# Tech-mining on Chinese Patents: Syntax and Terminology

LI Yixuan

## Résumé en français

### Fouille technologique dans les brevets chinois : syntaxe et terminologie

Cette thèse décrit un projet de recherche avancé qui cible l'automatisation de la création de variantes lexicales pour des termes techniques au sein des demandes de brevet chinoises. Il se distingue par deux avancées notables : premièrement, l'élaboration d'un outil d'analyse de dépendance au niveau des caractères, spécialement formé pour les textes de brevets chinois, facilitant l'analyse de la structure interne des termes techniques et abordant ainsi la problématique historique de segmentation en chinois. Deuxièmement, la construction d'une taxonomie technique s'appuyant sur les intitulés de la Classification internationale des brevets (IPC), qui permet d'offrir des substituts efficaces d’hyperonymes et d'hyponymes pour générer des variantes de textes de brevets.

## Introduction

L'introduction de la thèse met en exergue l'importance cruciale des brevets en tant que ressources principales d'informations techniques, englobant une grande partie des connaissances scientifiques et technologiques mondiales. Elle souligne également les défis liés à l'accès et à l'analyse de ces données, en raison de leur volume conséquent et de la complexité des documents. Le projet vise spécifiquement à améliorer l'accès et l'analyse automatique des textes de brevets chinois.

La "Tech-mining" ou la fouille technologique, est définie comme l'application de techniques avancées de traitement de texte aux ressources d'information scientifique et technologique. Elle est utilisée pour identifier des technologies émergentes et soutenir la prise de décision stratégique.

Le projet se focalise sur l'analyse technologique des brevets chinois, en réponse à la croissance significative des demandes de brevets dans ce pays. L'étude a pour but de traiter automatiquement les textes de brevets chinois, un domaine encore peu exploité malgré un intérêt croissant pour la traduction automatique et l'analyse textuelle.

L'objectif de la thèse est d'examiner en profondeur les caractéristiques linguistiques, particulièrement morphologiques et syntaxiques, des termes techniques chinois, et d'explorer les méthodes de reconnaissance des termes et de variation lexicale adaptées aux besoins spécifiques de la rédaction des revendications de brevets.

La structure de la thèse est organisée en cinq chapitres principaux (Figure 1), allant de l'introduction et la collecte de données, à l'analyse syntaxique détaillée des revendications de brevets chinois au niveau des caractères, en passant par la reconnaissance des termes de brevets chinois basée sur l'analyse syntaxique de dépendance, pour finir par l'élaboration et l'évaluation d'une taxonomie technique orientée vers la variation lexicale.

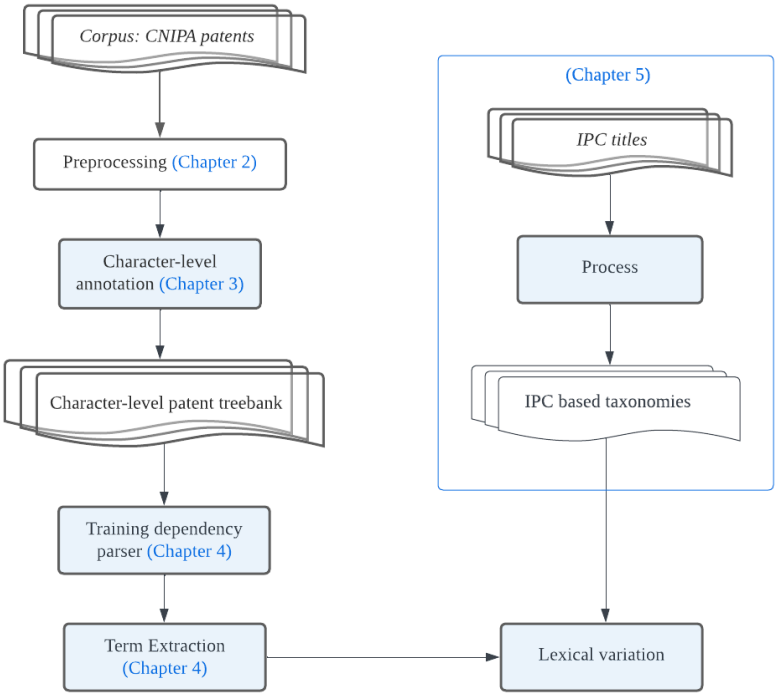


**Figure 1 - La Structure de la Thèse**

# Chapitre 1 - Le contexte linguistique et technique

Le premier chapitre de cette thèse dévoile les subtilités de la structure des brevets, l'importance de l'hyperonymie dans leur rédaction, ainsi que les divers systèmes de classification des brevets. Il établit un cadre linguistique et technique essentiel pour la compréhension des brevets.

La section 1.1 explore non seulement le brevet en tant que document complexe, mais également son rôle crucial dans la protection de la propriété intellectuelle. Il examine en outre les recherches antérieures en traitement automatique du langage et en exploration technologique, spécifiquement dans le contexte des brevets.

Dans la section 1.2, on se concentre sur l'analyse syntaxique et terminologique des brevets en langue chinoise. Cette analyse se divise en deux volets principaux : l'étude de la morphologie et de la syntaxe du chinois moderne, et les études terminologiques.

Dans la première partie de la section 1.2, on aborde les fondements de la formation des mots en chinois. Il décrit diverses théories sur la manière dont les mots sont formés, un aspect crucial pour comprendre la structuration des termes dans les brevets. on examine également les structures phrastiques typiques du chinois moderne, en mettant l'accent sur cinq structures de base essentielles pour l'analyse syntaxique des textes de brevets.

