# Repérage automatique d'énoncés définitoires avec Unitex pour l'aide à l'enrichissement de ressources terminologiques : retour d'expérience

Patricia Fener, Claude Dahdouh

INIST - CNRS

# Repérage automatique d'énoncés définitoires avec Unitex pour l'aide à l'enrichissement de ressources terminologiques : retour d'expérience

FENER Patricia_DAHDOUH Claude

# TABLE DES MATIERES

# 1. Problématique

Se fixer comme objectif l'identification puis l'extraction automatique des énoncés définitoires dans des données textuelles présuppose que les actes de langage laissent des traces dans les textes (Rebeyrolle 2000). Par traces langagières, « *textual clues* » (Péry-Woodley 1990), il convient d'entendre « *toute forme linguistique, expression figée, marqueur lexical, connecteur particulier, structure syntaxique ou mot-clef sémantique, marquant l'existence d'une explication et pouvant ensuite jouer un rôle « prédictif » pour cette existence* » (Prince 1994). Prenant la forme de schémas récurrents du langage, ces traces sont le matériau de base qui va permettre la construction de patrons lexico-syntaxiques (PLS) qui seront ensuite utilisés pour le repérage automatique de ces énoncés.

En terminologie, l'activité définitoire fait traditionnellement référence à la « *théorie classique de la définition* » (Brunschwig 1967) ou théorie des conditions nécessaires et suffisantes définie dans les *Topiques* d'Aristote, cinquième livre de l'*Organon*[1], selon laquelle « *une définition est formée de l'ensemble des conditions (traits ou propriétés) individuellement nécessaires pour l'existence de la chose définie et conjointement suffisantes pour la distinguer d'autres choses* » (Seppälä 2012).

L'intérêt porté aux énoncés définitoires vient du fait que c'est en leur sein que s'actualisent les relations sémantiques que les mots entretiennent entre eux. Ils sont « *les plus aptes à décrire le sens des mots et les réalités que ces mots recouvrent* ». De plus, leurs propriétés apparaissent « *suffisamment stables* » pour que leur repérage automatique puisse être envisagé de façon efficace (Rebeyrolle 2000).

Ces informations définitoires, relatives à un domaine d'activité, sont recherchées dans les publications scientifiques qui se caractérisent par le recours à des « *sociolectes* » se distinguant les uns des autres par des constructions phrastiques spécifiques (Tutin 2016). Elles impliquent de traiter des corpus textuels de grande taille, nécessitant l'utilisation d'outils issus du Traitement Automatique des Langues (TAL) qui permettent de révéler des informations morphologiques, syntaxiques et sémantiques sur les unités lexicales composant ces textes.

# 2. Objectif de l'étude

## *2.1. Extraction d'énoncés définitoires*

La finalité de cette étude expérimentale est d'améliorer l'ergonomie du travail terminologique et lexicographique par l'extraction d'information de type énoncé définitoire, au sein de corpus textuels spécialisés.

[1] L'Organon (« outil » ou « instrument » en grec ancien) est le nom scolastique utilisé pour désigner un ensemble de traités, principalement de logique, attribués à Aristote https://fr.wikipedia.org/wiki/Organon

La définition des concepts est un élément fondamental du travail terminologique. « *D'après l'ISO[2] 704[3], les définitions délimitent (ou distinguent), caractérisent ou décrivent le concept* » (Vachez 2021). A l'heure actuelle, cette tâche est encore souvent réalisée à la main ou consiste à interroger des contenus textuels au moyen d'un concordancier afin d'identifier le terme recherché. L'efficacité de cette procédure est loin d'être satisfaisante, ramenant de nombreux contextes sans lien avec une définition. L'automatisation de cette tâche conduirait à gagner en efficacité et en cohérence.

Notre objectif est donc de concevoir et de mettre en œuvre une chaine de traitement, capable de repérer des énoncés définitoires permettant la rédaction de définitions et de tester la faisabilité d'une utilisation dans tout type de contexte terminographique, indépendamment du domaine traité.

Pour atteindre cet objectif, nous lançons une expérimentation avec la suite logicielle Unitex[4] :

- il convient tout d'abord d'analyser la manière dont sont exprimés linguistiquement les énoncés définitoires dans les textes spécialisés puis de construire les patrons lexico-syntaxiques correspondants ;
- il s'agit ensuite de s'appuyer sur cette approche symbolique ou linguistique à base de règles pour concevoir des graphes syntaxiques dans Unitex, capables de reconnaître et d'extraire automatiquement ces segments textuels.

Notre ambition est bien sûr que ces graphes soient suffisamment généralistes pour être applicables, quel que soit le domaine de connaissance.

### *2.2. Enrichissement des ressources terminologiques de la plateforme Loterre*

Le résultat de cette démarche est destiné dans un premier temps au service Ingénierie terminologique de l'Inist-CNRS[5] pour enrichir en définitions les ressources terminologiques exposées sur sa plateforme Loterre[6] (Linked open terminology resources). Les terminologies sont créées par ses ingénieurs spécialistes en IST (Information scientifique et technique) ou par des partenaires, prioritairement issus de l'ESR (Enseignement Supérieur et de la Recherche) ou du monde académique, dans le respect de la Charte[7] définie pour Loterre.

Créée en 2018, Loterre est conçue pour exposer et partager des terminologies scientifiques multidisciplinaires (Sciences de la vie & Santé, Terre & Univers, Ingénierie & Systèmes,

---

[2] L'ISO (Organisation internationale de normalisation) est une organisation internationale non gouvernementale, indépendante, dont les 166 membres sont les organismes nationaux de normalisation. Par ses membres, l'Organisation réunit des experts qui mettent en commun leurs connaissances pour élaborer des Normes internationales d'application volontaire, fondées sur le consensus, pertinentes pour le marché, soutenant l'innovation et apportant des solutions aux enjeux mondiaux.

[3] ISO 704 : 2009 (Travail terminologique — Principes et méthodes) s'appuyant elle-même sur le vocabulaire de la norme ISO 1087 (Travail terminologique et science de la terminologie — Vocabulaire)
https://www.iso.org/obp/ui#iso:std:iso:704:ed-3:v1:fr

[4] https://unitexgramlab.org/fr

[5] https://www.inist.fr/

[6] https://www.loterre.fr/

[7] https://www.loterre.fr/wp-content/uploads/2020/02/Charte-Loterre.pdf

Physique, Chimie, Homme & Société) et multilingues (Français, Anglais, Espagnol). L'accès aux ressources, en consultation, interrogation et téléchargement est ouvert à tous.

S'appuyant sur un triplestore, elle se veut conforme aux standards du web des « *données ouvertes et liées*[8] » (Linked Open Data) et répond également aux principes dits « *FAIR* » (Findable, Accessible, Interoperable, Reusable), publiés par Wilkinson (Wilkinson 2016). Ils définissent les fondements d'un partage de données *« Faciles à trouver, Accessibles, Interopérables, Réutilisables* » qui ont récemment été renommés, suite à la recommandation de la Commission d'enrichissement de la langue française, en données « *Facilement Accessibles, Interopérables et Réutilisables* » (Legifrance 2021). Leur mise en œuvre s'inscrit dans une démarche de « *FAIRification* » (ou « *démarche de FAIRisation* ») (Khayari 2021).

# 3. De quoi parle-t-on ?

Ayant une formation scientifique et souhaitant éviter un repérage de ces traces langagières basé sur l'intuition, nous nous sommes interrogés sur la nature des procédures de définition dans les textes scientifiques.

## *3.1. Typologie des énoncés définitoires*

Selon Pearson (Pearson 1998), il existe trois grandes catégories d'énoncés définitoires : la **définition formelle**, la **définition semi-formelle** et la **définition non formelle**.

La **définition formelle** répond au patron « ***X = Y + C*** », dans lequel ***X*** représente le terme défini, ***Y*** le générique, et ***C***, la caractéristique qui sert à distinguer ***X*** des autres éléments qui ont le même générique.

> *ex : L'acro-ostéolyse autosomique dominante **(X)** est une dysplasie cranio-squelettique **(Y)** très rare, caractérisée principalement par (une petite taille, une acro-ostéolyse des phalanges distales révélée par un pseudo-hippocratisme digital, une ostéoporose généralisée, une dysmorphie cranio-faciale caractéristique et une parodontopathie) **(C)**.*

On peut faire le lien avec la **définition intensionnelle** qui, selon « *ISO 704* », doit indiquer le concept générique (ou superordonné) immédiatement supérieur (exprimé par sa désignation[9]), suivi du ou des caractères distinctifs qui le différencient des autres concepts subordonnés (coordonnés) de même niveau.

La **définition semi-formelle** est ainsi nommée en raison de l'absence d'un élément dans l'énoncé : **Y**, le générique. Elle renferme donc deux informations, ***X*** le terme défini, et ***C*** la caractéristique. Elle répond à la structure « ***X = C*** ».

> *ex : L'hémochromatose 4A **(X)** se caractérise par (une surcharge en fer macrophagique avec fer sérique et saturation de la transferrine normaux ou bas) **(C)**.*

La **définition non formelle**, appelée aussi définition par substitution, se présente souvent sous la forme :

[8] https://fr.wikipedia.org/wiki/Linked_open_data

[9] D'après l'ISO 704, « une *désignation* est une manière succincte de référencer le *concept*, tandis qu'une *définition* doit permettre d'en saisir l'*extension* et de distinguer le *concept* des autres concepts dans le domaine concerné. »

- d'un ***synonyme***

  *ex : maladie des chondromes multiples* ***(maladie d'Ollier)***

- d'une ***paraphrase***

  *ex : Certaines d'entre elles souffrent également de scoliose* ***(déviation de la colonne vertébrale)***.

Il convient d'ajouter la **définition par extension** qui, d'après « *ISO 704* », est utilisée lorsque le nombre de concepts à énumérer est fini, que la ***liste des concepts subordonnés est exhaustive*** selon un critère de subdivision et que les concepts subordonnés sont bien connus ou qu'ils peuvent être explicités par des définitions par intension.

*ex : vascularites des artères de moyen calibre* ***(PAN ou maladie de Kussmaul-Maier, maladie de Buerger et maladie de Kawasaki)***.

Pour Auger (Auger 1997), l'énoncé définitoire est un objet du discours. La relation entre l'objet à définir et sa définition doit par conséquent être explicitée, comme dans l'exemple suivant (Figure 1), par opposition aux énoncés que l'on trouve dans les dictionnaires, dans lesquels l'entrée et la définition ne sont pas unies par une relation explicite. Cette séquence linguistique, mettant en relation d'identification les deux éléments précédents, est appelée **relateur définitoire** (Cartier 2015).

| Terme | Définition du Larousse | Enoncé définitoire |
|---|---|---|
| Hémochromatose | Maladie métabolique consécutive à l'accumulation de fer dans les tissus de l'organisme. | L'hémochromatose comporte une surcharge en fer. |

***Figure 1: Comparaison « Définition vs Enoncé définitoire »***

On peut retenir que l'énoncé définitoire est un discours de forme phrastique qui détermine explicitement, en tout ou en partie, soit les caractères qui appartiennent à un concept, soit les éléments qui caractérisent une chose et qui met en œuvre un vocabulaire particulier (linguistique ou métalinguistique) qui explicite cette détermination.

Un énoncé définitoire est composé de 3 éléments (Labatut 2018) :

- l'objet de la définition, le **definiendum** ou **défini** ;
- une expression définissante, le **definiens** ou **définition** ;
- une **copule** ou un **terme** explicitant *« la nature du rapport de prédication définitionnelle »*.

### *3.2. Modèle définitoire retenu*

Tout comme Meyer (Meyer 2001), nous avons pris le parti de considérer qu'un énoncé est « définitoire » s'il contient au moins un élément susceptible de servir de base à la construction d'une définition lexicographique[10] ou qui donne au moins un élément sémantique propre à construire une telle définition. Comme le souligne Malaisé (Malaisé 2004), cet élément peut

[10]La définition lexicographique énumère et explique tous les sens (significations) du mot à définir. https://tel.archives-ouvertes.fr/tel-01725324/document

être par exemple l'hyperonyme (terme générique) ou une caractéristique permettant la différenciation avec d'autres termes proches dans le domaine.

# 4. Choix du domaine d'expérimentation

Nous avons choisi dans un premier temps de repérer des énoncés définitoires dans le domaine biomédical, en nous concentrant plus spécifiquement sur les maladies.

« La maladie est une altération des fonctions ou de la santé d'un organisme vivant. On parle aussi bien de *la* maladie, se référant à l'ensemble des altérations de santé, que d'*une* maladie, qui désigne alors une entité particulière caractérisée par des causes, des symptômes, une évolution et des possibilités thérapeutiques propres. »[11]

Comme le souligne Charlet (Charlet 2009), on ne peut que constater la dimension linguistique irréductible de la médecine qui s'exprime principalement en langue, avec des termes répertoriés dans des thésaurus[12] ou classifications médicaux.

Nous disposons dans Loterre d'un *« Thésaurus des pathologies humaines »*[13] (Figure 2), constitué de 7 605 termes (5 943 termes préférentiels + 1 662 termes synonymes) français/anglais, issus du vocabulaire de médecine utilisé jusqu'à la fin 2015 pour indexer les références bibliographiques de la base de données PASCAL[14,15] de l'Inist-CNRS.

Ce thésaurus est bien sûr non exhaustif, ne couvrant pas l'ensemble des termes de maladies humaines et pourra être enrichi au fur et à mesure du travail d'annotation, lors du repérage de noms de maladies ne figurant pas dans cette compilation initiale.


***Figure 2 : Extrait du Thésaurus des pathologies humaines***

[11] https://fr.wikipedia.org/wiki/Maladie

[12] L'ISO 25964 définit un thésaurus comme un vocabulaire contrôlé et structuré par une combinaison de relations hiérarchiques et associatives entre concepts, ainsi que des relations d'équivalence intra/inter-linguistiques entre les termes désignant ces concepts dans une même langue (préférentiels et synonymes) ou d'autres langues (traductions).

[13] https://www.loterre.fr/skosmos/VH8/fr/

[14] https://pascal-francis.inist.fr/

[15] Cette base de données bibliographiques en science, technologies et médecine, a été créée en 1971. L'acronyme PASCAL avait été créé avec la signification suivante : Programme Appliqué à la Sélection et à la Compilation Automatique de la Littérature

Nous utilisons les termes de ce thésaurus comme « amorce » pour la recherche de définition. En effet, chaque terme représente une maladie, un syndrome ou un symptôme, constituant ainsi un possible definiendum et, dans certains cas, tout ou partie du definiens.

# 5. Méthodologie

Notre expérimentation s'est déroulée progressivement en 13 étapes qui ont consisté à :

- constituer manuellement un corpus de publications scientifiques du domaine biomédical ;
- repérer et annoter manuellement les énoncés définitoires de maladies dans les documents sélectionnés, afin de constituer notre « corpus de référence » ;
- analyser la manière dont se manifestent linguistiquement les énoncés définitoires de maladies dans les textes (identifier les structures linguistiques qui expriment ces énoncés) ;
- constituer une liste de marqueurs spécifiques de ces énoncés définitoires, permettant leur repérage dans les textes ;
- construire des patrons lexico-syntaxiques à partir de ces marqueurs ;
- prendre en main Unitex, outil de traitement automatique du langage faisant appel à de nombreuses ressources linguistiques et permettant le traitement de gros corpus ;
- intégrer dans les dictionnaires électroniques (DELA)[16] d'Unitex, la liste des maladies figurant dans le « *Thésaurus des pathologies humaines* » ;
- implémenter nos modèles de PLS sous la forme de graphes dans Unitex ;
- évaluer les graphes prenant en entrée le corpus de référence et produisant en sortie les énoncés définitoires identifiés par les graphes respectifs ;
- apporter les modifications nécessaires pour améliorer les résultats de repérage ;
- évaluer la pertinence de ces graphes par un calcul du rappel et de la précision ;
- appliquer les graphes sur un corpus plus volumineux du même domaine pour valider leur réutilisabilité et mettre en évidence de nouveaux marqueurs ;
- appliquer les graphes sur un corpus de domaine différent, à savoir l'astrophysique, afin d'évaluer leur généricité.

# 6. Mise en pratique des différentes étapes

## *6.1. Constitution du corpus textuel*

Nous avons travaillé sur un corpus du registre scientifique, comprenant 48 documents francophones, représentatifs du domaine médical, sélectionnés manuellement, et disposant d'une version en texte intégral car l'expérience nous a montré que c'est essentiellement dans le corps du texte que l'on trouve les énoncés définitoires.

Comme le souligne Bodson (Bodson 2004), on observe moins de variations lexicales et grammaticales en langue de spécialité qu'en langue générale donc il est pertinent de travailler

[16] Les dictionnaires électroniques d'Unitex sont issus de travaux initiés sur le français par Maurice Gross au Laboratoire d'Automatique Documentaire et Linguistique (LADL). Ces travaux ont été étendus à d'autres langues au travers du réseau de laboratoires RELEX.

sur un corpus de petite taille, d'autant plus que le lexique des textes spécialisés est généralement plus dense que celui trouvé dans les textes de langue générale.

Chaque article relève d'un ou de plusieurs auteurs différents des autres afin d'éviter une sur-représentativité de formulations privilégiées par les uns et les autres.

Ces documents sont publiés dans des revues scientifiques issues de BibCNRS[17,18], portail d'accès aux ressources électroniques documentaires du CNRS, ou extraites de Google Scholar, à partir de l'équation de recherche « *(maladie OU pathologie OU *pathie) ET homme* ». Ayant pris le parti de nous intéresser uniquement aux énoncés définitoires de maladies, la sélection des textes a été conditionnée par la présence d'un nom de maladie figurant dans le titre et étant l'objet principal de l'étude, afin d'augmenter la probabilité de trouver des définitions s'y rapportant. Les articles dont les titres portaient sur des traitements ou des techniques d'exploration ont été rejetés.

Téléchargés en format PDF, les documents ont été convertis au format txt et encodés en Unicode UTF-8 pour être chargés dans l'outil Unitex en vue de leur traitement linguistique, avec un total de 621 422 tokens.

### *6.2. Annotation manuelle des énoncés définitoires de maladies*

L'objectif de cette annotation manuelle est de rendre compte, à travers le regard d'experts du domaine, de ce qu'est un énoncé définitoire.

L'intérêt de cette phase d'annotation sémantique est double, permettant à la fois de créer un matériau pertinent pour la description linguistique de l'énoncé définitoire mais aussi de repérer les marqueurs sur lesquels s'appuiera la construction des graphes syntaxiques dans Unitex.

Même en ayant connaissance de la structure d'un énoncé définitoire, il n'en reste pas moins que l'annotation manuelle représente une activité intellectuelle caractérisée par une certaine subjectivité. Celle-ci est en rapport avec une notion intuitive de pertinence des traits qui sous-tendent un contexte définitoire (Seppälä 2010). De plus, cette tâche est sujette à variation d'un individu à l'autre et parfois aussi pour un seul et même individu lorsqu'il est amené à reconsidérer une annotation effectuée auparavant.

Fastidieuse, cette étape a nécessité un travail de longue haleine, avec cependant un impératif constant de qualité.

L'annotation des énoncés définitoires a été réalisée dans NotePad++[19] par un seul annotateur, spécialiste du domaine médical. Un total de 411 énoncés définitoires a été repéré et la

[17] https://bib.cnrs.fr/

[18] BibCNRS est le portail multidisciplinaire, développé et géré par l'Inist-CNRS. Il met à disposition des unités et des chercheurs du CNRS un ensemble de revues et de livres électronique, ainsi que des bases de données, négociés soit pour l'ensemble des chercheurs soit pour des communautés scientifiques définies. Multidisciplinaire, il propose dix espaces thématiques couvrant chacun des dix instituts scientifiques du CNRS.

[19] [19]https://fr.wikipedia.org/wiki/Notepad%2B%2B

caractérisation des segments occurrences a été réalisée avec les balises ouvrante <TAG> et fermante </TAG>.

Lors de cette étape, nous avons mis en évidence des énoncés définitoires de maladies dont les noms ne figuraient pas dans le « *Thésaurus des pathologies humaines* » et qui ont donc été ajoutés à la liste initiale.

### *6.3. Analyse des structures linguistiques exprimant ces énoncés définitoires*

L'énoncé définitoire s'inscrit dans un vaste ensemble de procédés linguistiques qui apportent des éléments d'information sur un objet défini. L'analyse de ces procédés linguistiques nous a permis d'identifier deux types de construction phraséologique d'énoncés définitoires.

**Soit il existe une relation directe entre la maladie et sa définition, établie par un relateur de type verbal ou par un indice typographique, en l'occurrence la parenthèse.**

- ***Construction dans laquelle la relation entre la maladie et sa définition est établie par un relateur verbal.***

Dans ce type de construction, le verbe représente alors le pivot de la phrase. Il s'accorde en genre et en nombre avec le definiendum qui est son sujet et est suivi du definiens qui peut être un complément d'objet (direct ou indirect), un attribut du sujet (derrière le verbe être) ou une proposition subordonnée.

Nous retenons également les contextes dans lesquels le terme exprimant la maladie est repris sous forme anaphorique.

> *ex1 : La **maladie de Behçet** est classée parmi les vascularites de l'enfant. **Elle**_comporte une atteinte vasculaire…*

> *ex2 : **Hémochromatose de type 3**. **Elle** est due à une mutation du gène du récepteur de la transferrine de type 2 (chromosome 7). **C'**est une forme rare qui ressemble à l'hémochromatose de type 1.*

Le marqueur verbal contribue à exprimer un grand nombre de fonctions rhétoriques et discursives, comme le montrent les exemples suivants **:**

✓ Une fonction de **classification**

> *ex : La fièvre typhoïde est une bactériémie à point de départ digestif.*

✓ Une fonction de **définition**

> *ex : L'autisme se définit par l'association, débutant avant 3 ans, d'une déficience qualitative des interactions sociales réciproques, d'une déficience de la communication verbale et non verbale, et d'une déficience des activités imaginatives.*

✓ Une fonction d'**inclusion**

> *ex : L'hémochromatose 4A comporte une surcharge en fer macrophagique avec un fer sérique et un coefficient de saturation de la transferrine normaux ou bas.*

✓ Une fonction de **désignation**

*ex : Le terme de chondrocalcinose articulaire (CCA) désigne stricto sensu une calcification des cartilages articulaires.*

- ***Construction dans laquelle la relation entre la maladie et l'expression définissante est établie par un indice typographique de type parenthétique.***

Les parenthèses sont utilisées pour faire un ajout au discours principal. Ces signes d'énonciation qui appartiennent aux organisateurs para-linguistiques de la ponctuation (Dambreville 1995) sont « *le lieu du dédoublement de la voix de l'énonciateur, l'endroit où ce dernier décide de créer un espace graphique pour « ajouter par ailleurs » une information* quelconque » (François 2011). qui peut être de type définitoire

Védénina (Védénina 1989) range les parenthèses, au même titre que les tirets et les guillemets, dans les signes de ponctuation faisant partie de la syntaxe communicative et exprimant la valeur informationnelle d'une unité textuelle.

Il ressort des nombreux écrits portant sur la définition des parenthèses, la notion d'importance relative de l'unité textuelle marquée :

- ✓ « *La parenthèse permet d'introduire une explication indispensable ou non à la compréhension de la phrase, mais qu'on désire isoler pour ne pas rompre la continuité du texte* ». (Drillon 1991)
- ✓ « *Les parenthèses servent à intercaler dans une phrase une indication, une précision accessoire* ». (Imprimerie nationale 2002)
- ✓ « *Les parenthèses marquent l'insertion dans la phrase d'un segment accessoire et formant un sens à part* ». (Védénina 1989)
- ✓ « *Les parenthèses sont des signes spécifiques de l'écrit qui servent à introduire dans le discours une phrase ou une portion de phrase, qu'elles rendent incidentes par rapport au reste de l'énoncé, qu'elles portent d'emblée sur un autre plan* ». (Pasques 1980)

La parenthèse apparaît ainsi comme un outil privilégié dans la mise en relief d'informations relatives au definiendum.

