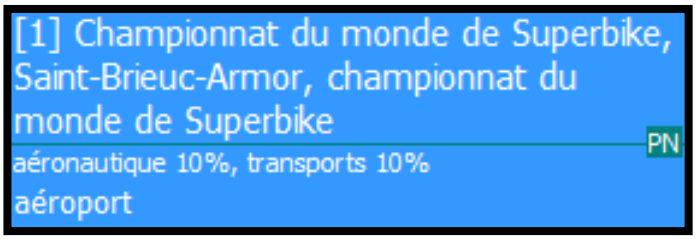
Lemme Sensi	Phrases test
gosier	Mes pieds sont complètement mouillés et une soif intense me dessèche le gosier.
	Le gosier de l'orgue est cassé.
	Le gosier est le siège de la voix, le prolongement du pharynx communiquant avec le larynx.
Palissade	Une palissade est un palis entrelacé de fil barbelé.
	Une palissade est une clôture faite d'une rangée de pieux, de perches ou de planches.
	Une palissade est un mur de verdure formé d'une rangée d'arbres ou d'arbustes spécialement taillés.
Monarchie	La monarchie est un régime dans lequel l'autorité politique réside dans un seul individu, le monarque, et est exercée par lui ou par ses délégués.
	La monarchie est par opposition à la république, un régime politique dans lequel le chef de l'État est un roi héréditaire.
	Une monarchie est un état gouverné par un seul chef, spécialement par un roi héréditaire.
Potin	Elles discutent des derniers potins du village.
	Ils font un potin d'enfer.
	Le potin est un alliage de cuivre, d'étain et de plomb.
Cyprès	Le cyprès est un arbre de la famille des conifères.
	On fait des tables en cyprès qui est le bois de l'arbre.
Gouttelette	Une gouttelette est une petite quantité de liquide.
	Une gouttelette est une petite goutte.
Rogne	Il est en rogne.
	La rogne est une coupe faite au massicot.
	Le sabotier creuse les sabots du cheval avec la rogne.
	La rogne se développe sur le bois et le détériore.
	Il est atteint de la rogne.
Cantonnement	Un cantonnement est l'action de cantonner des troupes.

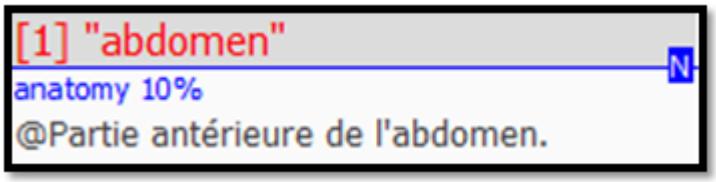
	Le cantonnement de l'esprit est le fait de se cantonner, de se limiter.
	Le cantonnement d'une saisie est la limitation à certains biens du débiteur.
	Un cantonnement est une circonscription forestière placée sous la responsabilité d'un Inspecteur des Eaux et Forêts.
	Un cantonnement est une opération par laquelle le propriétaire d'une forêt grevée d'un droit d'usage abandonne à l'usage la propriété d'une partie de cette forêt.
	Le cantonnement permet d'assurer l'espacement des convois circulant dans le même sens sur une même voie ferrée.
	Un cantonnement est une division en cantons ou sections.
Cultivateur	C'est un cultivateur de pomme de terre.
	Le cultivateur est une machine qui sert à labourer la terre.
	Le cultivateur est un outil à main.
	C'est un cultivateur de l'au-delà.
Vacherie	Il lui a dit des paroles méchantes, des vacheries.
	Je suis dans une situation difficile due à la vacherie de la vie.
	Une vacherie est l'ensemble des vaches d'une exploitation agricole.
Paravent	Elle se déshabille derrière le paravent.
	Au sens figuré, le paravent protège contre les atteintes de l'extérieur.
Bave	Il a de la bave autour de la bouche pendant ses crises d'épilepsie.
	Ses propos méchants et sa bave ne m'atteignent pas.
	Les huitres ont de la bave.
	C'est de la bave, des calomnies.
	Mon chien à plein de bave qui coulent de sa gueule.
Illumination	Une illumination est l'action d'éclairer, de baigner de lumière.
	Une illumination est une inspiration subite, un lumière soudaine qui se fait dans l'esprit.
	Une illumination est l'action d'illuminer occasionnellement par de nombreuses lumières décoratives.
	Une illumination est un lumière extraordinaire que Dieu répand dans l'âme d'un homme.
Rossignol	Les rossignols volent au printemps.
	Un rossignol est un livre sans valeur.
	Il a crocheté la serrure avec ce rossignol.
	Un rossignol est un objet démodé, une marchandise invendable.
	Il joue merveilleusement bien de la flûte et du rossignol.