Les cinq catégiries principales de relations syntaxiques et de relation interne des mots dans la langue chinoise sont décrites comme suit :

- Attributive (偏正式, piān zhèng shì) : Cette catégorie couvre les relations où un élément (généralement un adjectif ou un nom) modifie ou qualifie un autre nom. Elle est essentielle pour comprendre comment les caractéristiques ou attributs sont exprimés dans les phrases chinoises.
- Coordinative (联合式, lián hé shì) : Cette relation implique des éléments qui sont coordonnés ensemble, souvent à l'aide de conjonctions. Elle est utilisée pour relier des mots, des phrases ou des clauses de même importance dans le discours.
- Predicate-Object (述宾式, shù bīn shì) : Cette catégorie est cruciale pour identifier les relations entre un prédicat, généralement un verbe, et son objet. Elle aide à déterminer comment l'action d'un verbe est dirigée vers un objet dans une phrase.

- Predicate-Complement (述补式, shù bǔ shì) : Dans cette catégorie, un complément suit le prédicat pour fournir des informations supplémentaires à propos de l'état ou de l'action décrite par le prédicat. Ces compléments peuvent souvent indiquer le résultat, la direction ou le degré de l'action.
- Subject-Predicate (主谓式, zhǔ wèi shì) : Cette relation est la plus fondamentale dans les phrases chinoises, établissant une connexion entre le sujet et le prédicat. Elle détermine comment le sujet d'une phrase est lié à l'action ou à l'état exprimé par le prédicat.

Bien que la terminologie puisse varier, ces cinq catégories sont généralement acceptées et reconnues pour leur importance dans l'analyse grammaticale et syntaxique du chinois.

Un point important de cette section est le parallélisme observé entre la structure des mots composés et la structure syntaxique en chinois. Cette relation influence significativement l'analyse syntaxique au niveau des caractères, un domaine crucial dans le traitement automatique de la langue chinoise.

La section 1.2.1 soulève à la fin également les défis associés à la segmentation des mots en chinois, une tâche complexe en raison de l'absence de délimiteurs clairs entre les mots. Cette caractéristique unique de la langue chinoise rend l'analyse syntaxique de dépendance et l'utilisation des treebanks chinois particulièrement pertinentes.

La deuxième partie de la section 1.2 se concentre sur les études terminologiques en langue chinoise. On examine l'origine et le développement de la terminologie en tant que discipline indépendante, en soulignant son importance croissante dans le contexte des brevets. Il trace également une brève histoire de la terminologie en Chine, démontrant comment la recherche dans ce domaine a évolué au fil du temps.

Un aspect essentiel de cette section est la discussion sur les ressources disponibles pour les termes techniques chinois. On aborde l'existence de thésaurus, de bases de données terminologiques et de graphes de connaissances, tous des outils précieux pour les spécialistes des brevets et les linguistes travaillant avec des textes de brevets chinois.

La section 1.3 de la thèse est consacrée aux définitions des notions clés relatives aux brevets, notamment la distinction entre "terme" et "terme technique". On clarifie les concepts de "unithood" et "termhood" dans le cadre de l'extraction automatique de termes (ATE),

mettant l'accent sur l'identification des termes pertinents dans les textes de brevets grâce à l'indexation des composants.

Au cœur de la section 1.3, on réévalue le concept de "terme" pour éviter toute confusion avec le mot "mot" en chinois, qui peut être interprété de différentes manières, notamment en référence aux caractères individuels et aux phrases. Un "terme", selon on, est une désignation verbale d'un concept général pertinent dans un domaine spécifique. Cette perspective s'aligne sur les vues de théoriciens comme Wüster, qui considère les termes comme des étiquettes pour les concepts, et d'autres tels que Sager, Kageura, Umino, Cabré, et l'ISO, qui mettent en évidence leur rôle dans la communication spécialisée.

La section aborde également deux aspects essentiels des termes techniques : l'unité (unithood) et la spécificité terminologique (termhood). L'unité fait référence à la cohésion des combinaisons de mots, tandis que la spécificité terminologique indique le degré d'association d'une unité linguistique avec des concepts propres à un domaine spécifique. Dans le cadre des brevets, un "terme" est défini comme une unité lexicale substituable identifiée dans les corpus de brevets, allant des plus petites unités syntaxiques aux expressions multi-mots.

Une distinction claire est faite entre les termes spécifiques aux brevets et les termes techniques généraux. On examine des listes d'expressions souhaitables et indésirables dans les revendications de brevets, où les termes souhaitables sont directement liés à des techniques spécifiques et des composants de système d'un brevet donné. En revanche, d'autres expressions telles que "caractéristique" et "revendication" sont fréquemment rencontrées dans de nombreux brevets et ne sont donc pas spécifiques.

La section 1.3 met également en lumière l'importance des numéros de référence dans les descriptions et revendications de brevets. Ces numéros, qui se réfèrent aux symboles des composants dans les dessins des brevets, jouent un rôle crucial dans l'identification des termes pertinents pour l'extraction et la substitution dans les chapitres 4 et 5.

Enfin, la section 1.4 présente l'hypothèse, les difficultés et la méthodologie de sa recherche sur les termes techniques chinois utilisés dans les revendications de brevets. L'objectif est d'analyser en profondeur la structure interne de ces termes via l'annotation de syntaxe de dépendance et d'explorer la variation lexicale pour répondre aux exigences spécifiques de la portée des revendications. La recherche est divisée en deux parties principales : une analyse syntaxique et la reconnaissance des termes basée sur un analyseur au niveau des

caractères, suivie de la construction de taxonomies techniques pour faciliter les substitutions lexicales.