Dans notre corpus, seules les parenthèses ont été utilisées comme signe typographique par les auteurs pour formuler des définitions.

*ex1 : vascularites non nécrosantes des gros vaisseaux* ***(artérite de Takayasu et artérite à cellules géantes de Horton)****…*

*ex2 : insuffisance rénale* ***(urée > 400 mg/l ou créatininémie > 15 mg/l)****…*

**Soit c'est un agent extérieur qui érige une définition de la maladie.**

Dans cette construction on trouve le verbe **définir** qui a pour sujet un agent extérieur, auteur de la définition. La maladie à définir est dans ce cas complément d'objet direct (COD) du verbe **définir**.

L'agent extérieur peut être représenté par :

- un **organisme**

*ex : **NCEP-ATPIII** définissent le syndrome métabolique comme l'association de 3 critères parmi les 5 suivants : répartition androïde des graisses, hypertension systolo-diastolique, hyperglycémie modérée à jeun, hypertriglycéridémie, HDL-cholestérol bas.*

- une **personne**

*ex : C'est en 1943 que **Lichtenstein** et **Jaffe** ont défini le chondrosarcome. C'est un sarcome dont les cellules tumorales sont associées à une matrice cartilagineuse.*

- un **collectif plus diffus « on »**

*ex : **On** définit la maladie de Gaucher comme une « maladie de surcharge lysosomale».*

- un **référentiel**

*ex : Le **DSM-IV** définit les troubles du contrôle des impulsions non spécifiés comme l'incapacité à résister au besoin d'accomplir ce que l'on sait néfaste pour soi-même ou pour les autres, avec actes répétés, incorrigibles, sans motivation claire.*

## *6.4. Constitution et analyse d'une liste de marqueurs à intérêt définitoire*

A l'intérieur de ces constructions linguistiques, on peut mettre en évidence des mots (verbes, noms), des signes de ponctuation, des expressions qui sont des marqueurs d'un énoncé définitoire.

### 6.4.1. Classification des marqueurs en fonction du critère métalinguistique

Face à la diversité des énoncés définitoires, Auger (Auger 1997) choisit d'appliquer le critère métalinguistique afin de les départager en deux groupes principaux (Figure3) : les énoncés définitoires linguistiques et métalinguistiques.

Les énoncés définitoires métalinguistiques, également appelés énoncés définitoires directs, mettent en avant la fonction métalinguistique du langage et l'affichent en utilisant des verbes métalinguistiques (*appeler, baptiser, définir comme, dénommer, dénoter, désigner, nommer, signifier, vouloir dire*) (Malaisé 2004) qui explicitent la prédication définitionnelle entre la séquence définissante et le terme à définir. Le terme à définir y est toujours en usage autonymique[20]. De façon générale, on peut considérer que le métalangage (Lyons 1980) implique le discours qu'on utilise pour commenter ce que l'on a déjà dit. Il reflète souvent une réflexion cognitive de l'énonciateur sur la formulation linguistique ou le contenu déjà énoncé (Ji 2019).

Les énoncés définitoires linguistiques sont référentiels, ils ne comportent pas de mot métalinguistique, ni de sujet autonyme. Ils sont classés en 6 catégories : copulatifs, d'équivalence, de caractérisation, d'analyse, de fonction, de causalité.

Les énoncés définitoires métalinguistiques quant à eux peuvent être de désignation, de dénomination ou systémiques. (Auger 1997)

[20] Selon Bore (Bore 2009), l'autonymie est définie comme « la propriété du langage à parler de lui-même, à parler des mots qu'il utilise, à se citer lui-même, et le mot autonyme comme « l'homonyme du mot qu'il désigne ; ainsi /table /(= le mot »table ») n'est pas table (= le meuble) ».

De plus, comme le montre la typologie de base des énoncés d'intérêt définitoire (Figure 3), chacun de ces deux groupes possède ses marqueurs spécifiques.

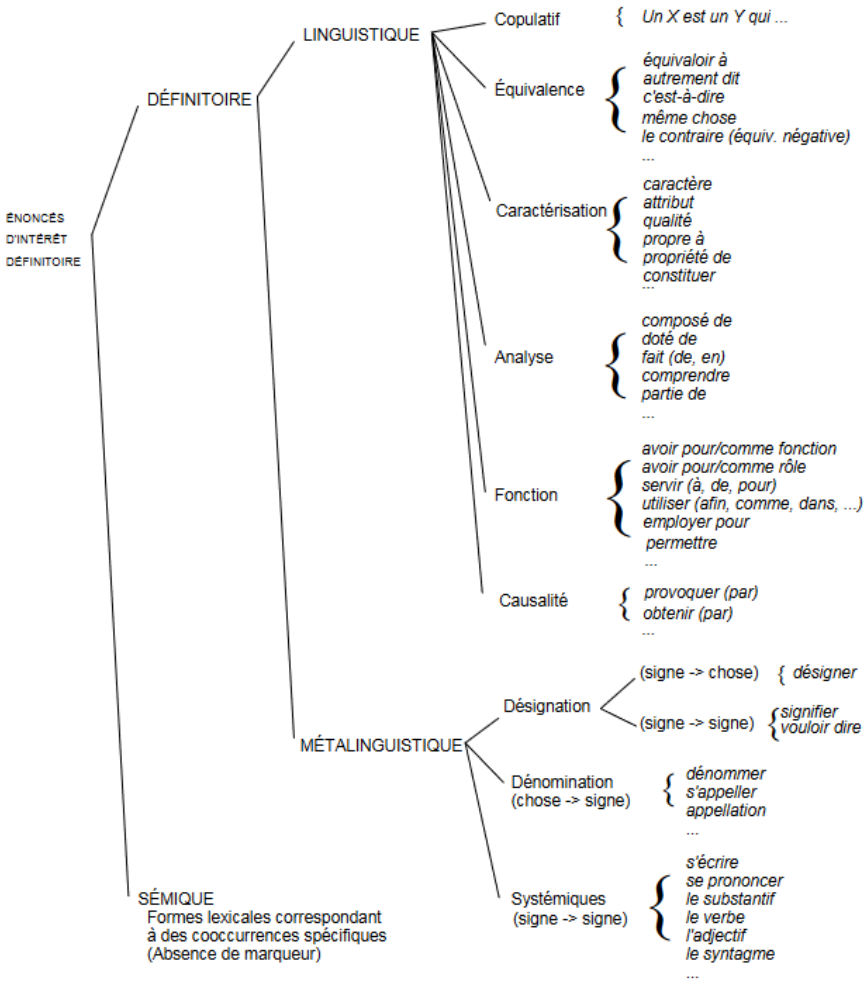



***Figure 3 : Typologie de base des énoncés d'intérêt définitoire***

Nous avons tenté de classer les marqueurs identifiés dans le corpus selon cette typologie.

#### *6.4.1.1. Marqueurs des énoncés définitoires linguistiques*

**Les marqueurs de l'analyse :**

Ce sont des verbes dont le sens a trait aux éléments constitutifs ou aux propriétés d'un objet :

**Comporter :** marqueur entrant dans la composition d'énoncés qui explicitent les éléments constitutifs d'un objet. Il est toujours conjugué au présent de l'indicatif, à la 3e personne du singulier ou du pluriel.

> *ex : La maladie de Behçet est classée parmi les vascularites de l'enfant. Elle **comporte** une atteinte vasculaire, mais toutes ses manifestations ne sont pas dues à une vascularite et les aspects histologiques sont variables selon les sites atteints.*

**Etre composé de :** marqueur présent dans le corpus, à la voix passive, au présent de l'indicatif.

*ex. : Le carcinome mucoépidermoïde bronchique **est composé de** 3 types cellulaires : des cellules mucineuses secrétant du mucus ; des cellules malpighiennes, et des cellules de types intermédiaires.*

**Associer :** marqueur utilisé au présent de l'indicatif.

*ex : Vascularite urticarienne hypocomplémentémique. Elle **associe** une urticaire, une vascularite leucocytoclasique et dans 50 % des cas un abaissement du taux sérique de C3.*

**Recouvrir :** marqueur conjugué au présent de l'indicatif.

*ex : Le terme vascularite d'hypersensibilité n'est plus utilisé aujourd'hui. Il **recouvre** les vascularites des petits vaisseaux, d'origine probablement allergique et présentant histologiquement de la leucocytoclasie vasculaire.*

**Consister en :** marqueur conjugué au présent de l'indicatif.

*ex : L'acro-ostéolyse longitudinale **consiste en** une ostéolyse de la houppe et de la diaphyse de façon concentrique, donnant un aspect en « sucre d'orge ».*

**Correspondre à :** marqueur utilisé au présent de l'indicatif.

*ex : Le syndrome d'activation macrophagique (SAM) peut être secondaire à de nombreuses infections, notamment virales. Il **correspond à** une atteinte inflammatoire sévère et peut engager le pronostic vital des patients.*

### Les marqueurs de la caractérisation :

Il s'agit de verbes mettant en évidence les caractéristiques d'une maladie ou encore certaines relations qui existent entre le défini et le définissant.

**Se caractériser par** : marqueur utilisé à la forme pronominale, au présent de l'indicatif.

*ex : Le syndrome d'Asperger **se caractérise par** un sex-ratio de neuf garçons pour une fille, un quotient intellectuel au moins subnormal, un langage oral préservé mais d'expression maladroite (style pédant, prosodie inadaptée, hermétisme vis-à-vis des jeux de mots et figures de style), …*

**Constituer :** marqueur utilisé au présent de l'indicatif.

*ex : La rétinochoroïdite toxoplasmique **constitue** la forme la plus fréquente d'uvéite d'origine infectieuse.*

**Représenter :** marqueur conjugué au présent de l'indicatif.

*ex : Les anémies hémolytiques auto-immunes (AHAI) à AAc chaud **représentent** la majeure partie des AHAI de l'enfant.*

**Etre défini par :** marqueur utilisé dans le corpus à la voix passive, au présent de l'indicatif.

*ex : L'uvéite postérieure **est définie par** une inflammation de la rétine et/ou de la choroïde.*

**Se distinguer de :** marqueur utilisé à la forme pronominale, au présent de l'indicatif.

*ex : L'artérite de Takayasu est une vascularite granulomateuse caractérisée par l'atteinte élective des gros vaisseaux, qui diffuse par continuité à partir de l'aorte thoracique. Elle **se distingue de** l'artérite temporale à cellules géantes du sujet plus âgé.*

**Appartenir à/au** : marqueur utilisé au présent de l'indicatif.

*ex : La maladie de Gorham-Stout **appartient au** groupe des malformations lymphatiques.*

**Se manifester par :** marqueur utilisé à la forme pronominale, au présent de l'indicatif.

*ex : La déficience qualitative des interactions sociales réciproques **se manifeste par** : une absence ou des anomalies majeures de toute modalité de communication non verbale : expressivité du regard ou du visage, mimique gestuelle, postures corporelles.*

**Se présenter (sous la forme, comme) :** marqueur utilisé à la forme pronominale, au présent de l'indicatif.

*ex : La méningite tuberculeuse **se présente sous la forme d'**une méningite d'installation plutôt progressive sur quelques semaines, isolée, ou associée à des signes neurologiques focaux, avec en particulier une atteinte évocatrice de la base du crâne.*

**S'exprimer par :** marqueur utilisé à la forme pronominale, au présent de l'indicatif.

*ex : La déficience de la communication verbale et non verbale **s'exprime par** : une déficience de langage oral, tant sur le versant expressif que réceptif ; une utilisation apragmatique, en cas de langage oral élaboré, empêchant sa fonction sociale intégrative : stéréotypies verbales, écholalie.*

**S'individualiser par :** marqueur utilisé à la forme pronominale, au présent de l'indicatif.

*ex : L'autisme dit « de haut niveau » **s'individualise par** les éléments suivants : sex-ratio de six garçons pour une fille, quotient intellectuel supérieur à 40, performances remarquables,…*

**Se traduire par :** marqueur utilisé à la forme pronominale, au présent de l'indicatif.

*ex : La maladie de Gorham-Stout appartient au groupe des malformations lymphatiques et **se traduit par** une prolifération de cellules lymphatiques sous l'influence de facteurs de croissance (VEGF) conduisant à un envahissement progressif de l'os.*

### Les marqueurs de l'équivalence :

**Il s'agit :**

*ex : L'anémie hémolytique auto-immune (AHAI) est l'une des premières pathologies ayant été reconnue comme maladie autoimmune. **Il s'agit** d'une destruction des globules rouges par des autoanticorps (AAc).*

### Le marqueur de l'énoncé définitoire copulatif :

**Etre :**

La copule **être** exprime une prédication définitionnelle qui établit l'identité des deux types de référents, à savoir le nom exprimant le definiendum et le nom exprimant le definiens.

*ex1 : Le chondrosarcome **est** une tumeur osseuse maligne.*

*ex2 : Maladie de Buerger. Appelée aussi thrombo-angéite oblitérante, **c'est** une vascularite des hommes jeunes, tabagiques, touchant principalement les artères et les veines de moyen et petit calibre des quatre membres, exceptionnellement les vaisseaux cérébraux et viscéraux.*

L'observation des variations morphologiques des verbes montre que la conjugaison prédominante est représentée par l'indicatif présent, à la 3$^{e}$ personne du singulier ou du

pluriel, selon que nous sommes en présence d'un sujet correspondant à une ou plusieurs maladie(s).

#### *6.4.1.2. Marqueurs des énoncés définitoires métalinguistiques*

Nous avons identifié des marqueurs de dénomination et de désignation.

« *Dénommer, c'est donner le nom d'une catégorie, catégoriser un référent en l'insérant dans une classe d'objets identifiés dans le lexique. Tandis que désigner renvoie à un élément du monde dont on désigne un aspect contingent* ».(Siblot 2001)

**Les marqueurs de la dénomination :**

**Appeler :** marqueur conjugué au présent de l'indicatif.

> *ex : Parfois, le diabète se développe pendant la grossesse et on l'**appelle** alors diabète gestationnel.*

**Dénommer** : marqueur conjugué au présent de l'indicatif.

> *ex : Pitotti (1932) **dénomme** « acrémoniose cutanée » des gommes des mains dont il isole « Cephalosporium acremonium Corda (Pollaci) ».*

**Nommer :** marqueur apparaissant dans les textes à la voix active et la voix passive.

- ✓ A la voix active, conjugué au présent de l'indicatif.

> *ex : …Weiss et al. décrivent le cas d'un patient présentant un syndrome hémorragique sévère et provoqué en rapport avec un défaut d'activité procoagulante plaquettaire qu'ils **nomment** syndrome de Scott…*

- ✓ A la voix passive, conjugué au passé composé de l'indicatif.

> *ex : Des manifestations neurologiques moins graves sont également décrites ; il s'agit de mouvements anormaux, chorée, acathésie, tremblements de type parkinsoniens, constituant ce qui **a été nommé** le « crack dancing ».*

**Les marqueurs de la désignation** :

**Désigner :** marqueur conjugué au présent de l'indicatif.

> *ex : Le terme de chondrocalcinose articulaire (CCA) **désigne** stricto sensu une calcification des cartilages articulaires.*

### 6.4.2. Classification des marqueurs en fonction du type de relation entre le definiendum et le definiens

Comme décrit précédemment, nous avons identifié deux types de relation entre ces différents composants de l'énoncé définitoire, soit une relation directe, soit une relation établie par un agent extérieur. Chacune de ces constructions a des marqueurs d'énoncés définitoires qui lui sont propres.

#### *6.4.2.1. Construction dans laquelle il y a une relation directe entre la maladie et sa définition*

Dans ce type de construction, on trouve les marqueurs verbaux listés précédemment et un marqueur de ponctuation, la parenthèse.

### Les marqueurs « *verbaux* » :

L'analyse syntaxique montre que certains verbes sont suivis d'un déterminant défini et/ou indéfini, d'autres sont utilisés avec une préposition.

- **Verbe + déterminant :** être (un(e), des, la, le, l', les) ; associer (un(e), des, la, le, l', les) ; comporter (un(e), des) ; caractériser (un(e), des) ; désigne un(e) ; présenter (un(e), des) ; représenter (un(e), des) ; constituer (un(e) le, la) ; recouvrir (un(e), des, la, le, les).
- **Verbe + préposition :** consister en ; correspondre à ; appartenir à ; se composer de ; il s'agit de ; se distinguer de ; se caractériser par ; s'exprimer par ; s'individualiser par ; se traduire par ; s'associer à ; se manifester par ; se présenter (comme, sous la forme de) ; se définir (par, comme).

Si l'on analyse cette liste de marqueurs verbaux, on trouve notamment :

- la copule **être** principal marqueur de l'activité définitoire qui possède une valeur métalinguistique, mais de très faible densité. « Être » constitue la forme de base de la prédication d'identité, puisqu'il permet la mise en rapport entre les signes et les choses : il projette « *sur l'axe syntagmatique (le discours) les éléments substituables qui sont dans l'axe paradigmatique (la langue)* » (Rey-Debove 1978) (Husson 2020).
  La construction des phrases dans lesquelles il apparait est plus simple que celles des autres marqueurs « verbaux », donc nous isolons « être ».

- le verbe **définir**, avec des traits de conjugaison spécifiques :

  ✓ Au présent de l'indicatif (3e personne du singulier ou du pluriel) à la voix passive :

  *ex : La fibromyalgie (FM)* ***est définie par*** *un syndrome polyalgique chronique, d'origine non inflammatoire, associé à la présence de points douloureux reproduits à la palpation.*

  ✓ Au présent de l'indicatif (3e personne du singulier ou du pluriel), à la forme pronominale :

  *ex : L'autisme* ***se définit par*** *l'association, débutant avant 3 ans, d'une déficience qualitative des interactions sociales réciproques, d'une déficience de la communication verbale et non verbale, et d'une déficience des activités imaginatives.*

  ✓ Au participe passé :

  *ex : une hypertension artérielle* ***: définie par*** *une pression systolique >140 mmHg et/ou une pression diastolique > 90 mmHg, ou un traitement antihypertenseur.*

- les autres verbes listés précédemment, conjugués au présent de l'indicatif, 3$^{e}$ personne du singulier ou du pluriel.

### Le marqueur de ponctuation, « *la parenthèse* » :

Le contenu des parenthèses revêt des formes linguistiques de nature différente, à savoir :

- Un **syntagme nominal**

  *ex : cécités légales* ***(acuité visuelle du meilleur oeil inférieure ou égale à 1/10)***

- Un **nom de maladie unique**

*ex : maladie des chondromes multiples* ***(maladie d'Ollier)***

- Une **suite de noms ou de syntagmes nominaux séparés par des virgules**

*ex : tableau d'hémolyse intravasculaire* ***(fièvre, lombalgie, urine rouge « porto » et parfois une insuffisance rénale)***

- Une **suite de noms de maladies séparés par des virgules**

*ex : déficits gonadotropes syndromiques* ***(Prader-Willi, BardetBiedl, Charge)***

#### *6.4.2.2. Construction dans laquelle la définition de la maladie est établie par un agent extérieur*

Deux types de marqueurs ont été mis en évidence dans cette construction, à savoir :

**Le marqueur verbal, « *définir* » :**

Dans cette construction agentive, le verbe **définir** est conjugué à l'indicatif présent ou passé composé (3e personne du singulier ou du pluriel), ainsi qu'au participe présent.

La forme conjuguée au présent de l'indicatif doit être envisagée avec une extension syntagmatique prenant en compte une autre unité lexicale, l'occurrence « comme », conjonction de subordination pouvant être séparée de **définir** par un certain nombre de mots :

*ex : NCEP-ATPIII (National cholesterol education program expert panel on detection, evaluation, and treatment of high blood cholesterol in adults)* ***définissent*** *le syndrome métabolique comme l'association de 3 critères parmi les 5 suivants : répartition androïde des graisses, hypertension systolo-diastolique, hyperglycémie modérée à jeun, hypertriglycéridémie, HDL-cholestérol bas.*

La forme conjuguée au passé composé s'explique par la présence d'une proposition subordonnée circonstancielle de temps exprimant l'antériorité « C'est en 1943 » :

*ex : C'est en 1943 que Lichtenstein et Jaffe* ***ont défini*** *le chondrosarcome qui est un sarcome dont les cellules tumorales sont associées à une matrice cartilagineuse.*

La forme au participe présent peut être utilisée au gérondif, précédée de « en » :

*ex : Fainzang va encore plus loin* ***en définissant*** *la maladie, dans sa dimension sociale, comme étant « … (la) désignation métonymique d'un désordre par sa fixation sur une partie du corps…».*

**Le marqueur nominal « *définition* » :**

Dans les énoncés définitoires comprenant le marqueur nominal **définition**, le terme à définir est toujours complément du nom **définition**.

*ex : On peut retenir comme* ***définition de l'asthme*** *difficile chez l'enfant la persistance d'exacerbations ou de symptômes d'asthme au moins trois jours par semaine, ou la persistance d'une obstruction bronchique.*

### 6.4.3. Nombre d'occurrences de chacun des marqueurs

Si l'on regarde la productivité de chaque marqueur d'énoncés définitoires, on remarque que dans notre corpus, les marqueurs « **verbaux** » ont le plus grand nombre d'occurrences, 151 sur 411 annotations, soit 36 % de l'ensemble des annotations, suivi du marqueur « **parenthétique** », 143 sur 411 annotations, soit 35 % de l'ensemble des annotations. Le verbe « **être** » avec 81 occurrences représente environ 20% des annotations. Les marqueurs « **définition + définir à la voix active** » totalisent 36 annotations, soit 9% des annotations.

## *6.5. Construction de patrons lexico-syntaxiques à partir de ces marqueurs*

Nous abordons maintenant la description des expressions linguistiques constituant les énoncés définitoires annotés manuellement, aboutissement d'une observation minutieuse des relations entre les termes.

### 6.5.1. Généralités

Les patrons lexico-syntaxiques ont été introduits par les travaux de Hearst (Hearst 1992), avec la construction d'un ensemble de PLS spécifiques à l'hyperonymie pour les textes en anglais. Appelés également patrons de relations sémantiques (PRS) ou « *knowledge-rich patterns* » (Meyer 2001), ils décrivent une expression régulière, formée de mots, de catégories grammaticales ou sémantiques et de symboles qui va permettre d'identifier des fragments de texte conformes à cette expression (Rebeyrolle 2000). Ces patrons explicitent un lien entre un terme et des éléments d'information sur son sens.

Leur construction implique un travail d'abstraction réalisé en amont, qui vise à définir une relation lexicale particulière et à identifier ses différents contextes d'apparition. Les éléments véhiculant la relation sont ainsi mis en évidence et synthétisés sous la forme d'un patron lexico-syntaxique. (Dragos 2010)

Ils sont formés d'unités linguistiques (ou marqueurs) indiquant la présence d'une relation lexicale et sous-tendus par un ensemble de contraintes que le contexte lexical ou syntaxique de ce marqueur doit remplir. Les PLS mettent en évidence le lien existant entre deux termes, et sont généralement représentés sous la forme d'un triplet « *Terme1-Marqueur-Terme 2* », dans lequel le marqueur exprime la relation existant entre les deux termes (Lefeuvre 2017).

