



**HAL**  
open science

# Tech-mining on Chinese Patents: syntax and Terminology

Yixuan Li

► **To cite this version:**

Yixuan Li. Tech-mining on Chinese Patents: syntax and Terminology. Linguistics. Université de la Sorbonne nouvelle - Paris III, 2024. English. NNT : 2024PA030015 . tel-04697099

**HAL Id: tel-04697099**

**<https://theses.hal.science/tel-04697099v1>**

Submitted on 13 Sep 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ SORBONNE NOUVELLE

École Doctorale (ED 622) Sciences du Langage  
Laboratoire de Phonétique et Phonologie (LPP), UMR 7018

A thesis submitted in partial fulfilment of the requirements  
for the degree of Doctor of Philosophy in Linguistics:

# Tech-mining on Chinese Patents

Syntax and Terminology

LI Yixuan

Directed by Kim Gerdes

Jury:

Agata SAVARY, Université Paris-Saclay, professeur, rapporteur

Pierre MAGISTRY, INALCO, maître de conférences, examinateur

Chunshan XU, Anhui Jianzhu University, professeur, rapporteur

Huaping ZHANG, Beijing Institute of Technology (BIT), professeur, examinateur

Kim GERDES, Université Paris-Saclay, professeur, directeur



# Abstract

## **Title: Tech-mining on Chinese Patents: Syntax and Terminology**

This thesis aims to contribute to the field of research by automating the generation of lexical variations for technical terms found in Chinese patent claims. It achieves this through two primary contributions. Firstly, a character-level dependency parser specifically pre-trained on Chinese patent claims is developed. This parser enables the analysis of the internal structure of the terms and thus avoids the long-existing segmentation problem in Chinese. Secondly, a technical taxonomy is constructed based on the titles of the International Patent Classification (IPC) system, providing promising hypernym/hyponym substitutes for the production of variants of a base claim text.

Chapter 1 serves as an introduction, providing the necessary linguistic and technical background for the research. In Chapter 2, the collection and preprocessing of the corpus used in the study are detailed. Chapters 3 and 4 focus on annotating the Chinese character-level dependency treebank and describe the training process used to bootstrap the parser. Chapter 5 presents the construction and evaluation of the technical taxonomy, which utilizes the IPC system. Finally, in the end of Chapter 5, the methodology for recognising and selecting lexical variations is demonstrated.

**Keywords: dependency parsing, Chinese morphology, terminology, automatic term extraction, lexical variation, term substitution, taxonomy, patent**

## **Résumé en français**

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

Cette thèse vise à contribuer au domaine de la recherche en automatisant la génération de variations lexicales pour les termes techniques présents dans les demandes de brevet chinoises. Elle réalise cela grâce à deux contributions majeures. Tout d'abord, un analyseur de dépendance au niveau des caractères, spécifiquement pré-entraîné sur les demandes de brevet chinoises, est développé. Cet analyseur permet d'analyser la structure interne des termes et évite ainsi le problème de segmentation qui existe depuis longtemps en chinois. Deuxièmement, une taxonomie technique est construite en se basant sur les titres de la Classification internationale des brevets (IPC), fournissant des substituts prometteurs d'hyperonymes/hyponymes pour la production de variantes d'un texte de demande de brevet de base.

Le chapitre 1 sert d'introduction, en fournissant les connaissances linguistiques et techniques nécessaires à la recherche. Le chapitre 2 détaille la collecte et la préparation du corpus utilisé dans l'étude. Les chapitres 3 et 4 se concentrent sur l'annotation de l'arbre de dépendance au niveau des caractères chinois et décrivent le processus d'entraînement utilisé pour démarrer l'analyseur. Le chapitre 5 présente la construction et l'évaluation de la taxonomie technique, qui utilise le système de la Classification internationale des brevets. Enfin, à la fin de chapitre 5, la méthodologie de reconnaissance et de sélection des variations lexicales est démontrée.