Les défis rencontrés comprennent le manque de ressources chinoises, comme les treebanks, les modèles, les wordnets et les bases de données terminologiques, et l'absence de lignes directrices cohérentes pour les annotations UD (Universal Dependencies) ou SUD (Surface-Syntactic Universal Dependencies) spécifiques au chinois. Pour surmonter ces obstacles, les auteurs développent leurs propres méthodes d'évaluation adaptées à la nature unique de leurs objectifs de recherche.

Les questions clés abordées dans la thèse sont :

- Peut-on décrire de manière cohérente la structure interne des termes techniques chinois en utilisant des liens syntaxiques standards ?
- Les analyseurs basés sur les caractères chinois donnent-ils de meilleurs résultats que les approches basées sur les mots ?
- La syntaxe aide-t-elle à reconnaître la terminologie, en particulier pour les nouveaux termes ?
- Comment évaluer une taxonomie ?

L'objectif global est de fournir des analyses complètes sur la composition et l'identification des termes techniques chinois, d'explorer le rôle de la syntaxe dans la reconnaissance terminologique et de réaliser des évaluations et comparaisons relatives aux taxonomies, y compris leurs distances et autres relations sémantiques.

## Chapitre 2 - Collection et Prétraitement des Textes de Brevets

Dans le domaine complexe et multidimensionnel de la propriété intellectuelle, le chapitre 2 de la thèse dédiée à l'analyse des brevets chinois joue un rôle crucial. Ce chapitre se penche de manière exhaustive sur la collecte et le prétraitement des textes de brevets, explorant les subtilités et les défis inhérents à ce processus.

La collecte des données de demandes de brevets chinois est le point de départ dans la section 2.1. Ces données, obtenues de la CNIPA (Administration Nationale de la Propriété Intellectuelle de Chine) et couvrant la période de novembre 2017 à mars 2021, forment un

corpus imposant de plus de 300 gigaoctets. Les fichiers XML, qui constituent la principale source de ces données, sont organisés annuellement et offrent un aperçu détaillé des demandes de brevets.

La section 2.2 explore également la structure typique d'une demande de brevet. Chaque dossier est constitué d'un fichier XML principal et de plusieurs fichiers JPG illustrant des aspects visuels tels que des formules et des schémas. Ces brevets sont soigneusement segmentés en sections distinctes, allant des données bibliographiques aux revendications, ce qui facilite leur analyse et traitement ultérieurs.

Et dans la section 2.3, un autre aspect crucial est la transformation des fichiers XML en textes bruts. Ce processus, réalisé via Python, permet d'extraire des informations essentielles telles que le titre, la date, la classe IPC, ainsi que le résumé, la description et les revendications de chaque brevet. En outre, un nettoyage et une normalisation approfondis sont effectués pour améliorer la qualité des textes transformés.

La classification des brevets selon la Classification Internationale des Brevets (IPC) est une étape importante pour une analyse systématique par secteur technologique dans la section 2.4. Cette classification couvre une gamme variée de domaines, allant de la médecine à l'électricité, offrant ainsi une perspective précieuse sur les tendances technologiques actuelles.

Enfin, la section 2.5 met en lumière les particularités linguistiques et stylistiques des brevets chinois. Les directives de rédaction imposent une structure complexe, notamment dans les phrases de revendication, qui se caractérisent par leur longueur, leur technicité et leur complexité grammaticale. Ces caractéristiques posent des défis uniques pour le traitement automatique du langage, nécessitant une analyse minutieuse des relations hiérarchiques et conceptuelles pour une interprétation correcte des revendications.

En conclusion, ce chapitre offre une base solide pour comprendre la complexité liée à la collecte et au prétraitement des textes de brevets chinois. Il aborde chaque aspect du processus de manière exhaustive, mettant en lumière les défis et les subtilités de l'analyse des données de brevets, un préalable indispensable pour toute recherche avancée dans le domaine de la propriété intellectuelle et de l'analyse des brevets.

# Chapitre 3 - Analyse Syntaxique au Niveau des Caractères des Revendications de Brevets Chinois

Dans la section 3.1, une approche novatrice pour annoter la banque d'arbres chinoise est proposée. Cette section détaille un schéma d'annotation pour les relations internes des termes et les parties du discours, adapté au cadre des Dépendances Syntaxiques Universelles de Surface (SUD). Des tests distributionnels et sémantiques sont établis pour déterminer les parties du discours des caractères et leurs relations inter-caractères.

Le processus d'annotation se divise en deux parties : l'annotation UPOS pour les caractères uniques et l'ExtPos pour les unités multi-caractères liées par une relation "@m". L'importance est accordée à la position distributionnelle des mots plutôt qu'à leur sens sémantique, un défi notable dans le contexte linguistique chinois où les mots peuvent avoir plusieurs étiquettes de partie du discours.

Des critères spécifiques pour chaque catégorie lexicale principale (noms, verbes, adjectifs et adverbes) sont détaillés, soulignant les caractéristiques uniques de la langue chinoise. Par exemple, les noms chinois ne prennent pas les marqueurs d'aspect, tandis que les verbes peuvent être suivis de marqueurs dynamiques.

Des tests pour les relations inter-caractères sont également présentés, établissant un lien entre les structures de mots et les structures de phrases dans les treebanks de dépendance. Chaque classe est définie avec des exemples tirés du corpus de revendications de brevets, en tenant compte de facteurs tels que la limite entre les relations au niveau des caractères et les relations syntaxiques conventionnelles, la structure interne du terme et les classes lexicales des termes entiers et de leurs composants.