Il s'agit de structures représentant des schémas récurrents du langage qui sont valorisées dans le domaine du traitement automatique de la langue pour repérer des schémas langagiers dont ils sont l'abstraction. (Dragos 2010)

L'approche par les patrons se justifie par le lien étroit entre lexique et grammaire et considère que le sens d'un mot est « *déterminé par les emplois co-textuels et contextuels antérieurs dans lesquels il apparaît, qu'il s'agisse de l'environnement lexical et des collocations, mais aussi de son environnement sémantique, syntaxique, pragmatique et discursif* ». (Legallois 2013) (Yan 2016)

### 6.5.2. Architecture des patrons lexico-syntaxiques

Nous avons mis au point des patrons lexico-syntaxiques à partir des marqueurs repérés manuellement et listés précédemment.

Nous spécifions tout d'abord les contextes droit et gauche du marqueur.

- Pour les patrons possédant un marqueur verbal, nous cherchons à exploiter les fonctions syntaxiques associées à ce verbe : nous extrayons son sujet et son objet.
- Pour les autres patrons, nous procédons à une exploration contextuelle : nous extrayons le groupe fonctionnel immédiatement à gauche et celui immédiatement à droite du marqueur.

Nous décrivons également la distance entre les composants de ce triplet, exprimée dans nos patrons sous la forme « *X unités lexicales* » et qui sera formalisée dans Unitex à l'aide de la boite **<TOKEN>**[21].

L'élaboration de ces PLS a nécessité de définir au préalable une liste des catégories morpho-syntaxiques (Figure 4) des éléments composant les patrons, ainsi que leurs attributs.

| Catégories morpho-syntaxiques | Autres notations |
|---|---|
| Vbe = Verbe | [o – x unité(s) lexicale(s)] |
| Det = Déterminant | |
| Nom = Nom | |
| Adv = Adverbe | |
| Conj = Conjonction | |
| Prep = Préposition | |
| Adj = Adjectif | |
| Pro = Pronom | |
| Ponc = Ponctuation | |
| Mal = Maladie | |
| Sm = Symbole mathématique | |
| Udm = Unité de mesure | |
| Ne = Nombre entier | |
| Nd = Nombre décimal | |

***Figure 4 : Liste des éléments composant les patrons morpho-syntaxiques***

Les déterminants figurant dans les PLS sont « génériques », représentés par des articles définis (le, la, l', les) et indéfinis (un, une, des, du).

La construction des PLS s'est faite en plusieurs étapes, avec comme objectif de minimiser le silence et le bruit. Pour cela, nous avons réduit les énoncés définitoires à une forme « canonique », ne conservant que les éléments essentiels à la constitution du modèle définitoire et supprimant les éléments accessoires (compléments circonstanciels, propositions subordonnées circonstancielles) qui sont des informations secondaires.

### 6.5.3. Patron lexico-syntaxique à partir du marqueur verbal « être »

Considérant les énoncés définitoires à structure copulative comprenant **être**, nous avons relevé l'existence de propositions simples et de propositions complexes et décrit chacune d'elles au moyen des formes lexico-syntaxiques la caractérisant.

[21] Un token est défini par toute unité lexicale.

Dans notre corpus, les différentes constructions autour du marqueur verbal **être** sont les suivantes :

*ex1 : Le chondrosarcome est une tumeur maligne.*

***Det, Mal ; Vbe (être) ; Det, Mal***

*ex2 : Le classique torticolis nasopharyngien, ou syndrome de Grisel, est un torticolis fébrile de début brutal.*

***Det, Mal ; Ponc ; Conj ; Mal ; Ponc ; Vbe (être) ; Det, Mal ; Prep ; Nom ; Adj***

*ex3 : L'artérite à cellules géantes (ACG) et l'artérite de Takayasu sont des vascularites granulomateuses affectant les vaisseaux de gros calibre.*

***Det, Mal ; Conj ; Det, Mal ; Vbe (être) ; Det, Mal ; Vbe (affecter) ; Det, Nom ; Prep ; Adj ; Nom***

*ex4 : L'anguillulose est une parasitose qui se caractérise par son polymorphisme clinique.*

***Det, Mal ; Vbe (être) ; Det, Mal ; Pro ; Vbe (se caractériser) ; Prep ; Det, Nom***

*ex5 : Chondrosarcomes, ostéosarcomes sont des tumeurs osseuses malignes.*

***Mal ; Ponc ; Mal ; Vbe (être) ; Det, Mal***

*ex6 : Les hémopathies associées aux vascularites sont principalement les leucémies à tricholeucocytes et les myélodysplasies, ainsi que les lymphomes malins, hodgkiniens ou non.*

***Det, Mal ; Vbe (associer) ; Prep, Mal ; Vbe (être) ; Adv ; Det, Mal***

L'observation de ces différents patrons morpho-syntaxiques nous amène à faire certaines remarques :

- Le definiendum se compose soit :
  - ✓ D'une maladie unique (ex1) ;
  - ✓ D'une maladie associée à son synonyme par la conjonction de coordination « ou » (ex2) ;
  - ✓ De deux, voire plusieurs maladies liées par la conjonction de coordination « et » (ex3) ou par une virgule (ex5).
- Le marqueur verbal **être** peut être séparé du syntagme nominal exprimant la définition par un mot, souvent un adverbe (ex6).
- Le verbe **être** s'accorde en genre et en nombre avec le definiendum qui est son sujet.
- Les déterminants entrant dans la composition des énoncés définitoires linguistiques copulatifs sont de valeur générique, excluant tout schéma syntaxique comportant des déterminants autres (adjectifs possessifs, démonstratifs, indéfinis et numéraux).
- Le definiens est représenté par un syntagme nominal.

Il est possible de trouver une unité lexicale insérée entre le marqueur **être** et le déterminant de la maladie composant le *definiens* qui est traduit dans le patron par la notation *[0-1 unité lexicale]*.

Ces premières constatations nous ont conduits à construire un patron lexico-syntaxique du type :
***Det, Mal ; [+ ou – (Conj ; Ponc ; Det)] ; Mal ; Vbe (Etre présent de l'indicatif, 3e pers. singulier ou pluriel) ; [0-1 unité lexicale] ; Det, Nom***

La transformation de ce patron, sous la forme d'un graphe syntaxique appliqué dans Unitex au « corpus de référence », a généré un certain nombre de faux positifs, le syntagme nominal positionné après la copule **être** ne représentant pas une définition, comme le montrent les exemples suivants :

*ex1 : Le torticolis est alors la posture compensatrice qui permet de diminuer la diplopie et retrouver un regard horizontal.*

*ex2 : La méningite est une complication précoce et insidieuse.*

*ex3 : L'épidémie de sida est un facteur prédisposant majeur.*

Constatant qu'au sein des énoncés définitoires le definiens est souvent représenté par l'hyperonyme du definiendum, ayant quant à lui le statut d'hyponyme (terme spécifique), nous avons ajouté une contrainte forte en restreignant le syntagme nominal de l'expression définissante à une maladie, conduisant au PLS suivant :

***Det, Mal ; [+ ou – (Conj ; Ponc ; Det)] ; Mal ; Vbe (Etre présent de l'indicatif, 3epers. singulier ou pluriel) ; [0-1 unité lexicale] ; Det, Mal***

La méthodologie appliquée pour la construction de l'ensemble des PLS a été la suivante (Figure.5), avec pour exemple le marqueur **être** :

- 1ère étape : liste des différents types d'annotation autour du marqueur ;
- 2e étape : écriture du PLS correspondant à chaque type d'annotation ;
- 3e étape : construction d'un patron « synthétique » regroupant les différents contextes textuels pour servir de modèle à l'élaboration du graphe dans Unitex.


***Figure 5 : Méthode de construction des patrons lexico-syntaxiques***

Le patron final autour du marqueur verbal « **être** » s'écrit de la façon suivante :

***Det, Mal ; [+ ou – (Conj ; Ponc ; Det), Mal] ; [0-1 unité lexicale] ; Vbe (Etre présent de l'indicatif, 3e pers. singulier ou pluriel) ; [0-1 unité lexicale] ; Det, Mal***

La distance *[0-1 unité lexicale]* entre le *definiendum* et le verbe *être* s'explique par le fait que l'on peut trouver après la maladie :

- un adjectif qualificatif exprimant la localisation anatomique :

*ex1 : L'atrésie **bronchique** est une malformation congénitale rare.*

*ex2 : Les métastases **oculaires** sont des tumeurs malignes rares des structures oculaires, notamment la choroïde…*

- un pronom démonstratif « c' » remplaçant le definiendum et prenant la fonction de sujet du verbe être :

*ex : Maladie de Buerger. Appelée aussi thrombo-angéite oblitérante, **c'**est une vascularite des hommes jeunes, tabagiques, touchant principalement les artères et les veines de moyen et petit calibre…*

### 6.5.4. Patron lexico-syntaxique autour du marqueur « parenthétique »

On remarque que le contenu des parenthèses positionnées après une maladie est hétérogène. Il peut s'agir de précisions, de commentaires métadiscursifs, de résultats, de références bibliographiques, de noms de personnes, de navigation intra- et intertextuelle, de noms de maladies, d'explicitation de la maladie.

Comme le souligne François (François 2011), les parenthèses sont à proprement parler « *des signes de l'énonciation, c'est-à-dire des signes marquant des phénomènes énonciatifs à tous les niveaux linguistiques des figures graphiques au texte* » et leur contenu a donc souvent peu à voir avec une définition

Les parenthésages que nous avons retenus pour leur contenu définitoire mettent en jeu des relations sémantiques de nature différente entre la maladie et le contenu parenthétique.

On trouve des relations :

- De **synonymie**, le synonyme pouvant être sous la forme,
  - ✓ d'un syntagme nominal

*ex : …selon Schwartz et Dorfman (1975), ce risque est de l'ordre de 25 % dans la maladie des chondromes multiples **(maladie d'Ollier)***

  - ✓ d'un syntagme comprenant une valeur numérique définissant l'équivalence avec le terme défini grâce à un symbole mathématique (<, >, =, ≤, ≥)

*ex : …ostéoporose **(T score < - 2,5)***

- D'**exemplification**

*ex :: …maniérismes moteurs stéréotypés et répétitifs **(par exemple battement ou torsion des mains ou des doigts, mouvements complexes de tout le corps)***

- D'**hyperonymie**

  *ex : …déficits gonadotropes syndromiques* ***(Prader-Willi, Bardet-Biedl, Charge)***

- D'**hyponymie**

  *ex :…thrombocytémie essentielle ou la myélofibrose primitive* ***(syndromes myéloprolifératifs)***

- D'**explicitation**

  *ex :… DRESS* ***(Drug Rash with Eosinophilia and Systemic Symptoms)***

Le contenu parenthétique peut se trouver à une certaine distance du definiendum.

*ex : …des troubles des conduites sociales au premier plan* ***(bizarrerie, psychorigidité, attitude solitaire)***

#### *6.5.4.1. Contenu des parenthèses sans relation avec un énoncé définitoire*

L'analyse des contenus parenthétiques suivant un terme de maladie a conduit à lister un certain nombre d'éléments sans relation avec un énoncé définitoire (Figure 6).

| **Type d'élément** | **Exemple** |
|---|---|
| Symbole | *, %, $ |
| Enoncé relatif à un nombre de patients, de cas | déterminant défini + (patient(e, s) ou cas) |
| Signe de navigation intratextuelle ou extratextuelle | figure(s), fig(s), image(s), table(s), tableau(x), flèche(s), encadré(s), suite, d'après |
| Précision méthodologique | (prélèvement 1/2h après l'injection) |
| Nom d'auteurs | Martin |
| Lieu | Europe, Italie, Rome,… |
| Organisme | OMS, HAS, … |
| Nombre décimal ou entier non suivi d'une unité de mesure | 1,5 ; 20 |
| Abréviation utilisée par l'auteur dans le contexte de l'article | MdP pour Maladie de Paget |
| Chiffre romain comme marqueur de renvoi au chapitre d'un ouvrage | I, II, III, IV, V,… |
| Résultat épidémiologique | Prévalence estimée à 17,8 p. 100. |

***Figure 6 : Contenus parenthétiques sans relation avec une définition***

Dans un souci d'efficacité, nous avons constitué un dictionnaire dit des « *mots rejetés* » **<MR>** (Figure.7) que nous avons intégré dans Unitex et que nous sélectionnons lors de l'application du graphe parenthèses afin d'exclure ces contenus parenthétiques sans lien avec un énoncé définitoire.

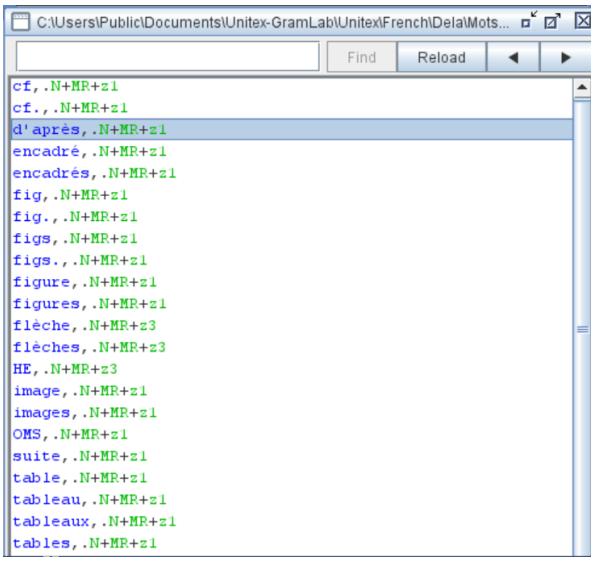



***Figure 7 : Dictionnaire des « mots rejetés »***

#### *6.5.4.2. Contenu des parenthèses constituant un énoncé définitoire*

En nous concentrant sur les contenus parenthétiques pertinents, nous avons dégagé certains éléments lexicaux et syntaxiques les caractérisant et formulé les patrons correspondants.

Le contenu des parenthèses peut être :

- Un **nom de maladie**

  *ex : la maladie des chondromes multiples* ***(maladie d'Ollier)***

  *Det, Mal ;* ***(Mal)***

- Une **suite de noms de maladies**

  *ex : déficits gonadotropes syndromiques* ***(Prader-Willi, BardetBiedl, Charge)***

  *Mal ;* ***(Mal ; Ponc ; Mal ; Ponc ; Mal)***

- Une **phrase nominale**

  *ex : tuberculose congénitale* ***(transmission de la mère au foetus pendant la grossesse)***

  *Mal ;* ***(Nom ; Prep ; Det, Nom ; Prep ; Nom ; Prep ; Det, Nom)***

- Un **groupe nominal suivi d'une unité de mesure**

  *ex1 : insuffisance rénale* ***(urée > 400 mg/l ou créatininémie > 15 mg/l)***

  *Mal ;* ***(Nom ; Sm ; Ne ; Udm ; Conj ; Nom ; Sm ; Ne ; Udm)***

  *ex2 : leucémie myélomonocytaire chronique* ***(monocytes circulants > 1x10$^9$/L)***

  *Mal ;* ***(Nom ; Adj ; Sm ; Ne ; Udm)***

  *ex3 : acidose métabolique`* ***(bicarbonates plasmatiques < 20 mmolil)***

  *Mal ;* ***(Nom ; Adj ; Sm ; Ne ; Udm)***

L'analyse des contenus à l'intérieur des parenthèses a permis la construction d'un PLS excluant un certain nombre d'éléments, formalisé par l'opérateur « **SAUF** », et intégrant une distance de [*0-5 unité(s) lexicale(s)*] entre le definiendum et la définition.

Le patron final autour du marqueur « **parenthétique** » comprend les éléments suivants :

***Mal ; [0-5 unités lexicales] ; [Mal OU Nom + Sm + (Ne ou Nd) + Udm SAUF (%, *, nombre de patients, références bibliographiques, cf, infra, supra, figure, image, tableau, flèche, table, précisions méthodologiques, personne, résultats épidémiologiques comme les taux de prévalence)]***

Les items lexicaux exclus par l'opérateur ***SAUF*** constituent une liste spécifique au domaine médical et au corpus étudié. Elle n'est bien sûr pas exhaustive et doit être adaptée à chaque corpus.

Il est vraisemblable que ce procédé d'exclusion d'un certain nombre d'éléments s'impose de façon générale dans le graphe parenthétique et soit spécifique du domaine étudié, nécessitant une étude systématique des parenthésages et un bon niveau d'expertise.

### 6.5.5. Patron lexico-syntaxique autour du marqueur nominal « définition »

En nous concentrant sur les énoncés définitoires comportant le marqueur **définition**, nous avons dégagé certains éléments lexicaux et syntaxiques les caractérisant et formulé les patrons correspondants.

On remarque que le marqueur nominal **définition** :

- Est dans la majorité des cas suivi d'un complément du nom exprimant la maladie à définir, comme le montrent les exemples suivants :

  *ex : La **définition des vascularites** est simple et anatomopathologique : elle se résume à une inflammation des vaisseaux sanguins quel(s) qu'en soi(en)t le(s) mécanismes.*

  ***Det, Nom « définition » ; Det, Mal ; Vbe***

- Fait parfois partie de la locution adverbiale « *Par définition* » qui est suivie du nom de la maladie à définir :

  *ex : **Par définition, le syndrome parkinsonien** comporte un symptôme central, l'akinésie, accompagné d'au moins l'un des signes suivants : rigidité extrapyramidale, tremblement de repos et instabilité posturale.*

  ***Prep ; Nom « définition » ; Ponc ; Det, Mal ; Vbe***

Le patron autour du marqueur nominal « **définition** » se résume à :

***Nom « définition » ; Det, Mal***

### 6.5.6. Patron lexico-syntaxique autour du marqueur verbal « définir » à la voix active

Il s'agit dans ce cas de la construction agentive du verbe **définir**. La maladie a la fonction de complément d'objet direct.

Les traits de conjugaison du verbe **définir** retrouvés dans cette construction sont le présent de l'indicatif (ex1), le passé composé (ex2) et le participe présent (ex3), comme le montrent les exemples suivants.

*ex1 : Les instances américaines du NCEP-ATPIII (National cholesterol education program expert panel on detection, evaluation, and treatment of high blood cholesterol in adults)* ***définissent*** *de façon unanime le syndrome métabolique comme l'association de 3 critères…*

*ex2 : C'est en 1943 que Lichtenstein et Jaffe* ***ont défini*** *le chondrosarcome. C'est un sarcome dont les cellules tumorales sont associées à une matrice cartilagineuse.*

*ex3 : Dans notre série, les résultats du frottis sanguin avaient révélé la présence : des hématies en rouleaux, une plasmocytose circulante chez 4 patients dont un avait un taux >20 %* ***définissant*** *une leucémie à plasmocytes.*

La séquence commune à ces différents exemples est ***:***

***Vbe « définir » ; [0-x unité(s) lexicale(s)] ; Det, Mal***

### 6.5.7. Patron lexico-syntaxique commun aux marqueurs « définition » et « définir »

Devant la similitude des constructions autour des marqueurs nominal **définition** et verbal **définir** à la voix active du présent ou passé composé de l'indicatif et au participe présent, nous proposons un patron unique.

Patron lexico-syntaxique commun au marqueur nominal « **définition** » et au marqueur verbal « **définir** » à la voix active :

***(Nom « définition » OU Vbe « Définir ») ; [0-x unité(s) lexicale(s)] ; Det, Mal***

### 6.5.8. Patron lexico-syntaxique autour des autres marqueurs « verbaux »

Le marqueur verbal identifié comme marqueur d'un énoncé définitoire d'une maladie peut se trouver positionné différemment par rapport à la pathologie à laquelle il apporte une information « *définitoire* ».

- Soit il se situe dans la proposition principale constituant l'énoncé définitoire et a comme sujet la maladie :

*ex : Le* ***diabète de grossesse*** *apparaît vers la fin du 2e et au 3e trimestre et* ***se manifeste par*** *une augmentation du taux de sucre dans le sang qui survient uniquement lors de la grossesse*

- Soit il a comme sujet une anaphore pronominale complète, c'est-à-dire un pronom personnel reprenant la visée référentielle de son groupe nominal antécédent, à savoir la maladie :

*ex :: L'****angéite granulomateuse allergique*** *ou syndrome de Churg et Strauss****. Elle se caractérise****, cliniquement,* ***par*** *l'existence d'un asthme grave, d'une hyperéosinophilie sanguine et d'une angéite nécrosante touchant les artères de petit calibre, les capillaires et les veinules et typiquement de manifestations extra-pulmonaires.*

La séquence commune à ces différents exemples est :

***Det, Mal ; [0-10 unité(s) lexicale(s)] ; Vbe***

Les verbes recueillis à partir des annotations du corpus sont regroupés dans une liste intégrée dans Unitex sous la forme d'un dictionnaire **<VERBES>** (Figure.8) incluant les traits de conjugaison repérés.

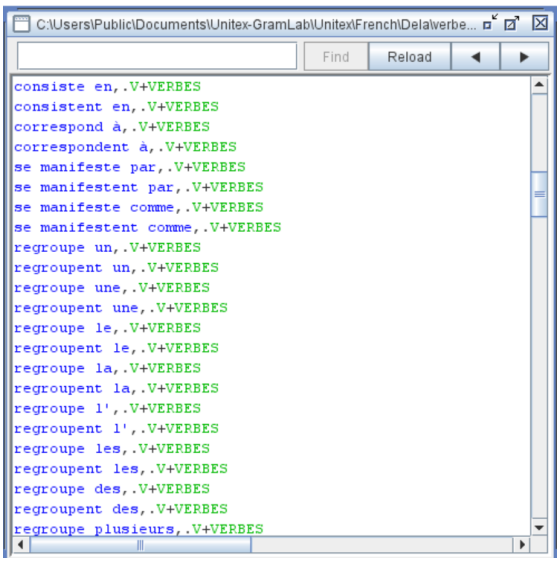



***Figure 8 : Dictionnaire des marqueurs verbaux***

## *6.6. Phase de travail dans Unitex*

Cette étude expérimentale comprend une prise en main d'Unitex qui ne reflètera que très partiellement les potentialités du logiciel et son champ d'utilisation.

### 6.6.1. Présentation d'Unitex

Développé par S. Paumier (Paumier 2011) en 2002, Unitex/GramLab est une suite logicielle open source (selon les termes de la Lesser General Public License (LGPL)[22]), multiplateforme (moteur TAL écrit en C++, l'IDE[23] Graphique écrite en Java), multilingue (conforme au standard Unicode 3.0[24]) qui permet de construire des descriptions formalisées de grammaires et d'utiliser des ressources telles que des dictionnaires de la langue à large couverture.

Il permet aux utilisateurs, non seulement d'implémenter des expressions régulières dans la recherche des informations, mais aussi de construire, de vérifier et d'appliquer des dictionnaires électroniques personnalisés.

Tous les objets traités par Unitex sont ou peuvent être transformés en des transducteurs à nombre fini d'états. Les opérations sur les textes, grammaires et dictionnaires renvoient à des

[22] http://www.gnu.org/licenses/lgpl-3.0.html
[23] Integrated Development Environment
[24] http://www.unicode.org/versions/components-3.0.0.htm

opérations sur des transducteurs. Un transducteur est un automate dont les transitions sont étiquetées par un couple de symbole : un symbole reconnu en entrée et un symbole produit en sortie. Un transducteur permet donc de reconnaître une chaîne en entrée et produit, en sortie, une autre suite de caractères (Tolone 2006).

Les grammaires sont représentées sous forme de graphes aisément réalisables grâce à un éditeur intégré.

### 6.6.2. Chargement et prétraitement du corpus textuel

Le corpus de 48 documents « ***Corpus_gold.txt*** » est chargé dans Unitex.

Les textes du corpus sont ensuite soumis à des opérations de prétraitement et d'application des dictionnaires électroniques d'Unitex.