**Mots-clé: analyse de dépendance, morphologie chinoise, terminologie, extraction automatique de termes, variation lexicale, substitution de termes, taxonomie, brevet**

## 中文简介

标题: 中文专利技术挖掘: 句法与术语

本论文旨在通过自动化生成中国专利申请中技术术语的词汇变体, 为研究领域做出贡献。它通过两个主要贡献来实现这一目标。首先, 我们开发了一个专门针对中国专利申请进行预训练的字符级依存句法关系解析器。该解析器可以分析术语的词内部结构, 从而避免了汉语中长期存在的分词问题。其次, 我们基于国际专利分类(IPC)系统的标题构建了一个技术分类树状词库, 为基于所给权利要求文本生成变体时提供可信的上义词/下义词替代词。

第1章作为引言, 为研究提供了必要的语言和技术背景。第2章详细介绍了所使用语料库的收集和预处理过程。第3章和第4章侧重于描述标注中国字符级依存树库和依存句法解析器的训练过程。第5章介绍了基于国际专利分类(IPC)标题的专利术语分类树状词库的构建和评估。最后, 在第5章的最后, 我们展示了识别和选择上下位词变体的方法。

关键词: 句法依存分析、汉语词法、术语学、自动术语提取、词汇变换、术语替换、分类法、专利

# Acknowledgements

First and foremost, I would like to express my heartfelt gratitude to my director, Kim Gerdes, for his unwavering support, guidance, and mentorship throughout the course of my doctoral journey.

During the course of my research journey, I encountered a series of formidable challenges that put my determination to the test. These challenges included the scarcity of Chinese language resources, encompassing treebanks, language models, WordNet equivalents, and terminology databases. The daunting nature of these obstacles necessitated innovative solutions. Conducting research from the vantage point of France, while advantageous in many aspects, brought about its own set of constraints. Access to certain critical resources and experts was limited, underscoring the need for adaptability and resourcefulness. Additionally, my initial unfamiliarity with the intricacies of patent-related studies compelled me to seek collaboration with domain experts. Their guidance and expertise proved to be invaluable in navigating the complexities of this specialized field. The complexity of conducting a cross-domain study without domain-specific knowledge presented a steep learning curve. It required a meticulous approach to ensure the validity and rigour of the research.

The insights and encouragement of you have been instrumental in shaping this research.

I would also like to express my heartfelt gratitude to every member of the Qatent research group who generously shared their expertise, provided guidance, and offered valuable insights throughout this thesis. Your collective wisdom has been a continuous source of inspiration and growth.

Moreover, I wish to take a moment to remember Isabelle Tellier, whose untimely passing due to cancer at the outset of this thesis left an irreplaceable void. Isabelle's presence, wisdom, and unwavering support are deeply missed, and her memory continues to inspire me.

I am immensely thankful to Thierry Poibeau, who extended a warm and supportive welcome during my initial year at Lattice after Isabelle's passing.

I extend my sincere appreciation to all the members of the jury who dedicated their time, expertise, and insights to evaluate this work, with special gratitude to the two pre-reporters.

To everyone who played a role in bringing this thesis to fruition, I extend my heartfelt thanks. Your contributions, unwavering support, and belief in this research are deeply appreciated.

Last but certainly not least, I would like to acknowledge the generous financial support provided by the CSC scholarship. It is through this support that my academic journey has been made possible, and I am profoundly grateful for the opportunity it has afforded me.