Le schéma d'annotation englobe toutes les 17 étiquettes UPOS issues du cadre UD. Cependant, le style d'écriture spécifique des textes de brevets signifie que certaines étiquettes, comme INTJ, n'apparaissent pas dans la banque d'arbres finale.

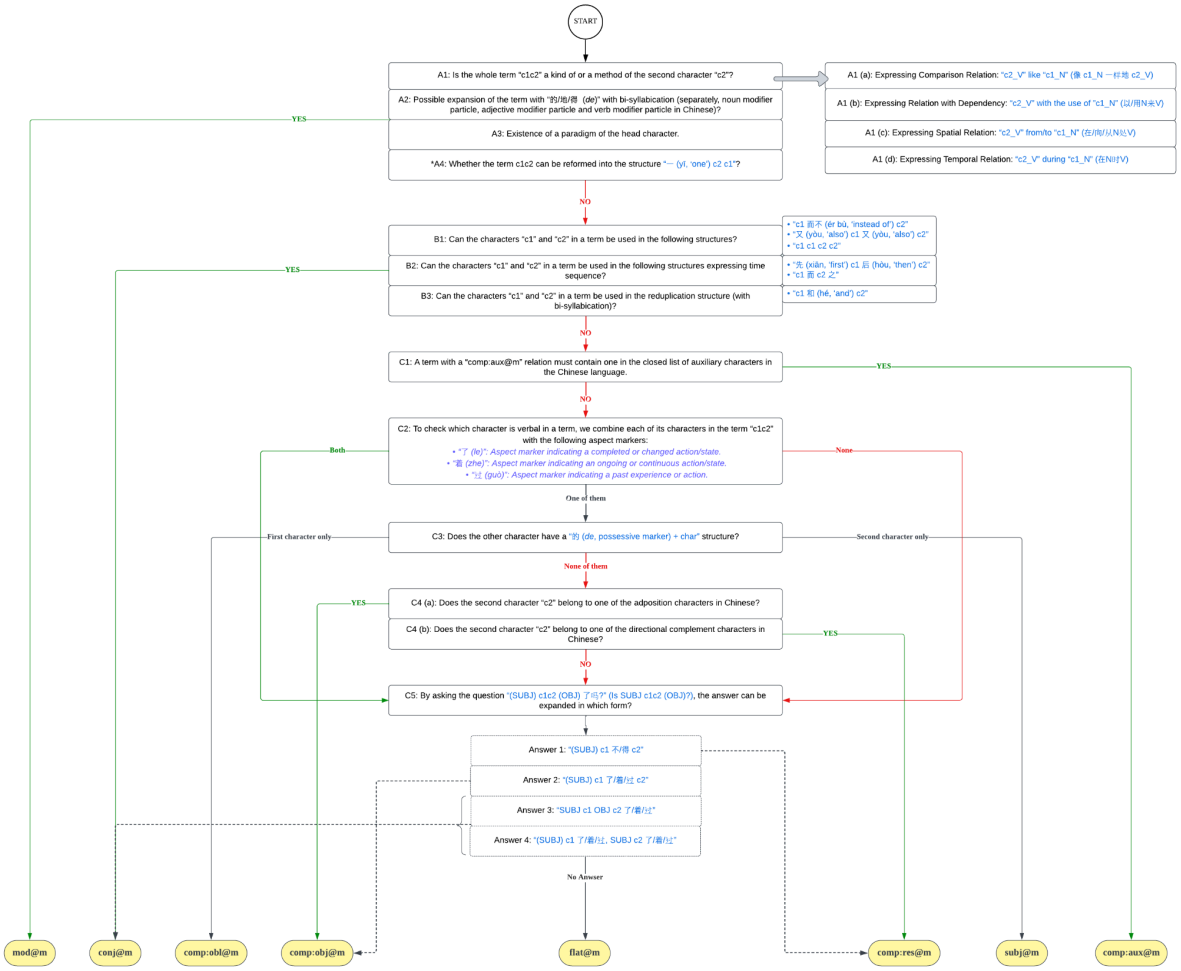


**Figure 2 - Arbre de décision finale**

La section aborde également la complexité et les défis de l'évolution de l'utilisation des caractères en chinois moderne, soulignant les difficultés de catégorisation et d'annotation basées uniquement sur des tests distributionnels.

Un arbre de décision complet (Fugure 2) est présenté pour guider l'annotation cohérente des termes internes, basé sur une analyse approfondie des propriétés syntaxiques et sémantiques associées à ces caractères. Ce processus implique l'utilisation d'un système automatique de balisage des parties du discours, soumis à une correction manuelle en utilisant les lignes directrices de balisage des parties du discours pour la Penn Chinese Treebank et le Dictionnaire Xin Hua pour l'annotation au niveau des caractères.

Enfin, cette section propose une hiérarchisation des étiquettes de relations morphologiques et syntaxiques pour les banques d'arbres au niveau des caractères, basée sur le schéma

SUD. Trois niveaux de granularité sont établis en fonction de la transparence de leur structure interne (Table 1) : relations syntaxiques régulières, relations morphologiques computationnelles (toutes les étiquettes avec "@m" sauf "flat@m") et relations syntaxiques entre caractères chinois qui n'existent pas en interne ou ne sont pas analogues aux structures syntaxiques (marquées avec "flat@m").