Le prétraitement (Figure.9) consiste à appliquer au texte une série d'opérations :

- normalisation des séparateurs[25] ;
- découpage en unités lexicales ou « *tokenisation* » ;
- normalisation de formes non ambiguës (Option *« Apply FST2 in REPLACE mode »*) ;
- découpage en phrases (Option *« Apply FST2 in MERGE mode »*) qui apparaitront dans le texte, précédées du symbole délimiteur de phrases {S} ;
- application des dictionnaires au format DELA (Option *« Apply All default Dictionaries »*).

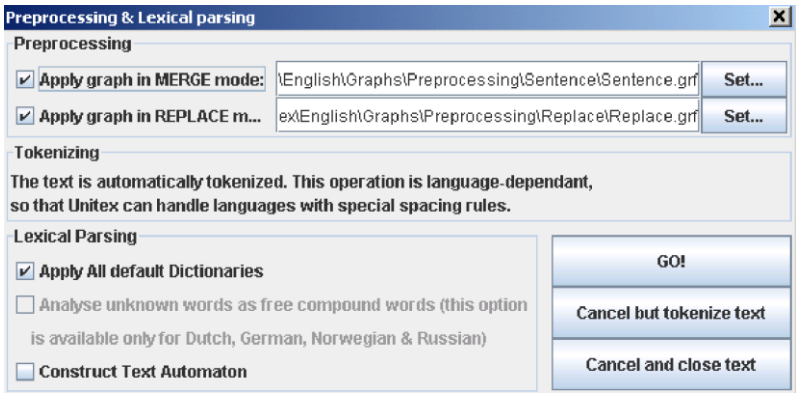



***Figure 9 : Fenêtre de prétraitement***

Le résultat de la normalisation du « ***Corpus_gold.txt*** » est un fichier situé dans le même répertoire que le .txt et dont le nom est « ***Corpus_gold.snt*** » (Figure.10).

[25] « Les séparateurs usuels sont l'espace, la tabulation et le retour à la ligne. On peut rencontrer plusieurs séparateurs consécutifs dans des textes, mais comme cela n'est d'aucune utilité pour une analyse linguistique, on normalise ces séparateurs selon les règles suivantes : toute suite de séparateurs contenant au moins un retour à la ligne est remplacée par un unique retour à la ligne et toute autre suite de séparateurs est remplacée par un espace. »

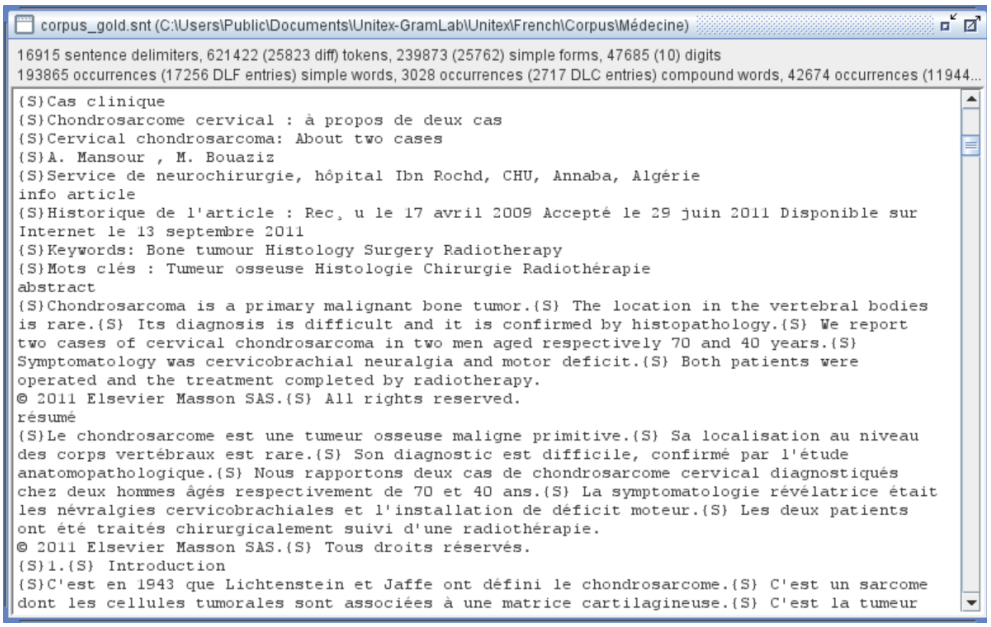



***Figure 10 : Extrait du Corpus_gold.snt***

### 6.6.3. Application des dictionnaires

#### *6.6.3.1. Dictionnaires électroniques utilisés par Unitex*

Les dictionnaires électroniques utilisés par Unitex utilisent le formalisme DELA « *Dictionnaires Electroniques du LADL* [26] » qui permet de décrire les entrées lexicales simples et composées d'une langue en leur associant de façon optionnelle des informations grammaticales, sémantiques et flexionnelles.

Il existe plusieurs types de dictionnaires électroniques. Celui que l'on utilise le plus couramment est le dictionnaire de formes fléchies, appelé DELAF (DELA de formes Fléchies) ou encore DELACF (DELA de formes Composées Fléchies) lorsqu'il s'agit d'un dictionnaire de mots composés.

Les ressources actuellement disponibles pour le français sont les suivantes : 67 tables de verbes distributionnels simples, 81 tables de noms prédicatifs simples et composés, 69 tables d'expressions figées (surtout verbales et adjectivales) et 32 tables d'adverbes, pour un total de plus de 78 000 entrées.

#### *6.6.3.2. Ressources linguistiques intégrées dans Unitex*

Comme nous l'avons précisé précédemment, il est possible de créer ses propres dictionnaires (dictionnaires externes) dans Unitex en suivant certaines normes d'écriture :

***« Forme fléchie, Forme canonique. Code Grammatical + Code Sémantique : Genre Nombre ».***

Seule la forme fléchie est obligatoire.

[26] Laboratoire d'Automatique Documentaire et Linguistique de l'Université Paris VII, intégré depuis l'année 2000, au sein du LIGM (Laboratoire Informatique Gaspard Monge) de l'Université paris-Est Marne-la-Vallée. https://igm.univ-gustave-eiffel.fr/

Pour les besoins de notre expérimentation nous avons créé 3 dictionnaires :

- dictionnaire des formes fléchies des noms de maladies « ***patho_avec_pluriels_2.dic*** », formalisé sous la forme **<PATHO>** dans Unitex ;
- dictionnaire des marqueurs verbaux avec leurs traits de conjugaison « ***verbes.dic*** », formalisé sous la forme **<VERBES>** dans Unitex ;
- dictionnaire des mots à exclure du contenu des parenthèses « ***mots_rejetés.dic*** », formalisé sous la forme **<MR>** dans Unitex .

Le dictionnaire des formes fléchies des noms de maladies n'est pas figé. La version utilisée pour cette expérimentation s'est construite en 2 étapes.

- Une construction initiale à partir des entrées du « *Thésaurus des Pathologies humaines* ».

  Chaque forme est suivie d'une séquence d'informations grammaticales et sémantiques : ***N*** signifiant qu'il s'agit d'un nom et ***z3*** d'un langage très spécialisé (Figure.11).

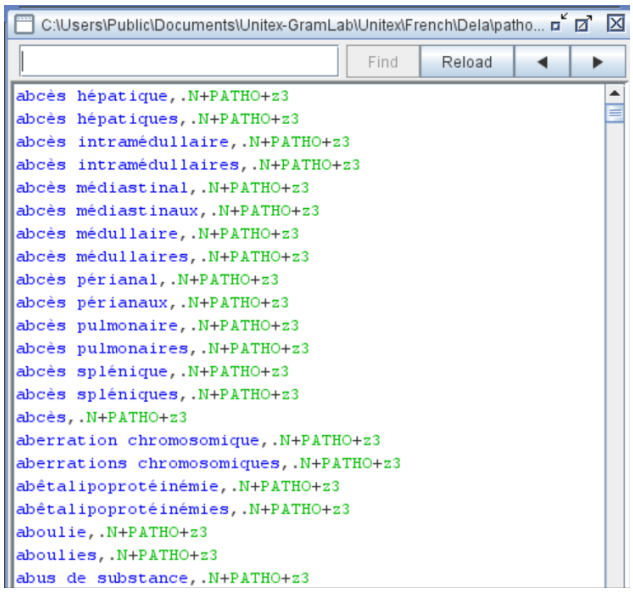



***Figure 11 : Dictionnaire des formes fléchies de maladies***

- Un enrichissement à partir du dictionnaire des mots non reconnus dans Unitex.

En effet, après la phase de prétraitement du corpus textuel par Unitex et l'application des dictionnaires, le programme Dico produit les fichiers suivants et les génère dans le répertoire du texte.

- ✓ dlf : dictionnaire des mots simples du texte
- ✓ dlc : dictionnaire des mots composés du texte
- ✓ err : liste des mots inconnus du texte

Dans la liste des mots inconnus (Figure.12) figurent des noms de maladies qui sont ajoutés au dictionnaire des formes fléchies de maladies, portant le nombre d'entrées à 15 189.

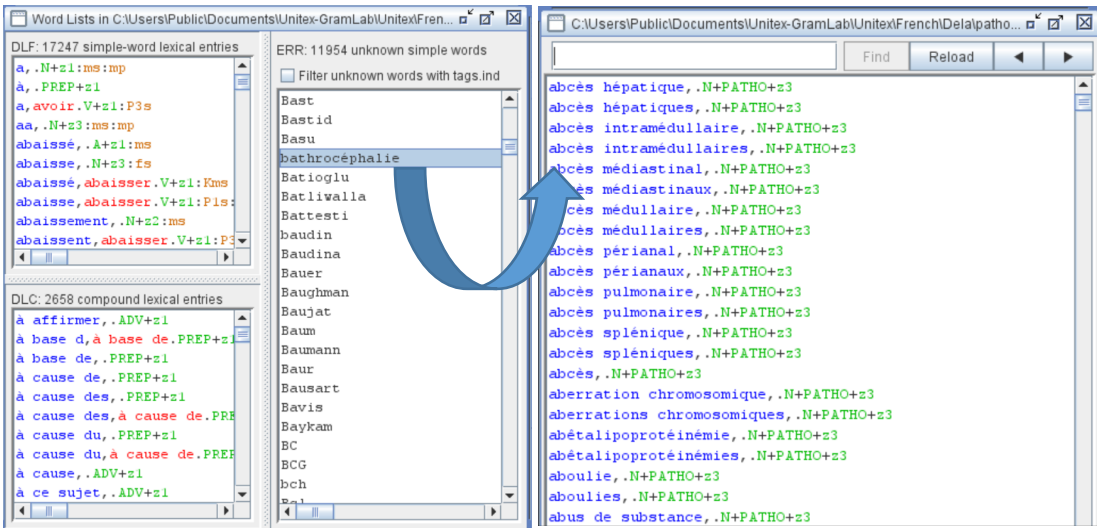



***Figure 12 : Liste des mots inconnus du texte***

### 6.6.4. Construction des graphes syntaxiques

Les graphes syntaxiques, également appelés grammaires locales, constituent un puissant formalisme pour décrire des motifs syntaxiques qui pourront ensuite être recherchés dans des textes. Ils possèdent une grande puissance d'expressions car ils permettent de faire référence aux dictionnaires.

#### *6.6.4.1. Création d'un graphe dans Unitex*

Un graphe (Figure.13) est créé dans un éditeur particulier qui s'ouvre sur deux symboles : le symbole en forme de flèche est le début du graphe, le symbole composé d'un cercle entourant un carré marque la fin du graphe.

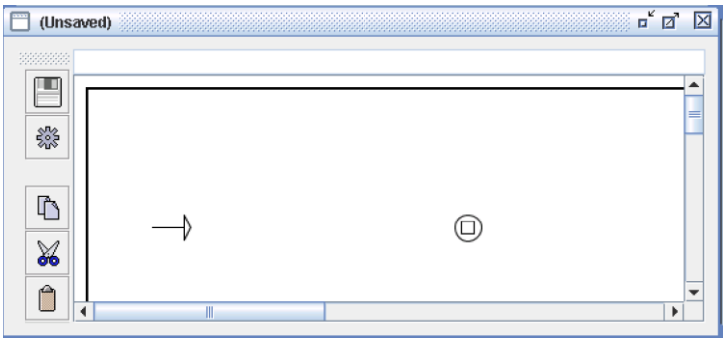



***Figure 13 : Graphe vierge***

Un graphe est une succession de boites reliées par des segments de droite et limité par ces deux symboles. A l'intérieur de ces boites, on peut mettre des mots, des expressions, des dictionnaires (internes à Unitex ou externes) ou des expressions régulières.

#### *6.6.4.2. Graphes pour le repérage des énoncés définitoires en corpus textuel*

L'élaboration des graphes que nous présentons s'est faite en plusieurs étapes, s'appuyant sur les PLS préalablement construits.

Le graphe général (Figure.14) est constitué de 5 branches que nous allons détailler.

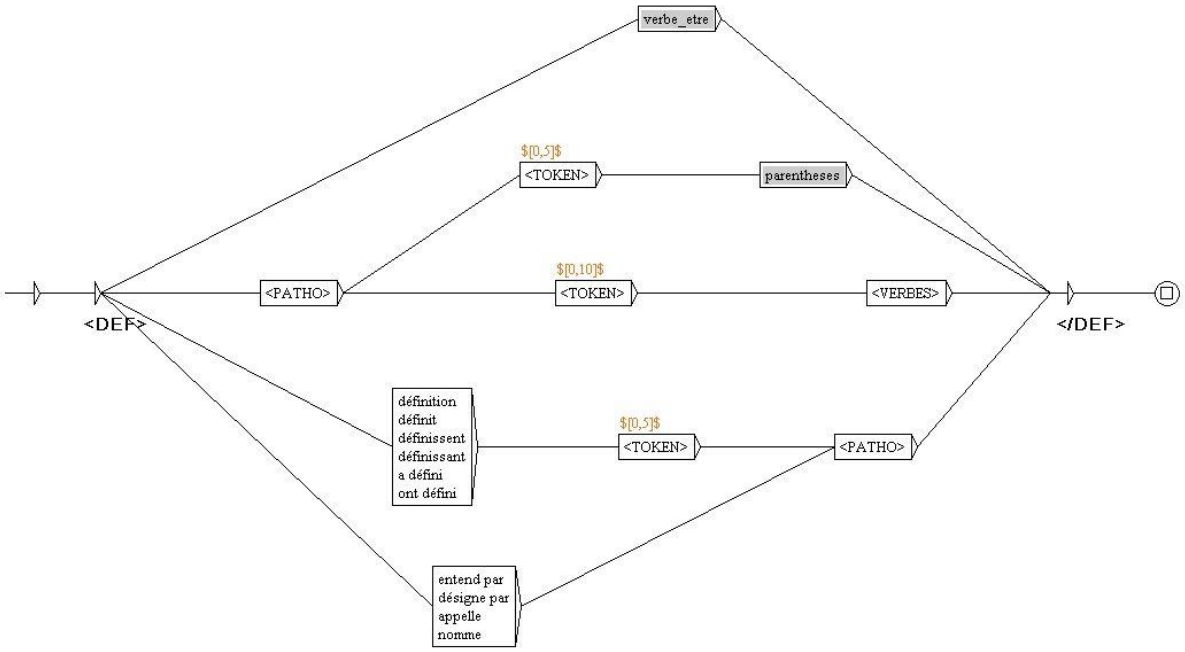



***Figure 14 : Graphe général***

Ce graphe est limité par des balises <DEF> et </DEF> qui vont entourer les définitions dans la sortie éditée du texte. Les boites grisées renvoient à des sous-graphes dessinés dans un autre éditeur afin de ne pas surcharger l'éditeur principal.

**<PATHO>** et **<VERBES>** sont les deux dictionnaires externes, **<TOKEN>** est un dictionnaire interne à Unitex qui contient les unités lexicales extraites du texte.

#### 6.6.4.2.1. Graphe du verbe « être »

Le graphe du verbe **être** (Figure.15) contient une boite grisée appelant un sous-graphe (Figure.16).


***Figure 15 : Graphe du verbe être***

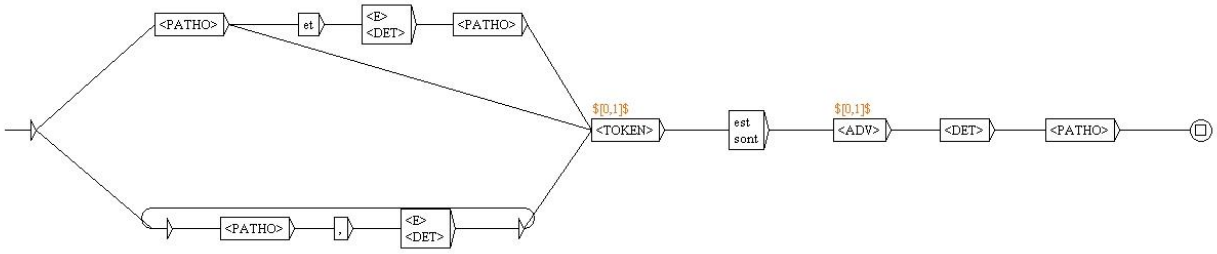



***Figure 16 : Sous-graphe du verbe être***

La première partie du graphe en forme de losange permet de rechercher une ou deux pathologies séparées par la conjonction de coordination « *et* » (branche du haut) ou « *n* » pathologies séparées par des virgules « , » (branche du bas).

Ces pathologies sont ensuite mises en relation, grâce aux déclinaisons du verbe **être** (est/sont), avec une autre pathologie, souvent plus générique, elle-même présente dans le dictionnaire **<PATHO>**.

Le verbe **être** est celui qui ramène beaucoup de définitions mais également celui qui produit le plus de bruit dans les réponses. Pour limiter au maximum ce bruit, nous avons donc imaginé la forme de relation ***« pathologie(s) spécifique(s) ↔ pathologie générique »*** décrite lors de la construction du PLS.

#### 6.6.4.2.2. Graphe du marqueur « parenthétique »

Le graphe du **marqueur parenthétique** (Figure.17) appelle plusieurs branches.

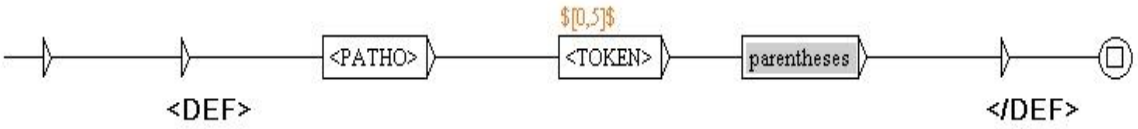



***Figure 17 : Graphe du marqueur parenthétique***

Les définitions entre parenthèses sont en général très proches de la pathologie (avec un maximum de 5 tokens entre les deux).

La boite grisée appelle à son tour le sous-graphe des parenthèses (Figure.18).

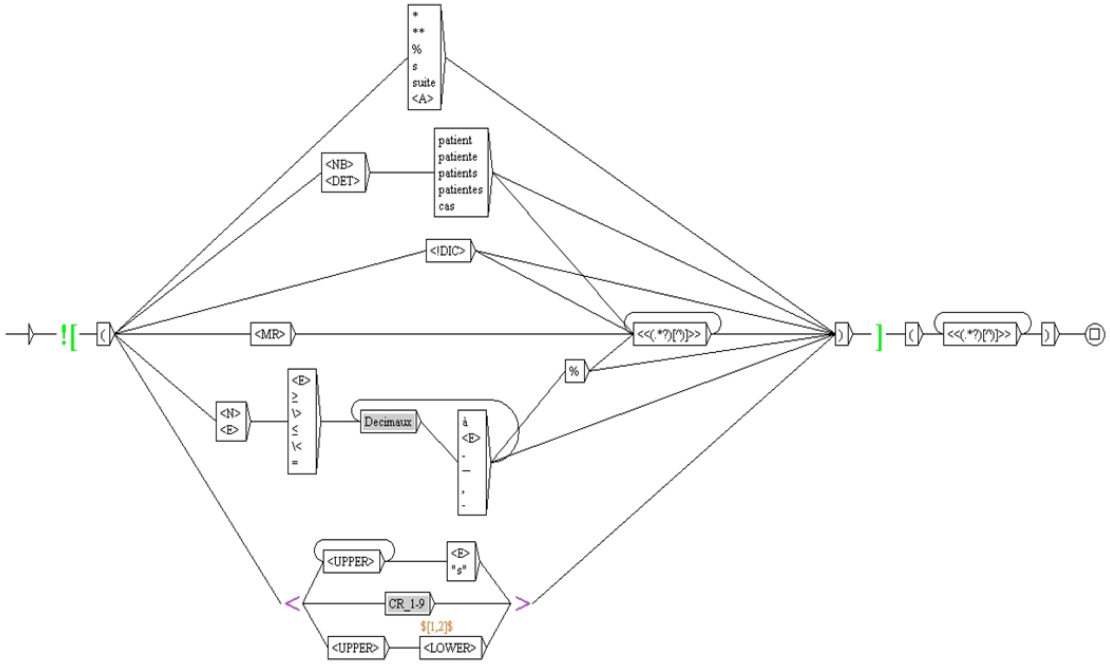



***Figure 18 : Sous-graphe du marqueur parenthétique***

L'étude du corpus de médecine a montré que les parenthèses peuvent contenir les objets les plus hétéroclites. En conséquence, il est impossible de construire un graphe-grammaire pour la recherche de définitions. Nous avons donc inversé la logique en récupérant le contenu de toutes les parenthèses grâce à une expression régulière *<<(.*?)[^)]>>* et supprimé ensuite ce qui ne constituait pas à fortiori une définition.

Le contenu supprimé apparaît entouré de crochets verts dans le sous-graphe.

**< !DIC>** est le dictionnaire des mots inconnus d'Unitex qui n'apparaissent ni dans un dictionnaire interne, ni externe : ce sont par exemple des noms propres, des lieux géographiques, des noms d'instituts, etc qui n'appellent pas de définitions.

Le renvoi des parenthèses à une figure, un tableau, une image, etc est supprimé grâce au dictionnaire externe **<MR>.**

Ont également été supprimés les décimaux tels que « *< 12-18 % des cas …* ». Le sous-graphe des décimaux incluent les nombres entiers.

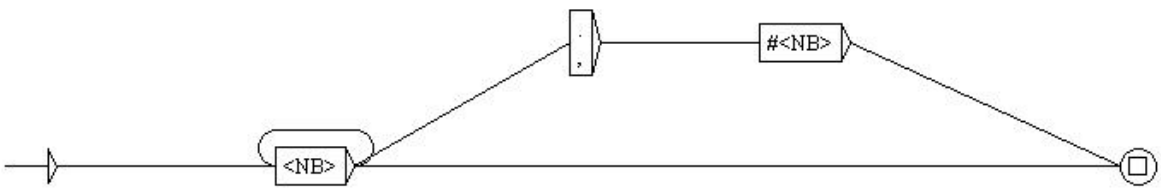



***Figure 19 : Branche des décimaux + unité de mesure***

Par contre, cette branche (Figure.19) permet de conserver les décimaux (ou entiers) suivis d'une unité de mesure, ces décimaux apportant une précision à une définition précédant la parenthèse.

*ex : Insuffisance rénale* ***(urée > 400 mg/l ou créatininémie > 15 mg/l)***

Le losange en bas du sous-graphe des parenthèses (Figure.20) permet d'éliminer les abréviations de termes quelquefois utilisés dans l'écriture d'un article lorsque ce terme est répété dans le corps de l'article.