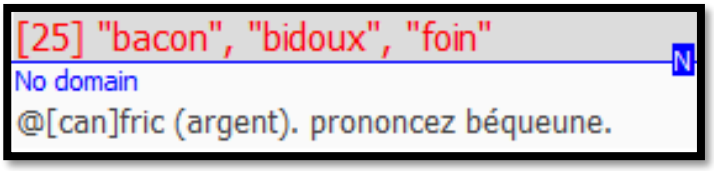
Transistor	Un transistor est un récepteur de radio.
	Le transistor est un composant électronique.
Urne	Il a introduit son bulletin de vote dans l'urne.
	L'urne contenant les cendres de son mari est sur la cheminée.
	L'urne de la mousse se détache à maturité.
	On a fait les statistiques sur l'urne.
	Nos ancêtres puisaient l'eau à l'aide d'une urne.
Sabbat	C'est le jour du sabbat des juifs.
	Les sorcières vont au sabbat.
	Ils font beaucoup de sabbat.
Jésus	Un jésus est une image, une statuette de l'enfant Jésus.
	Un jésus est un terme affectif à l'adresse d'un enfant.
	Un jésus est un gros saucisson.
	Un jésus est un jeune homosexuel qui se prostitue.
	Un jésus est un papier qui portrait en filigrane un monogramme de Jésus.
Décomposition	La gangrène de sa jambe a entraîné une décomposition rapide.
	La décomposition du FN va mener à la mort du partie.
	Le sujet de la décomposition de la lumière par le prisme est abordé dans ce livre.
	La mort de sa femme a été un choc à tel point qu'on put voir la décomposition de son visage en direct.
	La décomposition de son analyse en trois parties est bien faite.

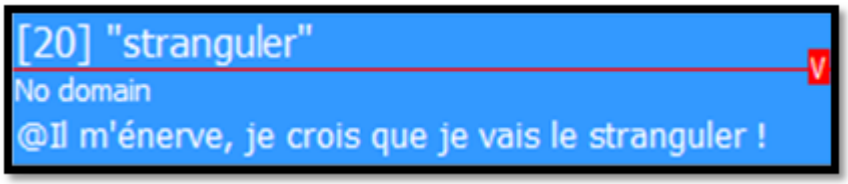
Annexe 9
Captures d'écran d'incongruités rencontrées lors du stage

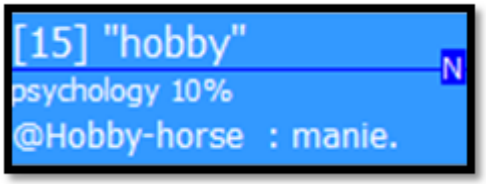












[1] arbre

N

No domain

@Ce qui a l'apparence d'un arbre.

[1] "gauloisement"

Adv

No domain

@D'une manière gauloise, avec une gaieté franche et un peu libre.