➔ **Niveau 1: relations synatxiques**

| **Etiquettes** | **Relations Synatxiques** |
|---|---|
| appos | Un modificateur appositionnel d'un nom est un nom qui suit immédiatement le premier nom et qui sert à définir, modifier, nommer ou décrire ce nom. |
| conj | Une conjonction est une relation entre deux éléments connectés par une conjonction de coordination, telle que "et", "ou", etc. |
| comp | La relation comp est utilisée pour les arguments des verbes, noms, adjectifs, adverbes, auxiliaires, adpositions et conjonctions. |
| comp:obj | La relation comp:obj est utilisée pour les compléments d'objet direct, y compris les compléments directs d'une adposition ou d'une conjonction subordonnée. |
| comp:obl | La relation comp:obl est utilisée pour les arguments obliques des verbes, adjectifs, adverbes, noms ou pronoms, quelle que soit leur forme. |
| comp:cleft | La relation comp:cleft est utilisée dans les phrases clivées pour la dépendance du chef de la phrase au chef de la proposition complétive. |
| comp:pred | La relation comp:pred est utilisée pour les arguments prédicatifs des verbes. |
| comp:aux | La relation comp:aux est utilisée pour l'argument des auxiliaires et correspond à la relation aux telle que définie par UD (Universal Dependencies). |
| comp:svc | La relation comp:svc est utilisée pour la construction de verbes sériels. |
| comp:dir | La relation comp:dir est utilisée pour les arguments directionnels des verbes. |
| comp:res | La relation comp:res est utilisée pour les arguments résultatifs des verbes. |

| | |
|---|---|
| subj | La relation subj est utilisée pour tous les sujets, quelle que soit leur forme (nominale ou verbale). Cette relation englobe à la fois les relations nsubj et csubj telles que définies par UD, comme le montrent les exemples suivants. |
| mod | La relation mod est utilisée pour les modificateurs de verbes, noms, adjectifs, adverbes, auxiliaires, adpositions et conjonctions. |
| flat | La relation flat est utilisée pour les expressions non chinoises, telles que « DIN ISO 4590-86 » « Bloking/Reacting Buffer ». |
| parataxis | La relation parataxis (du grec signifiant « placer côte à côte ») est une relation entre un mot (souvent le prédicat principal d'une phrase) et d'autres éléments, tels qu'une parenthèse sententielle ou une clause après un « : » ou un « ; », placés côte à côte sans aucune coordination, subordination ou relation d'argumentation explicite avec le mot tête. |
| punct | La relation punct est utilisée pour tout signe de ponctuation dans une clause, si la ponctuation est conservée dans les dépendances typées. |

➔ **Niveau 2: relations morphologiques**

| **Etiquettes** | **Relations Morphologiques** |
|---|---|
| conj@m | Type coordonnatif (联合型, lián hé xíng; 并列式, bìng liè shì) : Composé de deux morphèmes ayant des significations similaires, apparentées ou opposées. |
| comp:obj@m<br><br>comp:aux@m<br><br>comp@m | Type verbe-objet (动宾型, dòng bīn xíng; 支配式, zhī pèi shì) : Le premier morphème représente une action ou un comportement, tandis que le second morphème représente l'entité ou l'objet associé à cette action ou à ce comportement.<br>Type auxiliaire-verbe (助宾型, zhù bīn xíng) : Le premier morphème est un auxiliaire et le second est un verbe.<br>Type prédicat-complément (补充型, bǔ chōng xíng) : Le morphème suivant fournit des informations supplémentaires au morphème précédent. |

| | |
|---|---|
| *comp:obl@m<br><br>*comp:res@m | Un sous-type du prédicat-complément : Le morphème suivant est une préposition.<br><br>Un sous-type du prédicat-complément : Le morphème suivant est le complément résultatif du morphème précédent. |
| subj@m | Type sujet-prédicat (主谓型, zhǔ wèi xíng) : Le morphème suivant prédique la chose ou l'objet décrit par le morphème précédent. |
| mod@m | Type attributif (偏正型, piān zhèng xíng) : Le morphème précédent modifie de manière restrictive le morphème suivant. |

➔ **Niveau 3: flat@m**

| | |
|---|---|
| flat@m | Lorsque la relation interne du terme est ambiguë et ne peut être classée dans aucun des types précédemment mentionnés. Cela inclut des exemples comme les termes translittérés, les onomatopées, les termes redoublés et les termes fonctionnels qui ont été lexicalisés depuis longtemps sans aucune structure syntaxique discernable.<br><br>Ce type se présente comme la seule catégorie qui ne peut pas être divisée et se qualifie sans équivoque de « mots » dans cette étude. |

**Table 1 - Hiérarchie des Etiquettes**

La section 3.2 présente la construction et le traitement préliminaire automatique de la Banque d'Arbres de Revendications de Brevets Chinois au niveau des caractères. La banque a été établie en sélectionnant aléatoirement des phrases de revendications de brevets entre novembre 2017 et septembre 2018. Les 200 premières phrases provenaient de la section G - Physique de la Classification Internationale des Brevets. Ces phrases ont été segmentées en unités syntaxiques courtes. La longueur moyenne des revendications de brevets après segmentation était de 42,54 caractères par phrase.

Une annotation manuelle méticuleuse a été réalisée sur le premier ensemble de 100 phrases, tandis qu'un second ensemble de 100 phrases a subi une pré-annotation automatique suivie d'une correction manuelle. Ce processus a inclus la segmentation des mots, l'étiquetage des parties du discours (POS) et l'analyse de la dépendance, réalisée à

l'aide des pipelines de traitement du langage tels que SpaCy, Stanza et Trankit. Les résultats de segmentation des mots sont indiqués par le label "@m" dans les étiquettes de relation.