*ex* ***: MNs*** *pour maladies nosocomiales*

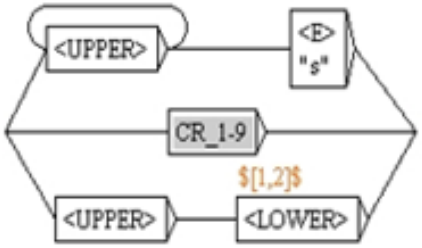



***Figure 20 : Branche de suppression des abréviations de concepts***

Le renvoi à des têtes de chapitres désignés par un chiffre romain est supprimé grâce au sous-graphe CR_1-9 (Figure.21) :

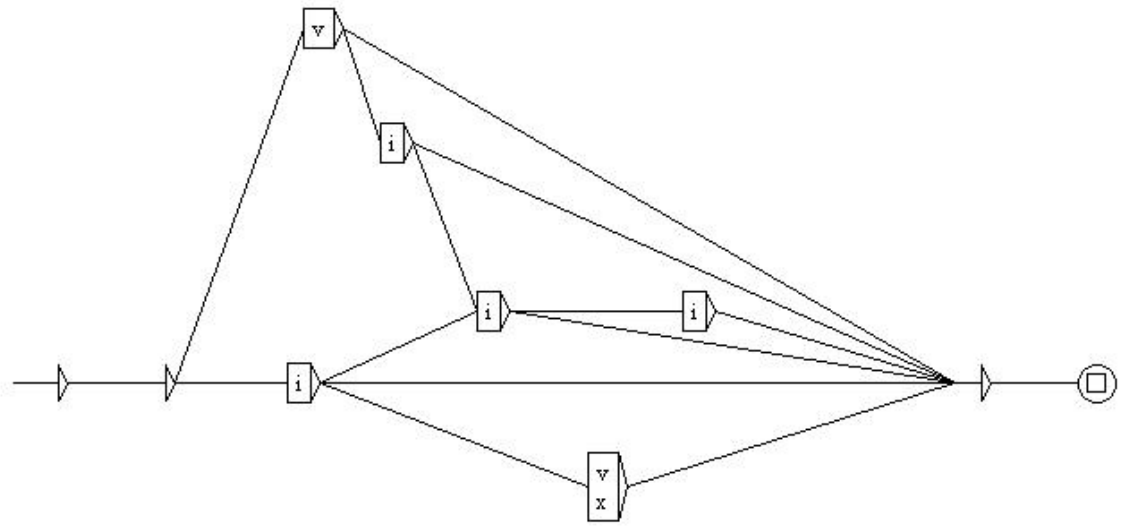



***Figure 21 : Sous-graphe CR_1-9***

Il est bien entendu que cette liste d'éléments supprimés dans les parenthèses ne peut être exhaustive et nécessite donc une relecture manuelle des résultats.

#### 6.6.4.2.3. Graphe des marqueurs « verbaux »

En ce qui concerne le graphe des **marqueurs verbaux** (Figure.22), les boites **<PATHO>** et **<VERBES>** peuvent être séparées de 10 tokens au maximum, conformément au PLS, réduisant ainsi le risque de décorrélation de la pathologie et du verbe associé.


***Figure 22 : Graphe des marqueurs verbaux***

#### 6.6.4.2.4. Graphe du marqueur nominal « définition » et de « définir » à la voix active

Nous appuyant sur le PLS construit autour du nom **définition** et du verbe **définir** à la voix active, nous avons réuni dans la première boite de ce graphe (Figure.23) le nom définition, ainsi que les traits de conjugaison du verbe définir retrouvés dans le corpus. Cette boite est séparée de la maladie par un maximum de 5 tokens.


***Figure 23 : Graphe « définition et définir »***

#### 6.6.4.2.5. Graphe des expressions particulières « entend par, désigne par, appelle, nomme »

Les marqueurs de ce graphe (Figure.24) n'ont pas été identifiés dans le corpus de 48 documents mais dans un corpus de 10 000 références qui est en cours d'analyse. Nous avons donc créé un graphe à partir de cet ensemble de **marqueurs d'appellation**, **de nomination et de désignation** mais sa pertinence n'a pas été évaluée.

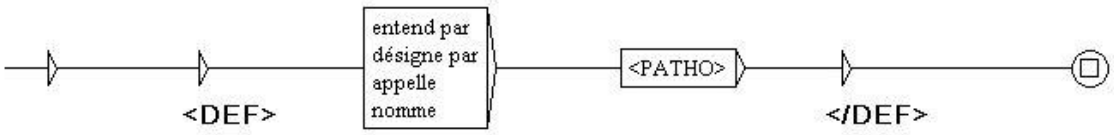



***Figure 24 : Graphe des expressions***

### *6.6.4.3. Lancement des graphes dans Unitex*

Dans Unitex, on sélectionne le corpus prétraité (Corpus_gold.snt) : « *Text > Open Tagged Text> Corpus_gold.snt ».*

Puis on applique les dictionnaires choisis (Figure.25) : « *Text > Apply lexical ressources ».* On dispose des ressources du système Unitex et des dictionnaires construits pour cette expérimentation.

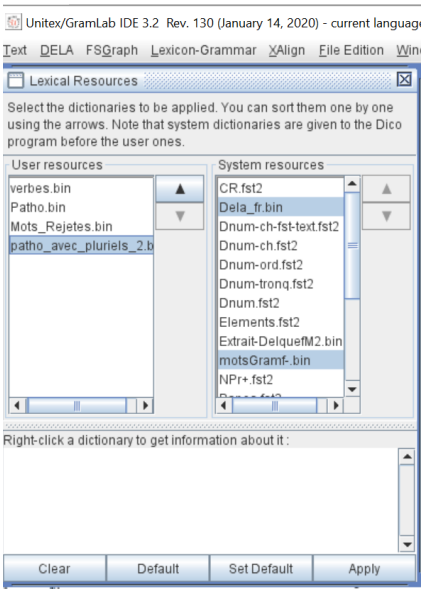



***Figure 25 : Unitex/Ressources lexicales***

Le dictionnaire « *patho_avec_pluriels_2.bin* », (rappelé par **<PATHO>** dans le graphe) est utilisé lors de l'application de chacun des graphes.

Le dictionnaire « *mots_Rejetés.bin* » (rappelé par **<MR>** dans le graphe), est sélectionné avec le graphe des parenthèses et « *verbes.bin* » avec le graphe des verbes (rappelé par **<VERBES>** dans le graphe).

L'étape suivante consiste à choisir un graphe et à l'appliquer sur le corpus (Figure.26) : « *Text > Locate Pattern »* avec balisage des occurrences trouvées dans le texte « *Merge with input text »*.

En terminologie, la recherche de contextes d'utilisation d'un terme est une tâche très importante pour réaliser des ressources terminologiques. Dans Unitex, elle se fait par le biais de la fonction « *Locate Pattern* » dans le menu Text. Il est possible de mener cette recherche de deux façons, soit par expression régulière (regular expression) ou par application de grammaires locales (graph) comme dans notre expérimentation.

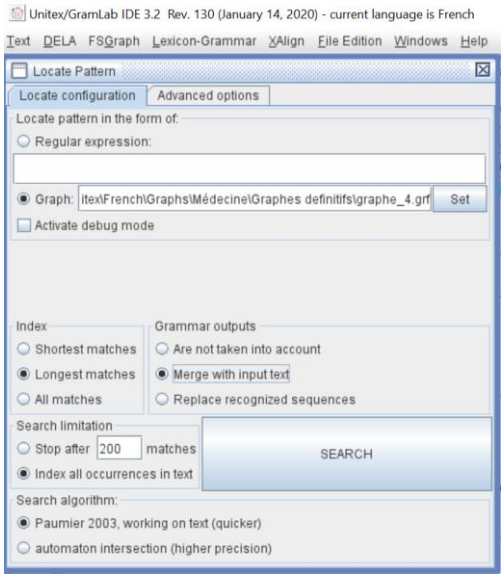



***Figure 67 : Unitex/Fonction « Locate pattern »***

La dernière étape permet de visualiser les concordances obtenues qui sont affichées en bleu, sous forme de liens hypertextuels (Figure.27).

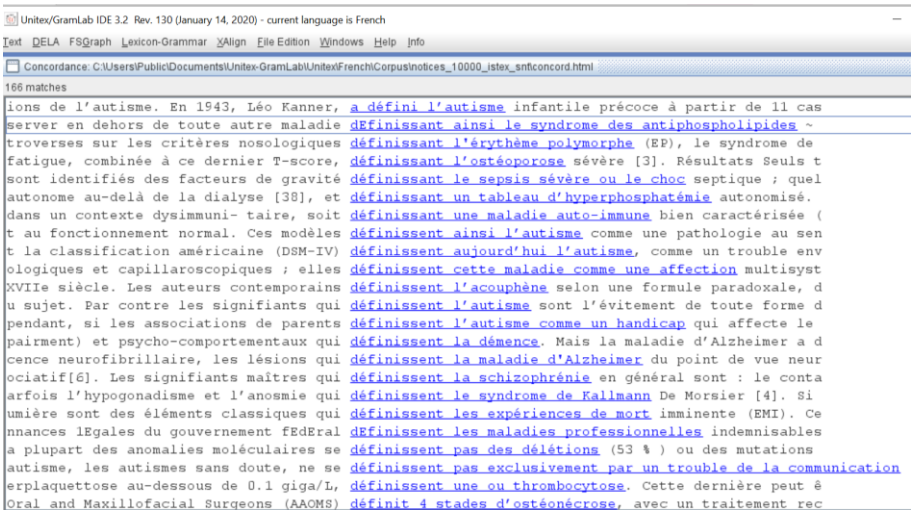
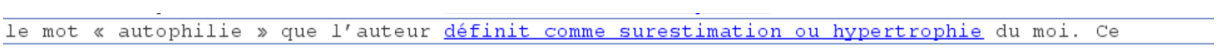



***Figure 27: Unitex/Affichage des concordances avec le graphe « définition »***

Il suffit ensuite de cliquer sur ces liens pour retrouver le contexte d'utilisation de la forme recherchée, qui apparaît surlignée en bleu turquoise (Figure.28). Dans la fenêtre sont indiqués le nombre d'occurrences trouvées, le nombre d'unités lexicales reconnues, ainsi que le rapport entre ce nombre et le nombre total d'unités lexicales du texte.



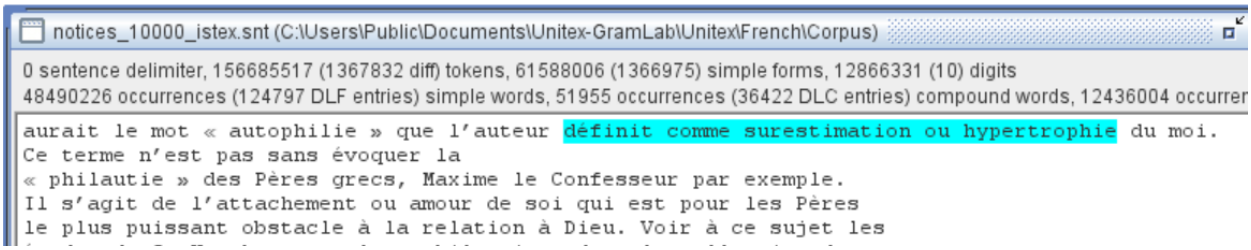



***Figure 28 : Unitex/affichage du contexte repéré***

On peut sélectionner le mode de tri à appliquer dans la liste « *Sort According to* ». Le mode « *Text Order* » affiche les occurrences dans l'ordre où elles apparaissent dans le texte (Figure.29). Les six autres modes permettent de trier en colonnes. Les trois zones d'une ligne sont le contexte gauche, l'occurrence et le contexte droit. Les occurrences et les contextes droits sont triés de gauche à droite. Les contextes gauches sont triés de droite à gauche. Le mode utilisé par défaut est « *Center, Left Col* ». La concordance est produite sous la forme d'un fichier HTML.

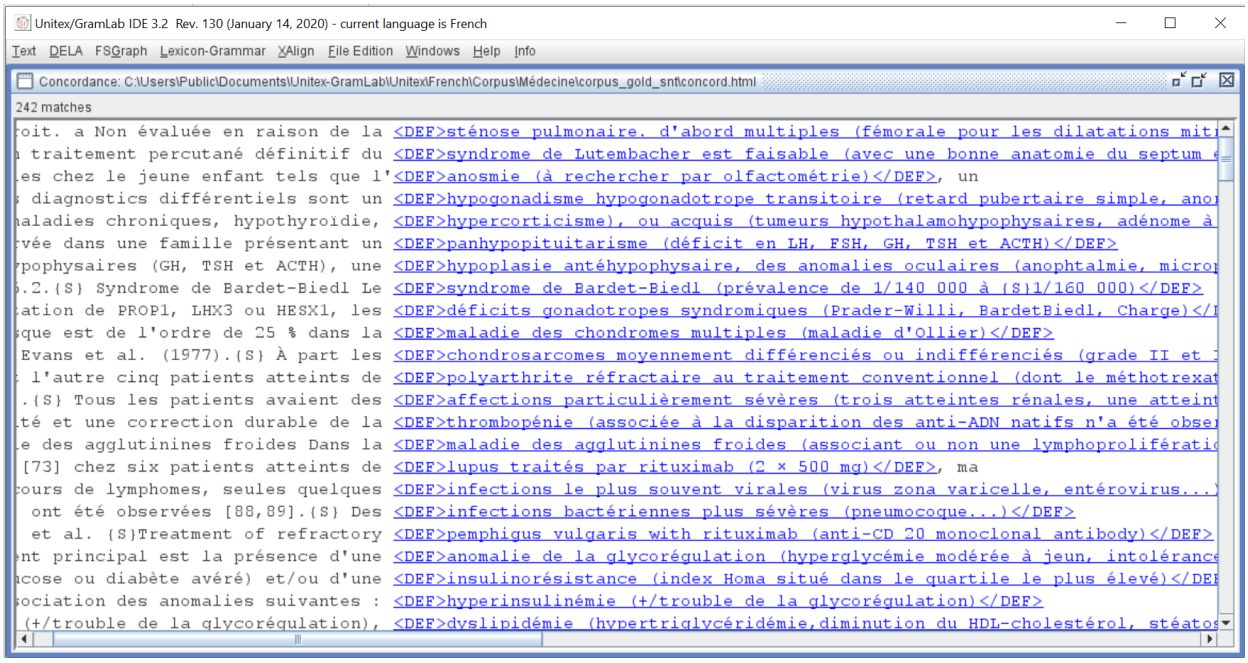



***Figure 29 : Affichage des concordances selon le mode « Text Order »***

#### *6.6.4.4.Résultats de l'application des graphes sur le corpus*

On note 229 concordances pour le *graphe parenthèses*, 171 pour le *graphe des verbes*, 28 pour le *graphe définition + définir* et 44 pour le *graphe être*.

# 7. Evaluation

## *7.1. Mesures d'évaluation de la pertinence des graphes syntaxiques : précision, rappel et F-mesure*

L'évaluation a pour objectif de mesurer la performance des graphes élaborés à l'aide d'Unitex dans la reconnaissance d'énoncés définitoires dans des textes scientifiques spécialisés.

La capacité du système à sélectionner des documents pertinents s'évalue par les mesures classiques que sont la précision (la proportion d'extractions correctes parmi les résultats obtenus à l'aide des graphes), le rappel (la proportion d'extractions pertinentes du système parmi les résultats attendus selon le corpus de référence annoté manuellement) et la F-Mesure, introduite par Van Rijbergen (van Rijsbergen 1979) qui est une moyenne harmonique entre rappel et précision.

Le rappel est d'autant plus élevé que le silence est faible.

La précision est d'autant plus élevée que le bruit est faible, c'est-à-dire qu'elle permet de mesurer à l'intérieur des énoncés détectés comme étant définitoires, la proportion qui est effectivement définitoire.

Comme le montre le tableau suivant (Tableau 1), le *graphe être* est celui qui est le plus précis en termes d'exactitude et de qualité des énoncés définitoires identifiés mais également le moins sensible en ce qui concerne l'exhaustivité des énoncés retrouvés, avec un silence (faux négatifs) important.

Le *graphe des parenthèses* est le moins précis, ce qui sous-entend qu'il ramène un certain nombre de faux-positifs, c'est-à-dire des contenus parenthétiques qui ne sont pas définitoires.

| Graphe | Nombre de définitions attendues (corpus de référence) | Nombre total de définitions trouvées avec Unitex | Nombre de définitions pertinentes trouvées avec Unitex | Précision | Rappel | F-Mesure |
|---|---|---|---|---|---|---|
| **Graphe « être »** | 81 | 44 | 43 | 0,97 | 0,53 | 0,68 |
| **Graphe « définition »** | 36 | 28 | 25 | 0,89 | 0,69 | 0,77 |
| **Graphe « verbes »** | 151 | 171 | 136 | 0,76 | 0,9 | 0,84 |
| **Graphe « parenthèses »** | 143 | 229 | 84 | 0,36 | 0,58 | 0,44 |

***Tableau 1 : Précision-Rappel des graphes***

Nous avons tenté d'analyser les résultats obtenus avec chaque graphe.

### 7.1.1. Graphe du verbe « être »

Précision : 0,97 et rappel : 0,53

En d'autres termes, un énoncé repéré avec le *graphe être* est un énoncé définitoire dans 97% des cas. Ce très bon taux de précision signifie que le fichier de sortie d'Unitex est plutôt fiable. Avec un rappel de 0,53, ce même graphe identifie correctement 53% des énoncés définitoires présents dans le corpus de référence.

On peut considérer que le graphe est précis, par contre il est moyennement sensible.

Le silence s'explique par le fait que nous avons restreint le definiens à une maladie et défini une proximité ou une distance entre les différentes unités lexicales constitutives du PLS. De plus, le dictionnaire externe des noms de maladies n'est pas exhaustif, conduisant à l'absence de repérage des maladies dont les noms sont absents de la liste.

On peut ainsi repérer des faux négatifs, comme les exemples suivants dans lesquels :

- le definiens est un groupe nominal ou une phrase nominale :

  *ex1 : Le torticolis postural est **le reflet d'une contrainte utérine excessive** qui se manifeste à la naissance par une asymétrie globale du nouveau-né.*

  *ex2 : La maladie de Gorham-Stout est **un classique de la rhumatologie**, appartenant au groupe des malformations lymphatiques et se traduisant par une prolifération de cellules lymphatiques sous l'influence de facteurs de croissance (VEGF) conduisant à un envahissement progressif de l'os.*

- la distance entre le déterminant et le nom de maladie du definiens n'est pas nulle :

  *ex1 : La pyélonéphrite emphysémateuse (PNE) est une **forme rare d'**infection du haut appareil urinaire et dont le pronostic reste sévère.*

  *ex2 : Le torticolis musculaire congénital (TMC) est **la troisième** déformation néonatale en termes de fréquence après la dysplasie des hanches et le pied-bot varus équin.*

- la distance entre le verbe **être** et le nom de maladie du definiendum est supérieure à un token :

  *ex : Le syndrome lymphoprolifératif **avec auto-immunité (ALPS)** est une maladie rare liée à une anomalie de la voie apoptotique Fas/Fas-Ligand.*

- le nom de la maladie constituant le definiendum ou le definiens est absent du dictionnaire **<PATHO>**

  *ex1 : Le **syndrome d'Hajdu-Cheney** est une forme très rare de dysplasie cranio-squelettique associée à…*

  *ex2 : Le torticolis est une **attitude vicieuse**.*

Le faux positif est représenté par l'exemple suivant :

*ex : …la seule **anomalie cytogénétique** retrouvée est une délétion du bras long du chromosome 5, avec au moins une délétion de la bande q31_q33.*

### 7.1.2. Graphe des « parenthèses »

Précision : 0,36 et rappel : 0,58

Ce qui signifie qu'un énoncé repéré avec le *graphe du marqueur parenthétique* est un énoncé définitoire dans 36% des cas et que le graphe met en évidence 58% des énoncés définitoires présents dans le corpus annoté manuellement.

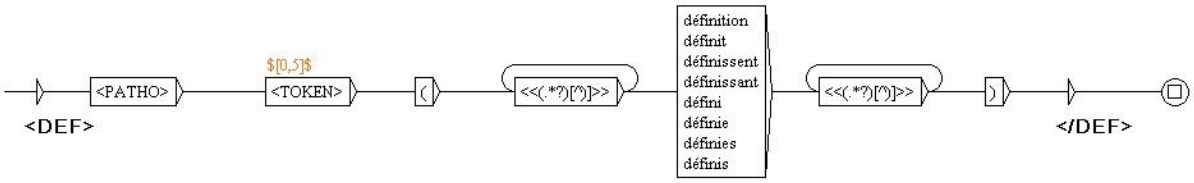
Comme l'on pouvait s'y attendre, le *graphe parenthétique* est celui qui a les moins bons résultats en termes de pertinence.

La méthode utilisée, consistant à prendre le contenu de toutes les parenthèses puis à retirer ce qui ne constitue pas des définitions, nous semble être l'une des meilleures puisqu'elle englobe toutes les définitions effectivement recherchées. Elle nécessite cependant la lecture d'un corpus assez volumineux en amont afin de rassembler le plus grand nombre de syntagmes *hors définitions* qui seront ensuite exclus des résultats.

Les exemples de syntagmes *hors définitions* déjà exclus des parenthèses sont : les acronymes, les décimaux sans unités, les mots rejetés, etc.

La partie du sous-graphe des parenthèses (Figure 19) entre les crochets verts est donc loin d'être exhaustive et explique les mauvais résultats en termes de pertinence.

Une autre approche possible consisterait à ne retenir dans les parenthèses, que les définitions potentielles amenées par le terme *définition* et ses dérivés grammaticaux *défini(es) par, définissant, etc*, ramenant ainsi la branche des parenthèses à un graphe beaucoup plus simple du type :




Un exemple de définition rapportée par ce graphe est :

> *... les caractéristiques de l'anémie réfractaire avec sidéroblastes en couronne (syndrome myélodysplasique) à une thrombocytose (**supérieure à 450 x 109/L, selon la définition de 2008**) et à une dysmégacaryopoïèse semblables à celle observée dans la thrombocytémie essentielle...*

Cependant, on perdrait toutes les définitions qui ne seraient pas contiguës aux dérivées du terme *définir*.

### 7.1.3. Graphe des « verbes »

Précision : 0,76 et rappel : 0,9

En d'autres termes, un énoncé repéré avec le *graphe des marqueurs verbaux* est un énoncé définitoire dans 76% des cas. Avec un rappel de 0,9 ce même graphe identifie correctement 90% des énoncés définitoires présents dans le corpus de référence.

On peut considérer que le graphe est relativement précis et que sa sensibilité est bonne.

En ce qui concerne la précision, les faux positifs sont représentés par une définition introduite par un verbe figurant dans la liste mais qui n'est pas en relation avec la maladie repérée par Unitex comme étant le definiendum ; il s'agit donc d'un problème de rattachement sémantique :

- Soit la maladie identifiée *definiendum* est en fait le complément d'un nom ayant la fonction de sujet du verbe figurant dans le graphe :

> *ex1 :* ***Le traitement de la chondrocalcinose*** *associe des mesures pharmacologiques et non pharmacologiques*

> *ex2 :* ***Le bilan standard d'un torticolis traumatique*** *associe un cliché de face, un cliché de profil et un cliché bouche ouverte.*

- Soit la maladie identifiée *definiendum* est à la fin d'une phrase et le verbe du graphe dans la phrase suivante, précédé de son sujet qui n'est pas une anaphore de la maladie :

> *ex : Le décès peut survenir par défaillance d'organe multiple ou en raison* **d'hémorragies massives** *(maladie identifiée par Unitex).* ***La prise en charge consiste en*** *celle d'un choc septique.*

Le silence s'explique par la distance entre le definiendum et le definiens limitée à 10 tokens. Pour quelques cas, la distance est plus grande.