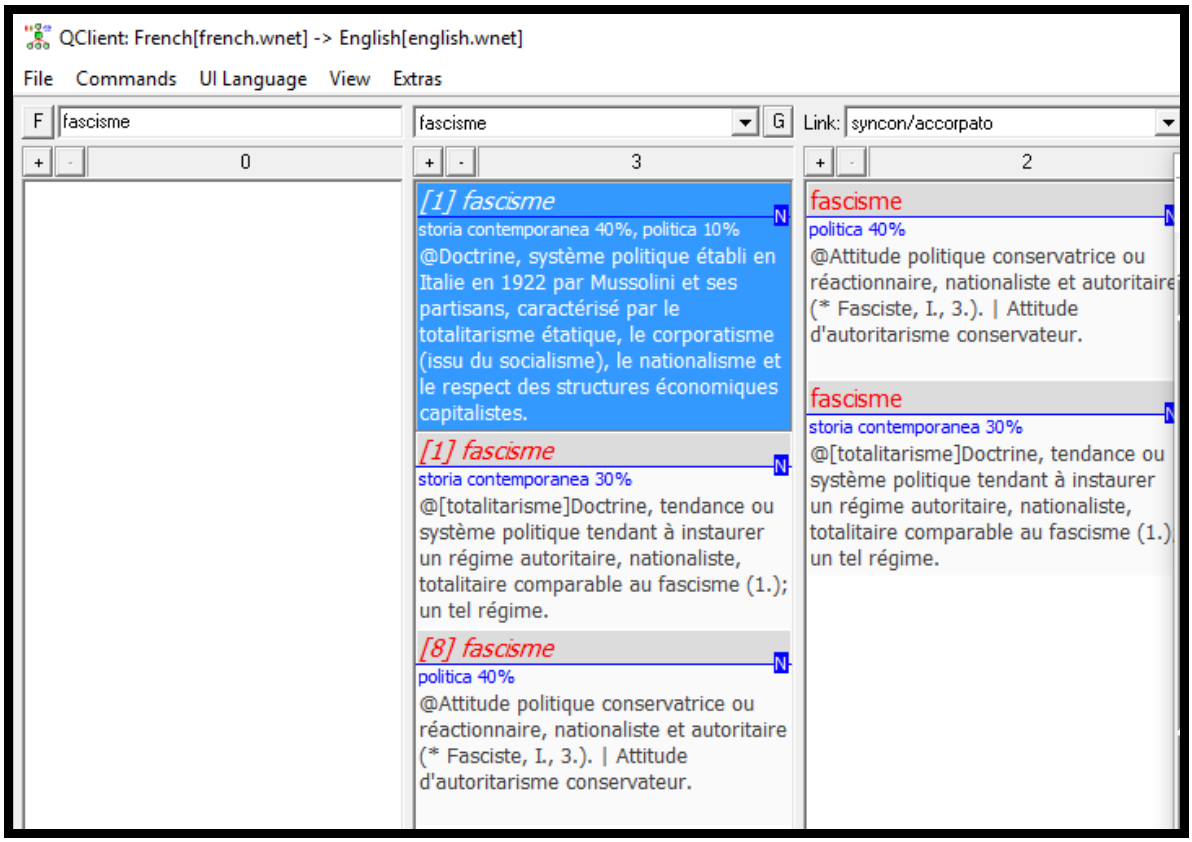


Table des matières

MOTS-CLÉS : Réseau sémantique multilingue, ontologie, désambiguïsation sémantique, entités nommées, traitement du français, Expert System.

RÉSUMÉ

Ce document résume le travail effectué durant 6 mois de stage dans le département de Recherche & Développement de l'entreprise Expert System. Le travail portait principalement sur l'enrichissement de la partie française d'un réseau sémantique multilingue appelé Sensigrafo étendu. L'objectif est d'améliorer la qualité de la désambiguïsation sémantique. Le Sensigrafo FR est la ressource sémantique utilisée par un moteur d'analyse sémantique qui permet d'effectuer différentes désambiguïsations (grammaticales, sémantiques et syntaxiques) sur des contenus textuels. Différentes tâches de tests et d'enrichissements de la ressource sémantique ont été effectuées. Les tâches de test ont pour but d'obtenir une analyse de la qualité des résultats des différentes désambiguïsations. Les observations des résultats de cette analyse ont permis d'établir une typologie d'erreurs à traiter en priorité et de réfléchir aux différentes stratégies de développement possibles.

KEYWORDS : Multilingual semantic web, Ontology, Semantic disambiguation, Entities, French language processing, Expert System.

ABSTRACT

This paper summarizes the work completed during a 6-month internship in the R&D department at Expert System. For the most part this work concerns the enrichment of the French part of a multilingual semantic web known as extended Sensigrafo. The objective of this enrichment is to improve the quality of the semantic disambiguation. The Sensigrafo web is the semantic resource used by the semantic analysis engine which conducts a series of disambiguations (grammatical, semantic and syntactic) on textual content. Both test and enrichment tasks have been performed on the semantic resource.
The test tasks provided us with the means of analysing the quality results of various disambiguations. The analysis of these results allow us to develop an error typology so as to determine those which need to be addressed in priority and to give thought to the various development strategies which could be implemented.