La section 3.3 aborde l'annotation manuelle et la correction des treebanks pré-annotées, en analysant les structures syntaxiques fréquentes dans les revendications de brevets chinois et en présentant l'annotation de structures fréquentes spécifiques à ces revendications. Une attention particulière est accordée aux cas problématiques rencontrés lors de l'annotation, notamment les termes de contenu et de fonction ayant une structure interne incertaine, ainsi que les expressions ayant des limites de mots floues et des structures syntaxiques ambiguës. Des exemples spécifiques de ces structures sont présentés pour illustrer les choix pratiques effectués lors de l'annotation.

Enfin, la section 3.4 présente la première banque d'arbres au niveau des caractères sur les revendications de brevets chinois, comprenant 200 phrases et 8 175 tokens. Cette banque d'arbres est disponible sur GitHub dans le cadre du projet SUD. Le chapitre conclut avec une évaluation de l'applicabilité des lignes directrices d'annotation, en présentant un score d'inter-annotateurs et en décrivant la conversion de la banque d'arbres en format UD conventionnel, en combinant chaque relation "@m" et en utilisant l'ExtPos comme UPOS pour les termes combinés.

## Chapitre 4 - Segmentation conjointe et analyse syntaxique des revendications de brevets chinois

Section 4.1 est sur l’analyse syntaxique automatique en chinois. L'analyse syntaxique et la segmentation des mots ont évolué depuis mes premières tentatives d'analyse des brevets chinois en 2018. Les méthodes traditionnelles de segmentation des mots et d'analyse syntaxique s'étaient révélées insuffisantes pour les textes de revendications de brevets. Depuis 2003, diverses normes de segmentation des mots chinois ont émergé. Meng et al. (2019) ont démontré que les modèles basés sur les caractères surpassent ceux basés sur les mots dans quatre tâches de référence en TAL. Nous avons donc réexaminé le concept de « mots », en formant le parseur de dépendance au niveau des caractères, permettant de dériver automatiquement des « mots » via l'étiquette “@m”. Nous avons développé un segmenteur-parseur conjoint basé sur nos arbres syntaxiques annotés au niveau des caractères.

L'analyse syntaxique, cruciale en linguistique et en TAL, vise à découvrir la structure grammaticale des phrases. Deux approches principales sont utilisées : les méthodes basées sur les règles et les méthodes statistiques. Le grammaire contextuelle libre (CFG) et la grammaire de dépendance sont deux cadres grammaticaux prédominants. Les modèles de parseurs de dépendance se divisent en catégories basées sur les graphes et les transitions. Les scores de parsing pour le chinois sont généralement inférieurs à ceux d'autres langues, ce qui pourrait s'expliquer par la segmentation préalable des mots et les modèles basés sur les caractères.

Dans la section 4.2 sur la segmentation des mots par analyse syntaxique avec “@m”, nous avons affiné un parseur de dépendance basé sur Bert sur notre corpus de revendications de brevets. En combinant toutes les relations “@m”, le parseur de dépendance sert également de segmenteur de mots. Les résultats d'évaluation montrent une performance supérieure en reconnaissance des tags de relation par rapport aux têtes. Les scores pour la segmentation des mots et le marquage des parties du discours atteignent presque 90%. Cependant, ces scores sont inférieurs aux résultats conventionnels, probablement en raison de la petite taille de l'ensemble de données d'entraînement.

Dans la section 4.2.1, on entraine le parseur basé sur Bert. Le parseur que nous avons utilisé remplace les premières couches Bi-LSTM par un modèle BERT. Nous avons comparé les résultats obtenus avec et sans modèle pré-entraîné. Bien que les résultats soient similaires, la principale différence réside dans la vitesse d'apprentissage.

La section 4.2.2 de la thèse, dédiée à l'analyse des erreurs dans le traitement automatique des textes de brevets chinois, se concentrant sur la segmentation des mots et l'analyse syntaxique.

Nous avons examiné les erreurs liées à la segmentation des mots et à l'analyse syntaxique. Les résultats indiquent qu'il n'y a pas de corrélation significative entre la longueur des phrases et la qualité de la segmentation des mots. La corrélation la plus forte est entre l'exactitude de l'analyse syntaxique et la qualité de la segmentation des mots. Nous avons également comparé notre segmenteur basé sur la dépendance avec des outils standards comme Jieba et NLPIR/ICTCLAS, observant des erreurs systématiques dans leurs segmentations. Jieba est connu pour son efficacité et son exactitude, utilisant une structure de dictionnaire de préfixes et un modèle de Markov caché pour traiter les mots inconnus. ICTCLAS, quant à lui, utilise un système d'analyse lexicale chinoise basé sur un modèle de HMM multicouche, avec une précision de segmentation impressionnante.

L'étude identifie trois types d'erreurs systématiques dans la segmentation de Jieba, liées aux numéros de série, aux déterminants et aux termes polysyllabiques. ICTCLAS présente également des différences dans la segmentation des termes polysyllabiques, mais pas pour les numéros de série et les déterminants.

L'étude évalue l'impact de la distance de dépendance sur la performance de l'analyseur syntaxique. Bien qu'il existe une corrélation entre la qualité de la segmentation des mots et la précision de l'analyse syntaxique, la longueur des phrases et la qualité du marquage des parties du discours ont un impact limité. La distance moyenne de dépendance (MDD) dans les revendications de brevet est comparée à celle d'autres corpus.