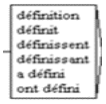
> *ex : Le diabète de grossesse* ***apparaît vers la fin du 2e et au 3e trimestre. Il*** *se manifeste par une augmentation du taux de sucre dans le sang qui survient uniquement lors de la grossesse.*

#### 7.1.4. Graphe « définition et définir à la voix active »

Précision : 0,89 et rappel : 0,69

En d'autres termes, un énoncé repéré avec le *graphe définition et définir à la voix active* est un énoncé définitoire dans 89% des cas. Avec un rappel de 0,69 ce même graphe identifie correctement 69% des énoncés définitoires présents dans le corpus de référence.

Le graphe apparait donc assez précis et moyennement sensible.

Le silence est dû à un nombre de tokens insuffisant entre les boites 

et 

comme le montre l'exemple suivant :

> *ex : L'American Thoracic Society a proposé en 2000 une définition* ***plus restrictive, en considérant comme sévère*** *un asthme recevant des doses élevées de corticoïdes inhalés.*

Les faux positifs concernent le marqueur nominal **définition** :

> *ex1 : La récente définition de ce groupe de syndromes myélodysplasiques ne permet pas encore de rapporter de données pronostiques précises.*

> *ex2 : La définition de l'asthme difficile est imparfaitement établie.*

> *ex3 : Cette proposition hybride rend cependant bien compte du malaise existant parmi le corps médical quant à la définition du syndrome métabolique à privilégier parmi celle de l'OMS et celle du NCEP-ATP III.*

Les énoncés définitoires composés du verbe **définir** ont tous été retrouvés avec le graphe mais leur faible nombre (5) ne permet pas de tirer de conclusion.

### *7.2. Projection des graphes sur d'autres corpus francophones*

La pertinence des graphes est évaluée sur un corpus du même domaine mais de volume plus important, ainsi que sur un corpus d'un autre domaine, l'astrophysique.

### 7.2.1. Corpus biomédical issu de l'archive ISTEX

#### *7.2.1.1. Présentation de la plateforme ISTEX*

ISTEX[27] (Initiative d'Excellence en Information Scientifique et Technique) est une plateforme numérique qui a pour objectif de permettre à la communauté ESR française de se connecter, via un accès en ligne, à une bibliothèque numérique regroupant 18,2 millions de documents couvrant tous les domaines de recherche.

Elle offre tous les moyens de recherche d'information et d'accès aux documents en texte intégral et fournit l'ensemble de ses services sous la forme d'une *API Web*[28] mais également via un *démonstrateur*[29] qui permet de se familiariser avec les formats et la syntaxe d'interrogation.

La plateforme ISTEX est un réservoir de publications scientifiques destiné à répondre aux besoins documentaires des documentalistes et chercheurs mais c'est également une ressource unique pour tous les chercheurs intéressés par la fouille de texte, le TAL et la recherche d'Information.

#### *7.2.1.2. Corpus de 10 000 références issues de l'archive ISTEX*

La démarche mise en œuvre pour créer ce corpus de grande taille se résume à l'écriture d'une équation de recherche basique « *(maladie OR *pathie) AND homme* » qui a permis de collecter 16 503 documents en texte intégral. Nous n'avons pas procédé aux étapes habituelles de nettoyage et de reformatage, indispensables à l'obtention de données homogènes, exploitables par les outils de TAL.

Nous sélectionnons les 10 000 premiers documents pour les charger dans Unitex.

#### *7.2.1.3. Projection des graphes sur le corpus issu d'ISTEX*

L'application de l'ensemble des graphes permet d'afficher 23 992 concordances.

##### 7.2.1.3.1. Graphe du verbe « être »

La projection du graphe du verbe **être** sur le corpus permet l'identification de 663 concordances. Bien que l'analyse quantitative des résultats n'ait pas été réalisée, ceux-ci sont globalement jugés pertinents.

Un certain nombre de faux-positifs apparaissent cependant, regroupant des énoncés définitoires dans lesquels la maladie mise en évidence n'est pas le sujet du verbe **être**, comme dans l'exemple suivant :

> *ex : Les anomalies cytogénétiques présentes chez un tiers des patients développant une **leucémie** sont le plus souvent une monosomie 7, une délétion partielle du bras long du chromosome 7 ou une duplication du bras long du chromosome 1.*

- La pathologie repérée par le graphe est *leucémie* puisque nous avons choisi de n'introduire qu'un token maximum entre la pathologie et le verbe **être**, alors que le sujet est *anomalies cytogénétiques*.

[27] https://www.istex.fr/
[28] https://api.istex.fr/
[29] https://demo.istex.fr/

- Les faux-positifs sont souvent dûs à la position de la pathologie ayant la fonction d'un complément du nom.
- Dans cet exemple la pathologie précédant le verbe **être** est l'hyperonyme *anomalies cytogénétiques*. Les pathologies positionnées après **être** sont des co-hyponymes.

On identifie un certain nombre de faux-négatifs :

- Le definiendum et le definiens sont absents du dictionnaire de **<PATHO>.**

> *ex : Le **délire de rédemption** est un **fantasme** qui nous est familier, il constitue très fréquemment le noyau de la paranoïa religieuse.*

- Le definiendum, bien que présent dans le dictionnaire de **<PATHO>**, n'est pas reconnu en raison de problèmes de conversion en format texte des documents chargés.

> *ex : L'insuffisance **r6nale** aigue (IRA) est un syndrome **caract~ris~** par la diminution rapide, en quelques heures ou quelques jours, du **d~bit** de filtration **glom~rulaire**.*

#### 7.2.1.3.2. Graphe des « parenthèses »

L'application au corpus du *graphe des parenthèses* révèle 18 507 concordances.

Comme l'on pouvait s'y attendre, le bruit est important.

L'analyse des résultats met en évidence de nouvelles unités lexicales, sans intérêt définitoire, qui pourront enrichir le dictionnaire des mots rejetés ou être rassemblées dans une nouvelle boite de rejet : ***n°, lettre de l'alphabet isolée faisant référence à un chapitre, page, p.x, ?, verbatim entouré de guillemets, illustration, adapté de, n = x***

#### 7.2.1.3.3. Graphe des « verbes »

L'application du *graphe des verbes* à ce même corpus ramène 4639 concordances.

Les faux positifs sont représentés par des phrases dont le verbe figure dans la liste mais qui n'est pas en relation avec la maladie repérée par Unitex comme étant le definiendum (problème de rattachement sémantique) et qui de plus n'introduit pas une définition :

> *ex1 : La recherche d'une éventuelle aberration chromosomique constitue l'indication majeure de l'examen cytogénétique prénatal.*

> *ex2 : L'inflammation allant du simple érythème à un véritable abcès centré par le molluscum correspond à un processus aseptique d'évolution favorable mais pouvant nécessiter une incision de décharge.*

#### 7.2.1.3.4. Graphe « définition et définir à la voix active »

La projection sur le corpus du *graphe définition et définir à la voix active* met en évidence 166 concordances.

Les faux positifs concernent essentiellement le marqueur nominal **définition** dont la position dans la phrase n'appelle pas un énoncé définitoire, comme le montrent les exemples suivants :

- Position au niveau d'un titre de chapitre :

> *ex : Définition, physiopathologie de la bronchite aiguë de l'adulte sans pathologie respiratoire préexistante.*

- Position en fin de phrase, ne représentant pas l'objet principal du discours :

*ex : L'originalité de ce cas clinique repose sur l'association d'un kyste osseux solitaire à une drépanocytose mais aussi à sa situation maxillaire et à la symptomatologie associée amenant une révision des critères de définition de ces pseudokystes.*

#### *7.2.1.4. Repérage d'autres marqueurs d'énoncés définitoires*

Ce corpus plus volumineux nous permet d'enrichir notre liste de marqueurs d'énoncés définitoires. Bien que nous n'ayons pas eu le temps de faire une analyse approfondie, nous avons pu tout de même pu mettre en évidence un petit nombre de nouveaux marqueurs, notamment verbaux, mais également des expressions, ainsi que la locution adverbiale « *c'est-à-dire* » comme le montrent les exemples suivants :

**Expressions « on appelle, on nomme, on entend par »**

*ex 1 : **On appelle** arythmie sinusale une accélération de la fréquence cardiaque à l'inspiration suivie d'un ralentissement à l'expiration.*

*ex 2 : **On nomme** botulisme l'ensemble des symptômes provoqués chez l'homme et chez les animaux par l'ingestion d'aliments contenant un microbe spécifique qui peut vivre en l'absence d'air.*

*ex 3 : Selon le DSM IV, **on entend par** jeu pathologique une « pratique inadaptée, persistante et répétitive du jeu qui entraîne une répercussion sur la vie familiale personnelle ou professionnelle ».*

L'application au corpus du graphe des expressions « *on appelle, on nomme, on désigne par, on entend par* » met en évidence 17 concordances, toutes pertinentes.

**Marqueurs verbaux**
**Réunir**

*ex : Le syndrome de Peutz-Jeghers **réunit** une polypose rectale et gastrointestinale, une lentiginose péri-orificielle et des fibromes ovariens.*

**Rassembler**

*ex : Les maladies hématologiques rares **rassemblent** la plupart des maladies du sang, de la moelle osseuse, des ganglions lymphatiques et de la coagulation.*

**Etre issu de**

*ex : Généralement, les émergences virales **sont issues de** la transmission de virus de l'animal à l'Homme : on parle alors de maladies zoonotiques ou zoonoses,..*

**Résulter de**

*ex : L'ostéogenèse imparfaite ou le syndrome de Marfan **résulte de** défauts qui touchent principalement une seule molécule matricielle, en l'occurence respectivement le collagène 1 et la fibrilline-1…*

**Etre classé**

*ex : L'autisme **est classé** parmi les troubles envahissants du développement.*

**Etre formé de**

*ex : L'aliénation **est formée de** troubles mentaux non fébriles.*

**S'intégrer à, aux**

*ex : La dysgraphie **s'intègre aux** autres pathologies* dys- *auxquelles elle est fréquemment associée (dyslexie, dyspraxie, dysorthographie).*

**S'organiser**

*ex : Dans les paraphrénies, le délire est riche, non structuré, et **s'organise** autour de l'imaginaire.*

**Se décomposer en**

*ex : les troubles bipolaires **se décomposent en** deux phases. La phase d'excitation est caractérisée par une hyperactivité, une euphorie, une volubilité mais aussi des troubles de l'appétit, une réduction du besoin de sommeil, une irritabilité…*

**Se diviser en**

*ex : L'ataxie spinocérébelleuse type 3 (ASC3) **se divise en** 3 formes…*

**Résulter de**

*ex : L'angiopathie amyloide cérébrale **résulte de** dépôts amyloïdes (protéine de conformation anormale) dans la paroi des petits et moyens vaisseaux cérébraux ce qui les fragilise.*

**Ressembler à**

*ex : Hankey et al, puis Calabrese et al ont proposé d'appeler cette affection « angiopathie cérébrale aigue bénigne », dont une forme particulière est l'angiopathie aigue du post-partum B qui **ressemble à** l'angiopathie cérébrale aiguë bénigne et qui est habituellement d'évolution courte et de très bon pronostic.*

**Englober**

*ex : Étymologiquement, le terme de pneumopathie **englobe** toutes les maladies des poumons. Toutefois, ce terme est habituellement utilisé pour évoquer une inflammation pulmonaire.*

**C'est-à-dire**

Cette locution adverbiale est le marqueur d'une reprise interprétative. « *Dans toute réalisation de la formule (A, c'est-à-dire B), le terme B doit pouvoir être de quelque manière pris comme une interprétation du terme A* ». (Murat 1987)

*ex : … lorsque la neutropénie est sévère, **c'est à dire** inférieure à 500 PNN/mm $^3$ de sang*

#### *7.2.1.5. Conclusion de l'application des graphes sur le corpus de 10 000 documents*

L'application à ces 10 000 documents des graphes construits à partir du corpus de 48 a permis de valider leur utilité et leur transportabilité pour la recherche d'énoncés définitoires dans le domaine biomédical, sans que leur pertinence n'ait été quantitativement évaluée.

De nouveaux marqueurs définitoires, notamment de type verbal, ont été mis en évidence et vont enrichir le dictionnaire des **<VERBES>**.

En ce qui concerne le graphe des parenthèses, d'autres mots à exclure ont été répertoriés et vont être ajoutés à la liste existante **<MR>**.

Des termes de maladies présentes dans ce corpus et non encore référencées dans **<PATHO>** vont y trouver leur place, ainsi que leur forme fléchie.

On remarque que le *graphe du verbe être* ramène des faux positifs lorsque le sujet du verbe est suivi d'un complément du nom comportant un nom de maladie.

### 7.2.2. Corpus d'astrophysique

Un test visant à déterminer l'applicabilité du graphe général, construit précédemment pour le domaine biomédical, à un autre domaine, a été effectué en prenant l'astrophysique pour témoin.

Le monde de la recherche en astrophysique publiant essentiellement en langue anglaise, nous n'avons pu constituer qu'un corpus de 13 articles francophones, extraits d'ISTEX et du portail d'astronomie de Wikipédia[30] . Les publications scientifiques ont été parcourues par un expert du domaine afin d'en extraire manuellement les définitions.

Le seul changement apporté au graphe est le remplacement du dictionnaire des pathologies par un dictionnaire d'astrophysique constitué de termes issus d'un thésaurus (Figure 30) en cours de construction pour le service « Ingénierie terminologique » de l'Inist-CNRS et qui sera prochainement exposé sur Loterre.

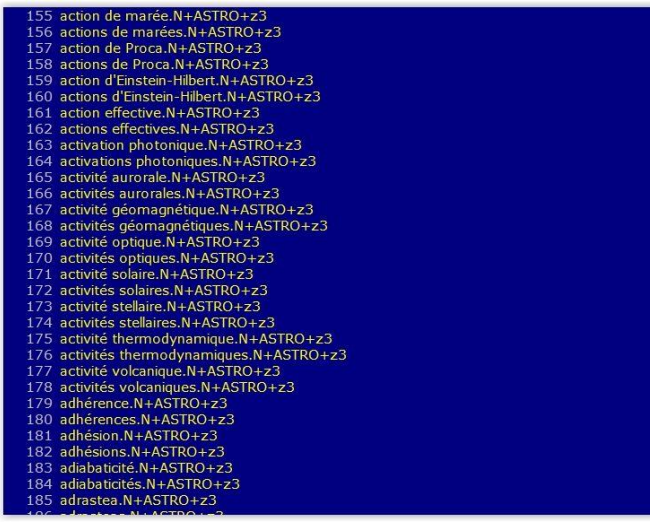



***Figure 30 : Extrait du thésaurus d'astrophysique***

Ces termes sont issus d'un vocabulaire d'indexation de l'ancienne base bibliographique PASCAL de l'Inist-CNRS, enrichis de termes extraits du thésaurus d'Astronomie de la NASA[31] et du Thésaurus International d'Astronomie (Unified Astronomy Thesaurus)[32].

En ce qui concerne l'application des graphes sur ce corpus d'astrophysique, le petit nombre d'articles nous permet d'exposer plus en détail l'ensemble des résultats.

#### *7.2.2.1. Application des graphes au corpus d'astrophysique*

Les résultats sont consignés dans le tableau suivant (Tableau 2) :

[30] https://fr.wikipedia.org/wiki/Portail:Astronomie
[31] https://www.sti.nasa.gov/nasa-thesaurus/
[32] https://astrothesaurus.org/

| Graphe | Nombre de définitions attendues (corpus de référence) | Nombre total de définitions trouvées avec Unitex | Nombre de définitions pertinentes trouvées avec Unitex |
|---|---|---|---|
| **Graphe « être »** | 21 | 8 | 4 |
| **Graphe « parenthèses »** | 0 | 65 | 0 |
| **Graphe « verbes »** | 1 | 16 | 0 |
| **Graphe « définition »** | 10 | 7 | 7 |
| **Graphe « expressions »** | 2 | 0 | 0 |

*Tableau 2 : Résultats de l'application des graphes au corpus d'astrophysique*

Une interprétation des résultats a été réalisée pour chaque graphe.

#### 7.2.2.1.1. Graphe du verbe « être »

**Définitions attendues :**

***1.*** *Une étoile variable de type AM Canum Venaticorum, ou étoile variable de type AM CVn, est un type rare d'étoile variable cataclysmique nommé d'après l'étoile prototype, AM Canum Venaticorum.*

***2.*** *Une exoplanète, ou planète extrasolaire, est une planète située en dehors du Système solaire.*

***3.*** *les objets avec une vraie masse en deçà de la masse limite permettant la fusion thermonucléaire du deutérium (actuellement calculée comme valant 13 fois la masse de Jupiter pour des objets de métallicité solaire) qui orbitent autour d'étoiles ou de rémanents stellaires sont des « planètes » (peu importe comment ils se sont formés). La masse/taille minimale requise pour qu'un objet extrasolaire soit considéré comme une planète devrait être la même que celle utilisée dans notre Système solaire*

***4.*** *les objets substellaires avec des masses vraies au-delà de la masse limite permettant la fusion thermonucléaire du deutérium sont des « naines brunes », peu importe comment ils se sont formés ou où ils se trouvent*

***5.*** *les objets flottant librement dans de jeunes amas stellaires avec des masses en deçà de la masse limite permettant la fusion thermonucléaire du deutérium ne sont pas des « planètes », mais sont des « sous-naines brunes » (ou quelque nom qui soit plus approprié)*

***6.*** *Les (étoiles) variables de type Mira sont une classe d'étoiles variables, caractérisées par des couleurs très rouges, des périodes de pulsation supérieures à 100 jours, et des amplitudes de luminosité supérieures à une magnitude.*

***7.*** *Ce sont des étoiles géantes rouges se trouvant dans les dernières étapes de leur évolution stellaire (la branche asymptotique des géantes rouges) qui finiront par expulser leur enveloppe externe en une nébuleuse planétaire et par devenir des naines blanches en quelques millions d'années.*

***8.*** *Cette classe d'étoile est nommée en référence à l'étoile Mira (o Cet).*

***9.*** *En astronomie, une planète naine est un objet céleste du Système solaire de classe intermédiaire entre une planète et un petit corps du Système solaire.*

***10.*** *Plus précisément, l'UAI explicite qu'une planète naine est « un corps céleste qui (a) est en orbite autour du Soleil, (b) a une masse suffisante pour que sa gravité l'emporte sur les forces*

*de cohésion du corps solide et le maintienne en équilibre hydrostatique, sous une forme presque sphérique, (c) n'a pas éliminé tout corps susceptible de se déplacer sur une orbite proche, (d) n'est pas un satellite ».*

***11.** (1) une planète est un corps céleste qui (a) est en orbite autour du Soleil, (b) a une masse suffisante pour que sa gravité l'emporte sur les forces de cohésion du corps solide et le maintienne en équilibre hydrostatique, sous une forme presque sphérique, (c) a éliminé tout corps susceptible de se déplacer sur une orbite proche ;*

***12.** (2) une « planète naine » est un corps céleste qui (a) est en orbite autour du Soleil, (b) a une masse suffisante pour que sa gravité l'emporte sur les forces de cohésion du corps solide et le maintienne en équilibre hydrostatique, sous une forme presque sphérique, (c) n'a pas éliminé tout corps susceptible de se déplacer sur une orbite proche, (d) n'est pas un satellite*

***13.** Les plutoïdes sont des corps célestes en orbite autour du Soleil à un demi-grand-axe plus grand que celui de Neptune qui ont une masse suffisante pour que leur propre gravité surpasse les forces rigides du corps donc leur permettant d'avoir une forme en équilibre hydrostatique (presque sphériques), et qui n'ont pas nettoyé le voisinage autour de leur orbite.*

***14.** La planète naine Cérès n'est pas un plutoïde puisqu'elle est située dans la ceinture d'astéroïdes entre Mars et Jupiter. Les connaissances scientifiques actuelles laissent à croire que Cérès est le seul objet de sa catégorie.*

***15.** Le paradoxe EPR, abréviation de Einstein-Podolsky-Rosen, est une expérience de pensée, élaborée par Albert Einstein, Boris Podolsky et Nathan Rosen, et présentée dans un article de 1935, dont le but premier était de réfuter l'interprétation de l'école de Copenhague de la physique quantique.*

***16.** En astronomie, le milieu interstellaire (en anglais, interstellar medium ou ISM) est la matière qui, dans une galaxie, remplit l'espace entre les étoiles et se fond dans le milieu intergalactique environnant.*

***17.** La classification de Harvard est celle qui attribue un type spectral à une étoile, et correspond globalement à une échelle de température.*

***18.** La classification de Yerkes est celle qui attribue une classe de luminosité à une étoile, et correspond globalement à une échelle de rayon (voir loi de Stefan-Boltzmann) pour une température donnée.*

***19.** Une (étoile) supergéante est un type d'étoile très massive, d'environ 10 à 70 masses solaires.*

***20.** Un quasar (source de rayonnement quasi-stellaire, quasi-stellar radiosource en anglais, ou plus récemment « source de rayonnement astronomique quasi-stellaire », quasi-stellar astronomical radiosource) est un noyau de galaxie extrêmement lumineux (noyau actif).*

***21** Bien qu'il y ait d'abord eu une certaine controverse sur la nature de ces objets jusqu'au début des années 1980, il existe maintenant un consensus scientifique selon lequel un quasar est la région compacte entourant un trou noir supermassif au centre d'une galaxie massive.*

**Définitions trouvées avec Unitex :**

***1.** Si **l'étoile hôte est une étoile simple**, on peut toujours considérer qu'elle possède un « A » dans sa désignation, bien qu'il ne soit habituellement pas écrit.*

***2.** Les deux causes principales **d'incertitudes sont la convection** dans les modeles de structure interne d'etoiles et les effets de la metallicite.*

***3.*** *Les derniers événements subis par ces* ***photons sont des diffusions*** *par les électrons libres à la fin de la recombinaison ...*

***4.*** *(1) une* ***planète est un corps céleste*** *qui (a) est en orbite autour du Soleil ...*

***5.*** *Une exoplanète, ou* ***planète extrasolaire, est une planète*** *située en dehors du Système solaire.*

***6.*** *(2) une* ***« planète naine » est un corps céleste*** *qui (a) est en orbite autour du Soleil ...*

***7.*** *En astronomie, une* ***planète naine est un objet céleste du Système solaire*** *de classe intermédiaire entre une planète et un petit corps du Système solaire.*

***8.*** *Un quasar (source de rayonnement quasi-stellaire, quasi-stellar radiosource en anglais, ou plus récemment « source de rayonnement astronomique quasi-stellaire », quasi-stellar astronomical* ***radiosource) est un noyau de galaxie*** *extrêmement lumineux (noyau actif).*

**Définitions pertinentes trouvées avec Unitex :**

(4) + (5) + (6) + (7)

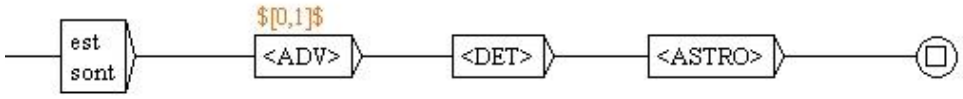
Les parties en gras dans les définitions correspondent aux déclencheurs du graphe.

La définition (8), bien que pertinente, n'a pas été retenue car il s'agit d'une coïncidence de proximité entre termes consécutifs présents dans le thésaurus (radiosource et noyau de galaxie) alors que c'est le quasar qui est un noyau de galaxie en non la radiosource.

Les phrases restantes trouvées par Unitex (1), (2), (3) et (8) collent effectivement avec le graphe mais ne sont pas des définitions, simplement des faits d'observation.