Finalement, nous avons analysé la précision du parsing pour chaque type de relation avec un matrix de confusion.

Notre étude souligne l'importance d'une segmentation précise des mots pour obtenir une analyse syntaxique de haute qualité.

Dans la section 4.3, on parle de la reconnaissance des termes de brevets chinois basée sur l'analyse syntaxique de dépendance

Section 4.3.1 est la présentation de la reconnaissance des termes. Dans cette section, nous explorons l'utilisation du parseur de dépendance pour reconnaître des termes dans des brevets chinois. Nous nous concentrons sur l'extraction de termes techniques à partir de revendications de brevets en utilisant l'analyse syntaxique de dépendance, plutôt que l'extraction automatique de termes (ATE) basée sur les parties du discours (POS). Nous comparons également nos résultats à d'autres techniques existantes, comme l'extraction de motifs POS.

Les systèmes ATE utilisent généralement une procédure en deux étapes : extraction d'une liste de candidats termes et détermination des termes corrects via des techniques supervisées ou non supervisées. La recherche en extraction de termes se concentre principalement sur l'amélioration des algorithmes de classement des candidats termes, avec peu d'attention accordée à la sélection des candidats termes. Yu et al. (2019) classent les méthodes de sélection des candidats termes en trois catégories : filtrage n-gramme, groupement de phrases nominales et correspondance de motifs de tags POS.

Dan la section 4.3.2, on fait une reconnaissance des termes basée sur les informations de relation de dépendance. Yu et al. (2019) ont introduit l'application de l'analyse syntaxique de dépendance à la reconnaissance des termes. Cette approche découvre des relations sémantiques de modification entre les mots dans les phrases en examinant les relations de dépendance. Nous utilisons notre parseur de dépendance au niveau des caractères pour extraire des termes des revendications de brevets, en regroupant toutes les relations internes marquées avec "@m" et en extrayant les tokens liés par la relation "mod", indiquant une relation de composé attributif.

Nous avons effectué des expériences séparées sur le corpus annoté manuellement et le corpus analysé automatiquement, en utilisant une liste de mots vides. Nous avons extrait 328 termes candidats du corpus annoté manuellement et 523 termes candidats du corpus analysé automatiquement, avec 220 termes candidats communs entre les deux listes.

Dans la section suivante de 4.3.3, on compare avec le système conventionnel basé sur POS. Nous évaluerons plus complètement les résultats d'extraction en utilisant la métrique C-value, en la comparant avec un reconnaîsseur de termes conventionnel basé sur le marquage POS. Nos résultats initiaux indiquent une précision de 16,30 % et un rappel de 63,64 % avant l'application de toute méthode de classement.

Nous cherchons à faciliter la variation lexicale pour l'invention augmentée, plutôt que d'acquérir uniquement un lexique technique adapté à un domaine spécifique. En définissant nos propres critères pour extraire des termes des revendications de brevets, nous évaluons la performance du reconnaîsseur de termes basé sur la dépendance en déterminant combien de termes indexés il identifie avec succès.

## Chapitre 5 - La variation lexicale dynamique orientée avec la construction d'une taxonomie de brevets

Le chapitre 5 offre une exploration approfondie de la création et de l'application d'une taxonomie technique pour les brevets. Cette taxonomie, construite à partir des titres de la Classification Internationale des Brevets (IPC), sert de base à un générateur d'hypernymes à base de transformateur, visant à faciliter la rédaction de brevets et à optimiser la variation lexicale dans ce domaine.

Ce chapitre souligne l'importance de développer une telle taxonomie en raison du manque de ressources accessibles et adaptées aux brevets. La nécessité d'une telle ressource est d'autant plus cruciale que les taxonomies existantes ne répondent pas entièrement aux besoins spécifiques de la rédaction de brevets, en particulier pour la généralisation des termes dans les revendications.

Las section 5.1 et 5.2 sont sur la construction d’une taxonomie technique.

La construction de la taxonomie implique une approche minutieuse : d'abord, l'identification des classes pertinentes de la IPC pour un domaine spécifique, puis l'extraction et la hiérarchisation des termes techniques issus des titres de ces classes. Cette approche transforme les titres en une structure arborescente de termes techniques, enrichie ensuite par un générateur de hypernymes pré-entraîné. Cette taxonomie constitue une ressource précieuse pour la substitution lexicale, en fournissant une liste de termes hyperonymes et hyponymes cruciaux.

La taxonomie technique est particulièrement utile pour diverses applications telles que le tech-mining de brevets, la classification de brevets, la récupération de passages et l'extraction de mots-clés. Elle offre un cadre structuré et spécifique à un domaine pour l'analyse des brevets, améliorant ainsi l'exactitude et l'efficacité des applications liées aux brevets.

En outre, la construction de cette taxonomie technique implique plusieurs étapes clés. La première étape consiste à créer des arbres taxonomiques, suivie de la classification et du raffinement des relations hyperonymes/hyponymes. Ensuite, un générateur d'hypernymes basé sur les taxonomies établies est formé, augmentant la richesse et la profondeur de la taxonomie.

La section 5.3 du chapitre met l'accent sur l'intégration d'un reconnaisseur de termes avec la taxonomie pour faciliter la rédaction de brevets. Ce processus (Figure 3) comprend l'identification des termes substituables dans un texte de revendication de brevet, puis la recherche de leurs hypernymes correspondants dans la taxonomie. Si un hypernym n'est pas trouvé, un modèle de Langage à Transformateur est utilisé pour générer automatiquement des hypernymes potentiels.