Quand on retire les 4 définitions pertinentes des 21 définitions attendues dans le corpus, il reste en définitive 17 définitions non détectées (faux négatifs) par Unitex. Les raisons de cette absence de détection sont les mêmes que pour le corpus biomédical, à savoir :

- la proximité au niveau du definiens du déterminant <DET> et du terme figurant dans le dictionnaire <ASTRO> ;

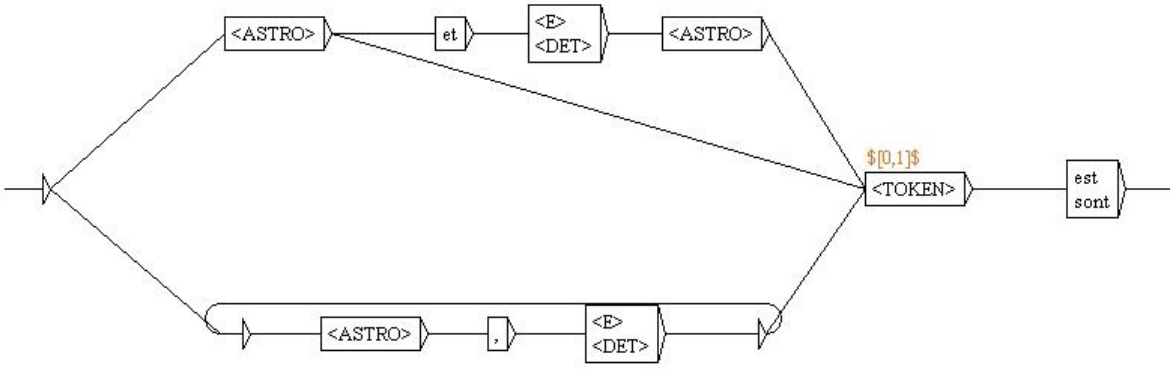



***Exemples :***

***1.*** *Une étoile variable de type AM Canum Venaticorum, ou étoile variable de type AM CVn, est un* ***type rare*** *d'étoile variable cataclysmique nommé d'après l'étoile prototype, AM Canum Venaticorum.*

***6.*** *Les (étoiles) variables de type Mira sont une* ***classe*** *d'étoiles variables, caractérisées par des couleurs très rouges, des périodes de pulsation supérieures à 100 jours, et des amplitudes de luminosité supérieures à une magnitude.*

- le nombre de tokens insuffisant entre le dictionnaire **<ASTRO>** et le verbe **être** ;




***Exemples :***

***4.*** *les objets substellaires* ***avec des masses vraies au-delà de la masse limite permettant la fusion thermonucléaire du deutérium*** *sont des « naines brunes », peu importe comment ils se sont formés ou où ils se trouvent*

***20.*** *Un quasar* ***(source de rayonnement quasi-stellaire, quasi-stellar radiosource en anglais, ou plus récemment « source de rayonnement astronomique quasi-stellaire », quasi-stellar astronomical radiosource)*** *est un noyau de galaxie extrêmement lumineux (noyau actif)*

- l'absence de certains termes dans le thésaurus ;

***Exemples :***

***3.*** *Les* ***plutoïde****s sont des corps célestes en orbite autour du Soleil à un demi-grand-axe plus grand que celui de Neptune qui ont une masse suffisante pour que leur propre gravité surpasse les forces rigides du corps donc leur permettant d'avoir une forme en équilibre hydrostatique (presque sphériques), et qui n'ont pas nettoyé le voisinage autour de leur orbite.*

***17.*** *La* ***classification de Harvard*** *est celle qui attribue un type spectral à une étoile, et correspond globalement à une échelle de température.*

***18.*** *La* ***classification de Yerkes*** *est celle qui attribue une classe de luminosité à une étoile, et correspond globalement à une échelle de rayon (voir loi de Stefan-Boltzmann) pour une température donnée.*

***21.*** *Bien qu'il y ait d'abord eu une certaine controverse sur la nature de ces objets jusqu'au début des années 1980, il existe maintenant un consensus scientifique selon lequel un quasar est la* ***région compacte*** *entourant un trou noir supermassif au centre d'une galaxie massive.*

#### 7.2.2.1.2. Graphe des « parenthèses »

L'absence de définitions entre parenthèses dans ce corpus d'astronomie n'a pas permis de tester cette partie du graphe.

Cependant, il est à prévoir de modifier fondamentalement cette branche qui ramènerait la majorité des formules mathématiques très fréquentes dans les articles d'astronomie et quasi-inexistantes en médecine.

Quelques exemples de parenthèses contenant des formules mathématiques ramenées par cette branche testée sur des articles d'astronomie extérieurs à ce corpus :

*Séries de Fourier : soit f* ***(x)*** *une fonction périodique …*

*Structure interne du Soleil : avec ρ = M /* ***(4/3 π R3)*** *où M est la masse du Soleil et R le rayon du Soleil, …*

*Equation du mouvement : ρ* ***(∂v/∂t + v.grad v)*** *= - grad P + ρg devient au premier ordre*

*Opérateur Laplacien Δ =* ***(∂²/∂x², ∂²/∂y², ∂²/∂z²)***

On constate immédiatement qu'il est en effet impossible de prévoir la syntaxe d'écriture des formules mathématiques, donc une nouvelle réflexion s'impose pour éliminer ce bruit des résultats.

#### 7.2.2.1.3. Graphe des « verbes »

**Définitions attendues :**

***1.*** *Il s'agit du fameux décalage vers le rouge cosmologique, noté z, qui est défini par le fait que la longueur d'onde observée λ obs , d'une radiation électromagnétique émise à la longueur d'onde λ, par une source distante, est supérieure à λ, dans le rapport (1 + z).*

**Définitions trouvées avec Unitex :**

***1.*** *Ces valeurs reposent sur un modèle cosmologique à constante cosmologique nulle, avec une constante de Hubble de 55 km/s/Mpc et une densité de l'Univers de 0,3 fois la densité critique* ***(la densité critique étant celle qui correspond à*** *un espace plat).*

***2.*** *la quantité* ***d'eau en hydrodynamique correspond*** *à la norme intégrée de la fonction d'onde de la mécanique quantique, qui reste constante, et les « motifs » qui se dessinent à la surface de l'eau correspondent aux mondes.*

***3.*** *Or la base des vecteurs de* ***l'espace de Hilbert qui correspondent*** *à des états macroscopiques bien définis ...*

***4.*** *Ceci modifie la structure de l'****étoile, ce qui se manifeste par*** *un changement de période.*

***5.*** *Les* ***étoiles les plus anciennes de la Galaxie présentent des*** *caractéristiques qui les différencient fortement des étoiles communes, comme notre Soleil.*

***6.*** *Certaines* ***étoiles présentent des*** *raies d'absorption et des raies en émission à des moments différents.*

***7.*** *Les (étoiles) variables de type Mira sont une classe d'****étoiles variables, caractérisées par*** *des couleurs très rouges ...*

***8.*** *La* ***poussière interstellaire se présente sous la forme de*** *grains extrêmement fins, dont la taille typique est de l'ordre d'une fraction de micron.*

***9.*** *Dans les deux cas, formation des* ***galaxies et des trous noirs correspondent à*** *des structures propres à la dynamique de ces théories*

***10.*** *Elle et son cortège de* ***galaxies satellites font partie du*** *Groupe local, lui-même rattaché au superamas de la Vierge appartenant lui-même à Laniakea.*

***11.*** *Le* ***milieu interstellaire se compose de*** *plusieurs phases, selon l'état de la matière (soit ionique, atomique ou moléculaire), sa température (millions de kelvins, milliers de kelvins ou dizaines de kelvins) et sa densité.*

***12.*** *Comme pour les* ***novas naines, l'état haut correspond*** *à un disque plus chaud avec un hélium ionisé optiquement épais ...*

***13.*** *Tres proches du* ***quasar, ils font partie de*** *son environnement.*

***14.*** *Il y a en entre une et deux par mille étoiles de population I, c'est-à-dire de composition proche de celle du* ***Soleil, qui constituent*** *l'écrasante majorité des étoiles visibles à l'oeil nu.*

***15.*** *De façon cohérente, les « planètes » du* ***Système solaire au sens de la définition précédente sont les huit objets définis comme*** *« planètes » au sens de la résolution du 24 août 2006.*

**Définitions pertinentes trouvées avec Unitex :**

La définition attendue n'a pas été détectée par Unitex. En effet, bien que l'expression verbale « *défini par* » soit présente dans la phrase, son déclencheur « *décalage vers le rouge cosmologique* » est absent du dictionnaire **<ASTRO>** (on y trouve par contre son préférentiel « *déplacement vers le rouge cosmologique* », d'où la nécessité d'enrichir en synonymes tous les termes d'un dictionnaire).

Lorsqu'on analyse les définitions potentielles trouvées par Unitex (avec les parties en gras correspondant aux déclencheurs du graphe), on voit que les expressions verbales du dictionnaire **<VERBES>** ne ramènent quasiment aucune définition contrairement au cas de la médecine. D'autres types d'expressions verbales devront donc être recherchés dans les articles d'astronomie et plus généralement de sciences exactes.

7.2.2.1.4. Graphe « définition et définir à la voix active »


**Définitions attendues :**

***1.*** *D'après la définition même des étoiles de la population III, celles-ci ne devraient contenir que les éléments produits dans la synthèse primordiale, avant que les supernovae n'aient pollué le milieu cosmique.*

***2.*** *La définition officielle d'une planète adoptée en août 2006 par l'Union astronomique internationale (UAI) ne concerne que les objets du Système solaire et ne s'applique pas aux exoplanètes*

***3.*** *À l'heure actuelle, la seule définition de l'UAI qui concerne les exoplanètes est une définition de travail donnée en 2002 et modifiée en 2003.*

***4.*** *Cette définition, plus générale et qui concerne toutes les planètes, y compris celles du Système solaire, contient les critères suivants*

***5.*** *De façon cohérente, les « planètes » du Système solaire au sens de la définition précédente sont les huit objets définis comme « planètes » au sens de la résolution du 24 août 2006.*

***6.*** *Et de la même façon, une « planète extrasolaire » est alors définissable comme une planète — toujours au sens de la définition générale juste au-dessus et uniquement celui-ci — située hors du Système solaire.*

***7.*** *Une définition alternative considère que les planètes devraient être distinguées des naines brunes sur la base de leur formation.*

***8.*** *La définition d'une planète repose alors sur son caractère sphérique, chose notamment défendue par Alan Stern, mais implique que le nombre de planètes pourrait croître rapidement avec les progrès de l'observation spatiale, les objets en équilibre hydrostatiques connus étant alors déjà estimés à une cinquantaine par Mike Brown.*

***9.*** *L'Union astronomique internationale, organisation chargée de la nomenclature astronomique, définit précisément un système avec trois catégories de corps célestes dans sa résolution no 5 adoptée le 24 août 2006*

***10.*** *Selon la définition, un corps céleste doit avoir « une masse suffisante pour que sa gravité l'emporte sur les forces de cohésion du corps solide et le maintienne en équilibre hydrostatique, sous une forme presque sphérique », mais les dimensions auxquelles un objet atteint un tel état ne sont pas spécifiées*

**Définitions trouvées avec Unitex :**

***1.*** ***Une définition alternative considère que les planètes*** *devraient être distinguées des naines brunes sur la base de leur formation.*

***2.*** *L'Union astronomique internationale (UAI) crée rapidement un comité de 19 spécialistes présidé par Iwan Williams — et composé d'experts tels que Michael A'Hearn, Alan Boss, Edward L. G. Bowell, Dale Cruikshank, Brian G. Marsden et Alan Stern — afin d'aboutir à une* ***définition d'une planète*** *...*

***3.*** *La définition d'une planète repose alors sur son caractère sphérique, chose notamment défendue par Alan Stern, mais implique que le nombre de planètes pourrait croître rapidement avec les progrès de l'observation spatiale ...*

***4.*** *D'après la* ***définition même des étoiles de la population III****, celles-ci ne devraient contenir que les éléments produits dans la synthèse primordiale, avant que les supernovae n'aient pollué le milieu cosmique.*

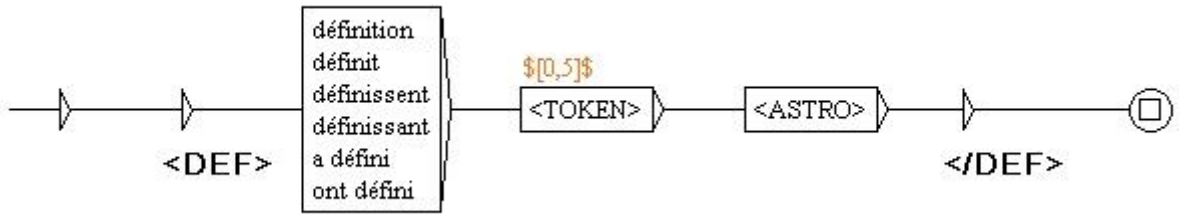
***5.*** *La* ***définition officielle d'une planète*** *adoptée en août 2006 par l'Union astronomique internationale (UAI) ne concerne que les objets du Système solaire et ne s'applique pas aux exoplanètes.*

***6.*** *D'autres planétologues ne sont pas d'accord avec une* ***définition permettant une inflation rapide du nombre de planètes****, suggérant que les facteurs dynamiques et l'environnement de l'objet devaient être considérés.*

***7.*** *Selon la* ***définition, un corps céleste*** *doit avoir « une masse suffisante pour que sa gravité l'emporte sur les forces de cohésion du corps solide et le maintienne en équilibre hydrostatique, sous une forme presque sphérique ».*

**Définitions pertinentes trouvées avec Unitex** :

L'ensemble des définitions trouvées avec Unitex est pertinent. Certaines d'entre elles (2), (5) et (6) peuvent être considérées comme des compléments d'informations rapportées aux définitions.




Qu'en-est-il des définitions restantes trouvées manuellement ?

> ***1.*** *À l'heure actuelle, la seule définition de l'UAI qui concerne les* ***exoplanètes est une définition*** *de travail donnée en 2002 et modifiée en 2003 ...*
>
> ***2.*** *Cette* ***définition, plus générale et qui concerne toutes les planètes****, y compris celles du Système solaire, contient les critères suivants ...*
>
> ***3.*** *De façon cohérente, les «* ***planètes » du Système solaire au sens de la définition*** *précédente sont les huit objets définis comme « planètes » au sens de la résolution du 24 août 2006.*
>
> ***4.*** *Et de la même façon, une « planète extrasolaire » est alors définissable comme une* ***planète — toujours au sens de la définition*** *générale juste au-dessus et uniquement celui-ci — située hors du Système solaire.*

Dans les exemples (1), (3) et (4), le terme **définition** suit son déclencheur figurant dans le dictionnaire **<ASTRO>** au lieu de le précéder. Dans l'exemple (2), il y a plus de 5 tokens entre le terme **définition** et le terme *planète* contenu dans **<ASTRO>**.

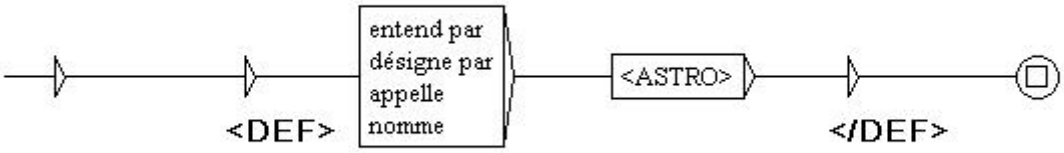
#### 7.2.2.1.5. Graphe des « expressions »

**Définitions attendues :**

> ***1.*** *Communément, on appelle « planète extrasolaire » toute planète orbitant autour d'une autre étoile que le Soleil.*
>
> ***2.*** *La frontière entre l'astrosphère d'une étoile (en particulier, pour le Soleil l'héliosphère) et le milieu interstellaire avoisinant se nomme l'astropause (en particulier, l'héliopause).*

**Définitions trouvées avec Unitex :**

Aucune définition n'a été détectée avec Unitex sur cette branche.




La définition (1) d'une planète extrasolaire n'a pas été détectée en raison de la présence de crochets entourant le terme.

La définition (2) est plus subtile dans le sens grammatical. En effet, habituellement on ne place pas d'article entre les expressions (entend par, désigne par, appelle et nomme) et le terme recherché dans le dictionnaire ; or dans l'exemple considéré il y a un article.

Il reste les définitions suivantes trouvées manuellement mais qui ne se projettent pas dans le graphe général :

***1.*** *Apres amplification, l'effondrement de surdensites initiales dans le fluide primordial est suppose conduire a la formation de halos massifs, au sein desquels naissent les galaxies et les amas. Plus connu sous le nom de CDM pour Cold Dark Matter, ce modele est forme a 90 % de mat&e sombre inerte (Blumenthal et al., 1984), n'interagissant que peu ou pas avec la matière.*

***2.*** *Cette variation de luminosité a une amplitude plus importante que la variation liée à la période orbitale et est appelée superhump.*

***3.*** *Par exemple, les planètes qui orbitent le pulsar PSR 1257+12 furent à l'origine nommées avec des capitales, en commençant à A (laissant l'étoile sans suffixe), plutôt qu'avec des minuscules.*

***4.*** *Les étoiles AGB pulsantes telles que les variables Mira subissent une fusion en coquilles alternant l'hydrogène et l'hélium, ce qui produit des convections profondes périodiques appelées dredge-ups.*

***5.*** *À partir des années 1990, les astronomes découvrent de nouveaux objets dans la même région de l'espace que Pluton, maintenant connue sous le nom de ceinture de Kuiper, et certains encore plus éloignés.*

***6.*** *(3) tous les autres objets en orbite autour du Soleil sont appelés « petits corps du Système Solaire ».*

***7.*** *La Voie lactée, aussi nommée la Galaxie (avec une majuscule), est une galaxie spirale barrée qui comprend de 200 à 400 milliards d'étoiles et au minimum 100 milliards de planètes.*

***8.*** *La partie du ciel occultée par la Voie lactée est appelée la zone d'évitement.*

***9.*** *La théorie d'Everett, appelée parfois théorie des états relatifs, théorie des mondes multiples (many-worlds) ou théorie des observateurs multiples (many-minds), est une formulation de la mécanique quantique fondée uniquement sur l'évolution déterministe de l'équation de Schrödinger, qui, appliquée à l'univers entier, régit son état quantique.*

***10.*** *Pour toutes ces raisons, la théorie d'Everett est appelée parfois théorie des états relatifs, théorie des mondes multiples (many-worlds) ou théorie des observateurs multiples (many-minds).*

***11.*** *Le Soleil et la plupart des étoiles du voisinage solaire ont une composition chimique qui diffère fortement de celle issue du Big Bang, que nous appelons composition « primordiale ».*

***12.*** *Nous appellerons nucléosynthèse stellaire (NS), cette deuxième exposition de la matière, au cœur des étoiles, à des conditions de température et de densité, capable de former de nouveaux éléments.*

***13.*** *Au début du XXe siècle, Ejnar Hertzsprung et Henry Norris Russell étudièrent la relation entre la luminosité et la température de couleur des étoiles. Ils arrivèrent indépendamment à la conclusion que la majorité des étoiles se trouvent dans une région précise d'un graphique luminosité-température. On désigne maintenant un tel graphique « diagramme de Hertzsprung-Russell » (ou plus simplement « diagramme HR »).*

***14.*** *Des quasars plus récents montrent qu'ils n'ont aucune région d'absorption mais plutôt des spectres contenant une zone avec un pic connu sous le nom de forêt Lyman-α.*

***15.*** *Cette variation de luminosité a une amplitude plus importante que la variation liée à la période orbitale et est appelée superhump.*

L'analyse de ces définitions confirme les conclusions des résultats précédents sur chacune des branches, à savoir :

- il y a des modifications à apporter à chacune des branches (rajout d'éléments grammaticaux et modification du nombre des tokens) ;
- il faut également rajouter des éléments de symétrie dans certaines branches (définition avant dictionnaire + définition après dictionnaire) ;
- il est indispensable d'enrichir le thésaurus d'astronomie en synonymes ;
- il faut également compléter le dictionnaire des verbes (ex : connu(es), nommé(es)) et trouver de nouvelles expressions verbales dans les articles d'astronomie.

#### *7.2.2.2. Conclusion de l'application des graphes sur le corpus d'astrophysique*

Les tests sur ce corpus d'astrophysique montrent que les graphes construits à partir du corpus de médecine n'apportent pas d'aussi bons résultats dans ce domaine, avec notamment trop de faux négatifs. Ils ne sont donc pas applicables quel que soit le domaine.

Il va falloir les adapter aux particularités d'écriture des articles dans les domaines des sciences exactes (par exemple les parenthèses ramènent le plus souvent des formules mathématiques, inexistantes en médecine).

### *7.3. Relations lexico-sémantiques au sein des énoncés définitoires*

L'analyse des contextes définitoires dans les textes spécialisés présente un intérêt supplémentaire pour le spécialiste en ingénierie terminologique, celui d'accéder à des relations lexico-sémantiques d'hyperonymie (Borillo 1996), de synonymie et de méronymie, utiles pour la construction et la structuration sémantique des terminologies.

Un énoncé définitoire peut être considéré comme une portion de texte riche en connaissances car contenant des informations conceptuelles permettant de préciser le sens du terme défini.

L'appellation « *Contextes Riches en Connaissances* » (CRC) a été proposée par Meyer (Meyer 2001) et désigne les contextes permettant de repérer, grâce à des éléments lexico-syntaxiques, des relations entre plusieurs termes. Il s'agit de portions de textes composées de termes d'un domaine spécialisé et de marqueurs explicitant les relations entre ces termes (Hmida 2018).

Lefeuvre et Condamines (Lefeuvre 2017) proposent la ressource « *MAR-REL* » (MARqueurs de RELations) qui est une base de marqueurs de relations conceptuelles pour la détection de « *Contextes Riches en Connaissances* », constituée lors du projet ANR CRISTAL[33] (Contextes RIches en connaissanceS pour la TrAduction terminoLogique). Ces marqueurs permettent d'identifier en corpus spécialisé trois types de relations, considérées comme structurantes et universelles, apportant des éléments de connaissance sur les termes d'un domaine. Il s'agit des relations d'hyperonymie, de méronymie et de cause.

[33] https://anr.fr/Projet-ANR-12-CORD-0020

En ce qui concerne notre étude, les relations sémantiques mises en évidence sont essentiellement des relations hiérarchiques de type hyperonymie/hyponymie, des relations de méronymie et de synonymie.

### 7.3.1. Relations hiérarchiques

#### *7.3.1.1. Hyperonymie /hyponymie*

La définition aristotélicienne de l'hyperonymie est la suivante : « *L'hyperonymie est une fonction qui, à partir d'un terme t, retourne un ou plusieurs termes plus généraux* ». L'hyperonymie est donc une relation hiérarchique entre des objets (des termes ou des concepts) qui est fondamentale dans le domaine de l'ingénierie des connaissances puisqu'elle constitue l'ossature de toute terminologie. (Aussenac 2020)

La relation hyperonymique a été longuement étudiée dans le domaine de la terminologie, avec une évolution dans son appellation. Nommée **relation d'inclusion** par Lyons (Lyons 1968), puis **relation générique** (generic relationship) par Sager (Sager 1990), elle établit un ordre hiérarchique entre termes d'une même classe sémantique.

Elle est souvent assimilée à la relation is-a, définie de la façon suivante : « *Si X est une classe d'objets, et X' une sous-classe de X, alors is-a(X', X) est vrai* » (Brachman 1989).

Dans ce type de relation sémantique, l'argument le plus général est appelé hyperonyme, tandis que l'argument le plus spécifique est appelé hyponyme.