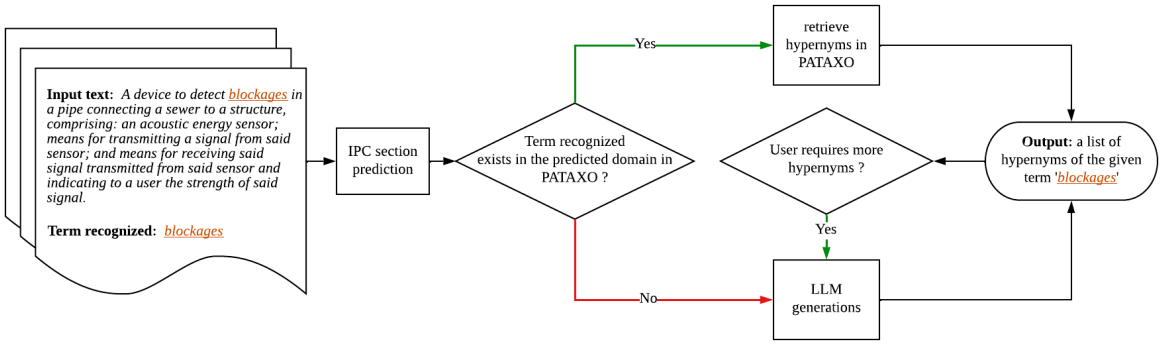


**Figure 3 - Processus de La variation lexicale dynamique orientée avec la construction d'une taxonomie de brevets**

Des modèles en anglais et en chinois ont été formés pour chaque domaine de la IPC. Pour l'anglais, le modèle FLAN-T5 a été affiné, et pour le chinois, le modèle zhuiyi-T5-pegasus a été utilisé. Ce dernier utilise un tokeniseur Bert adapté au chinois, avec un entraînement préalable en résumé de texte et paraphrase. Les taxonomies bilingues ont été transformées en paires terme-hypernym pour l'entraînement. Les performances des modèles ont été évaluées à l'aide de métriques telles que Hits@k et le Mean Reciprocal Rank (MRR).

Pour réaliser la variation lexicale, un corpus de test a été construit à partir de 1 000 phrases de revendications de brevets de chaque domaine de la IPC. Les termes substituables ont été extraits et des suggestions d'hypernymes ont été générées pour chaque terme. Cependant, les termes disponibles dans les taxonomies se sont souvent avérés trop généraux, manquant de la spécificité nécessaire pour des substitutions précises.

Une limitation notable est que les taxonomies se concentrent principalement sur les titres de niveau supérieur, ce qui limite la disponibilité d'hypernymes appropriés pour des termes plus spécialisés. Une solution serait d'intégrer des informations provenant de niveaux plus profonds des titres de la IPC. De plus, pour atténuer le problème de fréquence déséquilibrée des prédictions, il est suggéré de limiter les hypernymes aux niveaux inférieurs à "niveau 1" lors de la phase d'entraînement.

Le chapitre 5 conclut en soulignant l'importance et les défis de créer un système efficace pour la variation lexicale dynamique orientée dans le contexte de la rédaction de brevets. Bien que des progrès significatifs aient été réalisés, des améliorations sont encore nécessaires, notamment pour augmenter la spécificité des termes et équilibrer la fréquence des prédictions.

## Conclusion

En conclusion, cette étude présente une analyse critique de la recherche récente dans le domaine de la linguistique computationnelle et de l'analyse des brevets, tout en soulignant les perspectives d'avenir et les possibilités d'amélioration.

L'étude a réalisé des avancées notables, mais certains aspects nécessitent une attention et un développement supplémentaires. Premièrement, la quantité limitée de phrases dans notre arbre syntaxique (treebank) a conduit à des résultats de parsing sous-optimaux, soulignant le besoin d'élargir ce treebank. De plus, l'absence de numéros de référence d'images dans les phrases de l'arbre syntaxique et l'utilisation exclusive de la section “G” limitent la portée de notre analyse. Les recherches futures devraient approfondir la reconnaissance des termes et la variation lexicale, intégrant une évaluation manuelle par des experts pour améliorer la précision et la pertinence.

L'introduction de la nouvelle version de la Classification Internationale des Brevets (IPC) et la potentielle inclusion de la Classification Coopérative des Brevets Chinois (CPC) représentent des avancées significatives dans notre domaine. Ces développements reflètent non seulement l'évolution des classifications de brevets, mais offrent également de nouvelles opportunités pour l'exploration linguistique et analytique.

À l'avenir, l'application de la méthode tf-idf pour l'identification des termes candidats, l'exploration de différentes méthodes de sélection telles que les variations de la C-value et de l'Information Mutuelle Ponctuelle (PMI), et la mise en œuvre d'un parseur au niveau des caractères sur les titres de la taxonomie sont des étapes essentielles. Ces techniques permettront une analyse plus nuancée et précise des documents de brevets.

De plus, la fusion des taxonomies avec d'autres ressources pour créer une base de données dynamique et plus complète est un axe de recherche crucial. Cette approche, associée à un traitement plus détaillé de domaines spécifiques (comme les substances chimiques dans la section “C” de la IPC) et à la gestion des conjonctions dans les taxonomies, améliorera considérablement la profondeur et l'utilité de nos données.

Bien que cette étude ait jeté des bases solides et introduit de nouvelles perspectives, le chemin à parcourir est riche en opportunités pour une exploration et un raffinement

supplémentaires dans l'interaction complexe entre la linguistique computationnelle, l'analyse des brevets et la science des données.