Dans la base « *MAR-REL* » la liste de candidats-marqueurs relative à l'hyperonymie permet d'identifier 6 types différents : les marqueurs attributifs, appositifs, coordonnés, d'inclusion, d'exemplification, et les marqueurs avec « Utiliser ».

Dans notre corpus annoté, la relation d'hyperonymie est exprimée de façon variée, dans des structures combinant des indices lexicaux et morpho-syntaxiques.

Dans les patrons suivants, la variable ***Y*** correspond à l'hyperonyme (élément générique), la variable ***X*** correspond à l'hyponyme (élément spécifique).

- **Construction de type attributif :**

La structure prédicative attributive est la plus évidente et la plus courante pour exprimer la relation d'hyperonymie. Elle peut prendre différentes formes :

> ***Det, X ; Vbe « être » ; Det, Y***
>
> *ex : La sarcoïdose est une maladie systémique.*

> ***Det, X ; Vbe « être » ; Det, Y + caractéristiques***
>
> *ex : Le déficit en adénylosuccinate lyase est une anomalie du métabolisme des purines se manifestant précocement par un retard psychomoteur sévère, voire un syndrome autistique, le plus souvent associé à une épilepsie et un syndrome pyramidal des membres inférieurs.*

***Det, X ; Vbe « se présenter » ; Prep « comme, sous la forme de » ; Det, Y + caractéristiques***

*ex : La méningite tuberculeuse se présente sous la forme d'une méningite d'installation plutôt progressive sur quelques semaines, isolée, ou associée à des signes neurologiques focaux, avec en particulier une atteinte évocatrice de la base du crâne.*

Rq : « Sous la forme d' », locution prépositive, a été taggée avec l'attribut « Prep » pour faciliter la fluidité dans la construction des patrons.

Dans les deux exemples précédents, le definiens se compose de deux parties : un hyperonyme et des différences spécifiques, reflétant la vision classique de la définition. Ces différences (ou encore propriétés), spécifiques au definiendum, permettent de le distinguer de ses co-hyponymes.

- **Construction de type inclusif :**

***Det, X ; Vbe « constituer, représenter, consister en » ; caractéristiques {spécialisation/spécification/précision/ détermination} ; Det, Y***

*ex : La maladie d'Alzheimer constitue la plus fréquente des affections dégénératives.*

***Det, X ; Vbe « constituer, représenter, consister en » ; Det, Y ; caractéristiques {spécialisation/spécification/précision/ détermination}***

*ex : L'acro-ostéolyse longitudinale consiste en une ostéolyse de la houppe et de la diaphyse de façon concentrique, donnant un aspect en « sucre d'orge ».*

- **Construction de type appositif :**

***Det, Y (Det, X ; Ponc ; Det, X')***

*ex : Les pneumonies aiguës (pneumonie franche lobaire aiguë, pneumonie atypique)…*

***Det, Y (Det, X ; Conj ; Det, X')***

*ex : Les vascularites non nécrosantes des gros vaisseaux (artérite de Takayasu et artérite à cellules géantes de Horton)…*

#### *7.3.1.2. Méronymie/holonymie*

La relation de méronymie est la relation qui s'établit entre une partie et son tout. Cette relation, tout comme la relation hyperonymique (et la relation inverse, à savoir l'hyponymie) peut aussi être vue sous deux angles, soit du tout vers la partie (méronymie), soit de la partie vers le tout (holonymie). (Bodson 2004)

Dans les patrons suivants, la variable ***Y*** correspond « *au tout* », la variable ***X*** correspond à « *l'expression des parties* ».

- La méronymie apparaît généralement comme information principale de l'énoncé définitoire, dans un patron générique du type :

***Det, X ; Vbe « est un(e) (partie, ...) de » ; Det, Y***

*ex : Les troubles des interactions sociales sont un des symptômes de la triade autistique.*

- L'holonymie est exprimée de manière converse par les patrons suivants :

***Det, Y ; Vbe « (être) {formé/constitué/composé} de » ; Det, X***

*ex : Le spectre autistique, aussi connu sous le nom de « triade autistique », est composé de trois principaux domaines : des anomalies de la communication orale et/ou non verbale, des anomalies des interactions sociales, ainsi que des centres d'intérêt restreints.*

***Det, Y ; Vbe « (comprendre/abriter/comporter/compter/inclure/intégrer) » ; Det, X***

*ex : La triade athérogénique comporte l'association d'une élévation de l'insulinémie à jeun, d'une augmentation de l'apolipoprotéine B (apo B) et d'une diminution de la taille des particules LDL.*

Il est également possible de rencontrer un troisième patron (Cartier 2015) combinant expression hyperonymique et méronymique :

***Det, Y ; Vbe « est un(e) Hyperonyme (composé de, associant, …) » ; Det, X***

*ex : Le syndrome néphrotique est une maladie rénale qui associe une protéinurie supérieure à 3 g/24 h chez l'adulte (> 50 mg/kg/j chez l'enfant), une hypoalbuminémie inférieure à 30 g/L, et le plus souvent une hyperlipidémie avec augmentation du cholestérol.*

### 7.3.2. Relation d'équivalence, la synonymie

Cette relation sémantique est surtout présente dans les énoncés définitoires de type parenthétique.

Comme le souligne Doualan (Doualan 2013), la synonymie est communément définie comme étant une relation d'équivalence sémantique entre deux ou plusieurs unités lexicales dont la forme diffère mais qui peuvent se substituer l'un à l'autre.

*ex1 : dysplasie ostéofibreuse de Campanacci (ou adamantinome juvénile)*

*ex2 : ostéoporose (T-score <_ 2,5)*[34].

# 8. Discussion

Ce retour d'expérience expose les différentes étapes de mise en place d'une méthode d'extraction des énoncés définitoires dans les textes scientifiques.

Portant sur deux corpus de spécialités différentes, la médecine et l'astrophysique, il met en évidence la richesse des procédés concernant la pratique définitoire mais aussi les spécificités linguistiques et sémantiques propres à chaque domaine de connaissance. Il montre également l'intérêt de ces énoncés pour les spécialistes de l'IST travaillant sur la constitution et/ou l'enrichissement de terminologies. Outre les informations relatives au sens d'un terme, ils sont porteurs de relations sémantiques permettant de structurer un vocabulaire.

En ce qui concerne la phase d'extraction, les graphes élaborés dans Unitex répondent assez bien à notre objectif de repérage d'énoncés définitoires francophones de maladies. Comme nous l'avons montré, les résultats doivent toutefois faire l'objet d'une validation humaine. La réutilisabilité et l'adaptabilité des graphes sont des atouts majeurs dans ce processus de recherche d'énoncés définitoires. Ils restent évidemment perfectibles et leur amélioration repose sur l'exploration de corpus plus volumineux et diversifiés.

[34] Cette valeur de densité minérale osseuse est la définition ostéodensitométrique de l'ostéoporose.

Grâce à sa grande puissance de description lors de la construction des patrons linguistiques, Unitex apparaît donc comme un outil intéressant d'aide à l'extraction automatique d'énoncés définitoires. Il pourrait ainsi trouver sa place dans une chaine de traitement (Figure.31) appliquée sur les collections documentaires contenues dans l'archive ISTEX, afin d'enrichir les terminologies exposées sur Loterre.

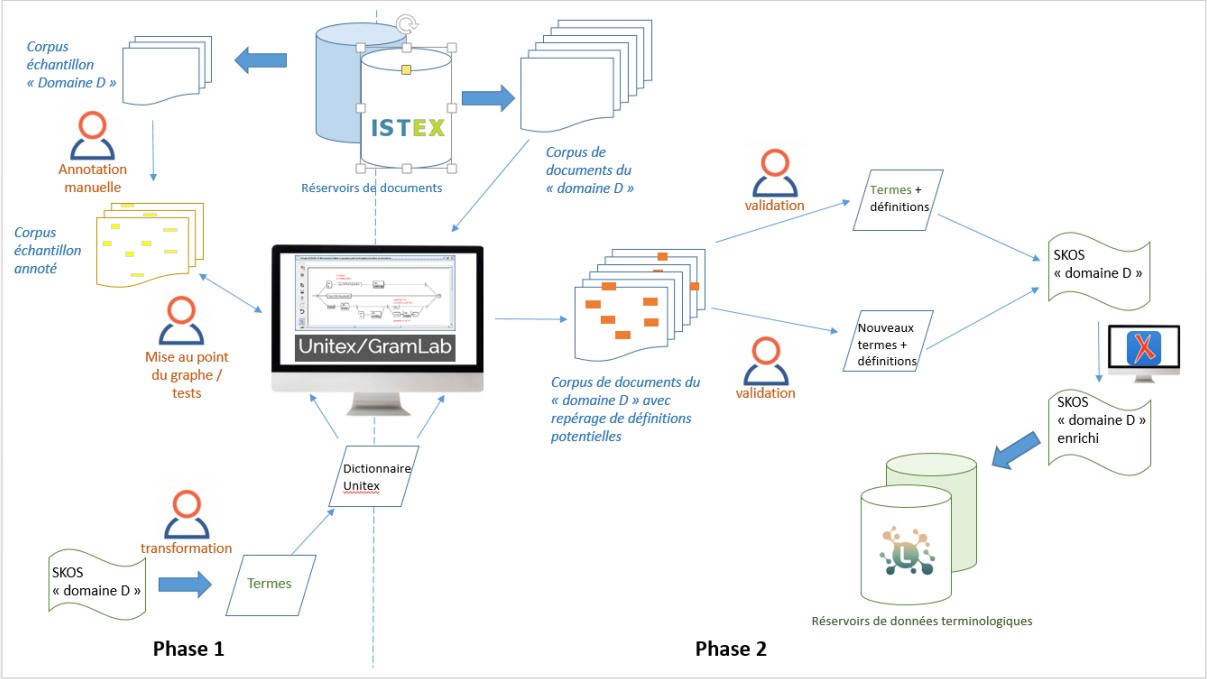



***Figure 31 : Chaine de traitement***

L'analyse des résultats de cette expérimentation renvoie au fait que le bon fonctionnement de certains graphes est fortement sous-tendu par le domaine du corpus sur lequel ils sont appliqués, ce qui vient contrecarrer notre objectif initial de généricité des graphes construits. Ce constat est particulièrement vrai pour le graphe des parenthèses qui a une portabilité quasi nulle.

Les tests dans le domaine de l'astrophysique nécessitent d'être étendus à un corpus plus volumineux afin d'évaluer plus précisément l'adaptabilité des graphes et apporter les améliorations nécessaires.

Pour un domaine donné, on voit qu'il n'est pas facile de trouver le juste niveau de contrainte des patrons lexico-syntaxiques pour parvenir à un silence informationnel acceptable.

Sur le plan méthodologique, plusieurs améliorations devront être envisagées. Il conviendra de corriger les problèmes de conversion en format texte des documents chargés, notamment l'absence d'accentuation, les défauts de segmentation du texte ou d'incohérence de certains passages. Il s'agira également d'effacer les résumés en anglais et les tableaux présents, afin de ne pas compromettre le traitement des textes avec des informations peu pertinentes.

Enfin, l'annotation manuelle a été réalisée dans chaque domaine par un seul annotateur, ce qui n'a pas permis de valider la qualité des données annotées. L'implication d'un (ou plusieurs)

autre(s) annotateur(s) permettrait d'obtenir des données de référence plus consensuelles. Le corpus pourrait alors être considéré comme fiable et représentatif de la tâche que l'on s'est fixée.

# 9. Perspectives

La prochaine étape (Figure.31) va donc consister en la mise en production des graphes dans le cadre de l'enrichissement du « *Thésaurus des pathologies humaines* ». L'extraction des définitions et de leur source a pour finalité leur injection dans le fichier SKOS (Simple Knowledge Organization System)[35] de la terminologie.

Dans certains domaines scientifiques, notamment en sciences exactes, la littérature spécialisée est exclusivement en langue anglaise. Nous profitons donc de l'expérience acquise pour poursuivre notre expérimentation dans le domaine de l'astrophysique en élaborant des graphes permettant l'extraction d'énoncés définitoires dans des corpus anglophones.

L'appropriation d'Unitex est bien sûr à poursuivre et à conforter, avec pour objectif l'acquisition d'un savoir-faire concernant l'élaboration de graphes syntaxiques nous permettant de pouvoir répondre à des besoins spécifiques.

# 10. Conclusion

Malgré les limites identifiées, cette première expérimentation sur la détection automatique d'énoncés définitoires dans le domaine biomédical a permis de mettre en place une méthodologie et des critères de recherche de ce phénomène qui pourront être enrichis en analysant d'autres corpus. La reproductibilité possible du format d'analyse apparaît prometteuse car en adaptant les règles à d'autres marqueurs, nous pourrions envisager d'analyser sur le même format d'autres phénomènes linguistiques.

Le système de traitement, basé sur l'outil Unitex a permis de détecter des énoncés à caractère définitoire et répond en ce sens à notre besoin. Les différentes structures morpho-syntaxiques rencontrées dans l'écriture des définitions, une fois modélisées par les patrons lexico-syntaxiques, peuvent être parfaitement décrites par les grammaires qui sont manipulées par Unitex sous la forme de graphes. Il reste ensuite à trouver le juste équilibre entre la syntaxe de l'écriture des définitions et leur représentation sur chacune des branches du graphe pour optimiser les paramètres statistiques que sont la précision et le rappel afin de rapprocher aux mieux le nombre de résultats pertinents attendus lors de la lecture du corpus et le nombre de résultats pertinents trouvés avec Unitex.

A notre connaissance, Unitex est le seul outil qui permette de concilier l'approche morpho-syntaxique des définitions et une technique qui imite cette approche en donnant des résultats aussi satisfaisants. C'est en ce sens un outil précieux pour la fouille de texte.

Nos résultats préliminaires sont très encourageants mais d'autres études devront être menées pour approfondir les différentes facettes de ce domaine complexe.

[35] https://fr.wikipedia.org/wiki/Simple_Knowledge_Organization_System

# Références


Auger, A. (1997). *Repérage des énoncés d'intérêt définitoire dans les bases de données textuelles.* (Doctoral dissertation, Université de Neuchâtel). [Consulté le 01 septembre 2021]. Disponible à l'adresse : https://doc.rero.ch/record/473

Aussenac-Gilles, N., Fabre, C., Ghamnia, A., Kamel, M., & Trojahn, C. (2020). *SEMPEDIA: Sémantisation à partir des documents semi-structurés-Enrichissement de DBPédia (Rapport sur les travaux de thèse d'Adel Ghamnia) Rapport de fin de contrat de la région Midi-Pyrénées Convention 620402C5266* (Doctoral dissertation, Université de Toulouse-le-Mirail). [Consulté le 07 septembre 2021]. Disponible à l'adresse : https://hal.archives-ouvertes.fr/hal-02960440/document

Bodson, C. (2005). Termes et relations sémantiques en corpus spécialisés: rapport entre patrons de relations sémantiques (PRS) et types sémantiques (TS). (Doctoral dissertation, Université de Montréal). [Consulté le 01 septembre 2021]. Disponible à l'adresse : https://papyrus.bib.umontreal.ca/xmlui/

Bore, C. (2007, February). Dénommer, désigner en classe: aspects du métalangage et interactions. In *Dénommer, désigner en classe: aspects du métalangage et interactions*.

Borillo, A. (1996). Exploration automatisée de textes de spécialité: repérage et identification de la relation lexicale d'hyperonymie. *Linx*, *34*(1), 113-124. [Consulté le 07 août 2021]. Disponible à l'adresse : https://doi.org/10.3406/linx.1996.1421

Brachman, R. J., & Schmolze, J. G. (1989). An overview of the KL-ONE knowledge representation system. *Readings in artificial intelligence and databases*, 207-230. [Consulté le 27 juillet 2021]. Disponible à l'adresse : https://doi.org/10.1016/B978-0-934613-53-8.50019-4

Brunschwig, J. (1967). Les Topiques d'Aristote et la dialectique platonicienne. La méthodologie de la définition.

Cartier, E. (2015, June). Extraction automatique de relations sémantiques dans les définitions: approche hybride, construction d'un corpus de relations sémantiques pour le français. In *Conférence annuelle Traitement Automatique des Langues Naturelles*. [Consulté le 4 août 2021]. Disponible à l'adresse : https://halshs.archives-ouvertes.fr/halshs-01412736/document

Charlet, J., Baneyx, A., Steichen, O., Alecu, I., Daniel-Le Bozec, C., Bousquet, C., & Jaulent, M. C. (2009). Utiliser et construire des ontologies en médecine. Le primat de la terminologie. *Tech. Sci. Informatiques*, *28*(2), 145-171.

Dambreville, S. C. (1995). *Rôle des organisateurs para-linguistiques dans la consultation des documents électroniques* (Doctoral dissertation, Université Stendhal-Grenoble III). [Consulté le 2 août 2021]. Disponible à l'adresse : https://tel.archives-ouvertes.fr/tel-00451634/document

Doualan, G. (2013). La synonymie, relation d'équivalence, un artefact de la pensée?. *Equivalences*, *40*(1), 15-42. [Consulté le 2 août 2021]. Disponible à l'adresse : https://doi.org/10.3406/equiv.2013.1378

Dragos, V., & Jaulent, M. C. (2010). Apprentissage de patrons lexico-syntaxiques à partir de textes. In *EGC* (pp. 615-620).

Drillon, J. (1991). *Traité de la ponctuation française* (pp. 293-325). Paris: Gallimard.

François, G. (2011). Étude comparée du fonctionnement des parenthèses et des tirets. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (9).

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Coling 1992 volume 2: The 15th international conference on computational linguistics*.

Hmida, F. (2017). *Identification et exploitation de contextes riches en connaissances pour l'aide à la traduction terminologique* (Doctoral dissertation, Université de Nantes). [Consulté le 2 août 2021]. Disponible à l'adresse : https://tel.archives-ouvertes.fr/tel-01725324/document

Husson, A. C. (2020). Activité définitoire folk et argumentation en contexte polémique. *Corela. Cognition, représentation, langage*, (HS-31).

Imprimerie nationale (France). (2002). *Lexique des règles typographiques en usage à l'Imprimerie nationale*. Imprimerie nationale.

Ji, Y. & Tutin, A. (2019). Les routines métalinguistiques dans les écrits scientifiques en français. *Studii de Lingvistica,* 9(2), 177-200.

Khayari, M., Reszetko, V., Vachez, D., Vedovotto, N., Yon, J., & Aubin, S. (2021). De TermSciences à Loterre: comment l'Inist-CNRS a rendu les terminologies ouvertes plus conformes aux principes FAIR. [Consulté le 21 septembre 2021]. Disponible à l'adresse : https://hal.archives-ouvertes.fr/hal-03176063

Labatut, F. (2018). Énoncés définitoires et subjectivité dans les débats sur l'évolution du mariage aux États-Unis. *Les Carnets du Cediscor. Publication du Centre de recherches sur la didacticité des discours ordinaires,* (14), 67-76.

Lefeuvre, L. (2017). *Analyse des marqueurs de relations conceptuelles en corpus spécialisé: recensement, évaluation et caractérisation en fonction du domaine et du genre textuel* (Doctoral dissertation, Toulouse 2)

Lefeuvre, L., & Condamines, A. (2017). MAR-REL: une base de marqueurs de relations conceptuelles pour la détection de Contextes Riches en Connaissances (MAR-REL: a conceptual relation markers database for Knowledge-Rich Contexts extraction). *In Actes des 24ème Conférence sur le Traitement Automatique des Langues Naturelles.* Volume 2-Articles courts (pp. 183-191).

Legallois, D., & Tutin, A. (2013). Présentation: Vers une extension du domaine de la phraséologie. *Langages,* (1), 3-25.

Legifrance (2021). Vocabulaire de l'informatique (liste de termes, expressions et définitions adoptés). [Consulté le 2 août 2021]. Disponible à l'adresse : https://www.legifrance.gouv.fr/

Lyons, J. (1968). Introduction to theoretical linguistics (Vol. 510). *Cambridge university press.*

Lyons, J. (1980). Josette Rey-Debove, Le métalangage: Étude linguistique du discours sur le langage. (Collection L'ordre des mots.) Paris: Le Robert, 1978. Pp. 318. *Journal of Linguistics*, *16*(2), 292-300.

Malaisé, V., Zweigenbaum, P., & Bachimont, B. (2004, April). Repérage et exploitation d'énoncés définitoires en corpus pour l'aide à la construction d'ontologie. *In Actes de la 11ème conférence sur le Traitement Automatique des Langues Naturelles.* Articles longs (pp. 149-158).

Meyer, I. (2001). Extracting knowledge-rich contexts for terminography. *Recent advances in computational terminology,* 2, 279.

Murat, M., & Cartier-Bresson, B. (1987). C'est-à-dire ou la reprise interprétative. *Langue française,* (73), 5-15.

Paumier, S. (2011). Unitex-manuel d'utilisation.

Pearson, J. (1998). Comment accéder aux éléments définitoires dans les textes spécialisés ? *In Terminologies nouvelles* (19): 21-28.

Péry-Woodley, M.P. (1990). Textual clues for user modeling in an ITS. Thesis for the degree of Master in Cognitive Science, University of Manchester.

Prince, V. (1994). Indices linguistiques pour la construction d'un modele automatique d'analyse et de production des explications. *Actes de l'atelier de recherche GENE du PRC IA*, 141-153.

Rebeyrolle, J., & Tanguy, L. (2000). Repérage automatique de structures linguistiques en corpus: le cas des énoncés définitoires. *Cahiers de grammaire,* 25, 153-174.

Rey-Debove, J. (1978). Le métalangage. Etude linguistique du discours sur le langage, Paris, Le Robert.

Sager, J. C. (1990). Practical course in terminology processing. *John Benjamins Publishing.*

Seppälä, S. (2010, July). Automatiser la rédaction de définitions terminographiques: questions et traitements. In *Actes de la 17e conférence sur le Traitement Automatique des Langues Naturelles. REncontres jeunes Chercheurs en Informatique pour le Traitement Automatique des Langues* (pp. 93-102).

Seppälä, S. (2012). Contraintes sur la sélection des informations dans les définitions terminographiques: vers des modèles relationnels génériques pertinents. (Doctoral dissertation, University of Geneva). [Consulté le 21 août 2021]. Disponible à l'adresse : https://archive-ouverte.unige.ch/unige:21874

Tolone, E. (2006). Extraction d'entitées nommées par les graphes d'Unitex. [Consulté le 21 août 2021]. Disponible à l'adresse : http://www-igm.univ-mlv.fr/~tolone/publi/Rapport_Stage_2006_Tolone.pdf

Tutin, A., & Kraif, O. (2016). Routines sémantico-rhétoriques dans l'écrit scientifique de sciences humaines: l'apport des arbres lexico-syntaxiques récurrents. *Lidil. Revue de linguistique et de didactique des langues,* (53), 119-141.

Vachez, D. (2021). Etude comparative de thésaurus en Sciences de l'Environnement-Bonnes pratiques de conception et FAIRisation de thésaurus. [Consulté le 30 septembre 2021]. Disponible à l'adresse : https://hal.archives-ouvertes.fr/hal-03264803/document

Van Rijsbergen, C. J. (1979). *Information Retrieval*, 2nd edition. London, Butterworths

Vedenina, L. G., Védénina, L. G., & Védénina, Ludmilla. G. (1989). *Pertinence linguistique de la présentation typographique* (Vol. 21). Peeters Publishers.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, *3*(1), 1-9.

Yan, R., & Hatier, S. (2016). L'extraction et la modelisation de patrons lexico-syntaxiques pour leur enseignement en FLE: un exemple a partir du verbe montrer. *Linguistik online,* (78).