# Table of contents

<a href="#">Abstract</a>	3
<a href="#">Acknowledgements</a>	5
<a href="#">Table of contents</a>	6
<a href="#">Introduction</a>	1
<a href="#">What is Tech-mining?</a>	1
<a href="#">Motivation: Why tech-mining on Chinese patents?</a>	2
<a href="#">Objective and Plan of the Thesis</a>	4
<a href="#">Chapter 1 - Linguistic and technical background of the research</a>	5
<a href="#">1.1 Patents as Cutting-Edge Technological Innovations</a>	6
<a href="#">1.1.1 What Is a Patent?</a>	6
<a href="#">1.1.1.1 The General Structure and Components of a Patent</a>	7
<a href="#">1.1.1.2 Hypernymy in the Patent Drafting</a>	9
<a href="#">1.1.2 Patent Classification Systems</a>	10
<a href="#">1.2 Domain of the Study</a>	13
<a href="#">1.2.1 Review of Modern Chinese Morphology and Syntax</a>	13
<a href="#">1.2.1.1 The Phrasal Structures in Modern Chinese</a>	14
<a href="#">1.2.1.2 The Parallelism Between Compound Word Structure and Syntactic Structure in Chinese</a>	17
<a href="#">1.2.1.3 Wordhood in Chinese</a>	28
<a href="#">1.2.1.4 Dependency Syntax Analysis and Existing Chinese Treebanks</a>	34
<a href="#">1.2.1.5 Segmentation and Syntactic Parsing of Chinese Patents</a>	40
<a href="#">1.2.2 Review of Terminology Studies</a>	41
<a href="#">1.2.2.1 The Origin and Development</a>	41
<a href="#">1.2.2.2 A Very Short History of Terminology in China</a>	43
<a href="#">1.2.2.3 Resources for Chinese Technical Terms</a>	46
<a href="#">1.3 Defining Notions</a>	50
<a href="#">1.3.1 Unithood in the context of patents</a>	52
<a href="#">1.3.2 Termhood: distinguishing patent-domain specific terms from technical domain specific terms</a>	52
<a href="#">1.4 Hypothesis, Difficulties and Methodology</a>	57
<a href="#">Chapter 2 - Collection and Preprocessing of Patent Texts</a>	59
<a href="#">2.1 Collecting patent application data from CNIPA</a>	60
<a href="#">2.2 The general structure of a patent application</a>	62
<a href="#">2.3 From XML files to less space-consuming raw texts in a unified format</a>	65
<a href="#">2.4 Classification into IPC domains</a>	69
<a href="#">2.5 The Writing Style and Linguistic Specificities of Chinese Patents</a>	73
<a href="#">Chapter 3 - Syntactic Analysis of Chinese Patent Claims on Character-level</a>	81
<a href="#">3.1 A New Character-level Annotation Schema of Morphosyntactic Relations in Chinese</a>	82
<a href="#">3.1.1 Possible Distributional and Semantic Tests for Chinese Character Part-of-Speech and inter-characters Relations</a>	82
<a href="#">3.1.1.1 Tests for Part-of-Speech of the Characters</a>	82
<a href="#">3.1.1.2 Tests for Inter-characters Relations</a>	90
<a href="#">3.1.1.2.1 Coordination compounds</a>	90

## Table of contents

3.1.1.2.2	Arributive compounds	95
3.1.1.2.3	Subject-predicate compounds	98
3.1.1.2.4	Predicate-object compounds	98
3.1.1.2.5	Predicate-complement compounds	99
3.1.1.2.6	Non-compound Terms with Unclear Internal Structures	100
3.1.1.2.7	Terms Composed of More Than Two Characters	101
3.1.2	Hierarchizing the Morphological and Syntactic Relation Labels for Character-level Treebanks	105
3.1.3	The Final Decision Tree	108
3.2	Construction and Automatic Pre-annotation Processing of the Character-level Chinese Patent Treebank	115
3.3	Manual Annotation and Correction of the Pre-Annotated Chinese Treebanks	117
3.3.1	Frequent Syntax Structures Specific to Chinese Patent Claims	117
3.3.2	Practical Choices for Challenging Examples	119
3.3.2.1	Content terms with @m	120
3.3.2.2	Function words constructed with @m	126
3.3.2.3	Choices involving the assignment of @m	130
3.3.2.4	Problematic syntactic structures	132
3.4	The first treebank on Chinese patent claims at character-level	142
Chapter 4	- Joint Segmentation/ Dependency Parsing of Chinese Patent Claims	145
4.1	Automatic Syntax Analysis in Chinese	146
4.2	Word Segmentation by Parsing with “@m”	152
4.2.1	Fine-tuning the Bert-based Parser	152
4.2.2	Error Analysis	155
4.3	Dependency Parser-based Term Recognition for Chinese Patents	165
4.3.1	A Brief Presentation of Term Recognition	165
4.3.2	Term Recognizer based on Dependency Relation Information	167
4.3.3	Comparison with the POS Matcher	173
Chapter 5	- Lexical Variation with the Construction of a Patent-related Taxonomy	177
5.1	Leveraging Technical Knowledge in IPC Titles	179
5.1.1	The Original Format of the IPC Titles	179
5.1.2	Alignment of English-Chinese IPC Titles	182
5.2	Mining Hypernyms/Hyponyms in IPC Titles	184
5.2.1	The Construction of the Tree Structure	184
5.2.1.1	Filtering Inappropriate Titles or Irrelevant Information in the Titles	185
5.2.1.2	Extraction of Technical Expressions	189
5.2.1.3	Nominalization of Incomplete Technical Expressions	191
5.2.2	Evaluation and Pruning	194
5.3	Oriented Dynamic Lexical Variation	197
5.3.1	Training a Transformer-based Hypernym Generator	197
5.3.2	Substitution of Terms	199
Conclusion and Outlook		206
References		207
Appendix		219





# Introduction

Patents represent the largest source of technical information in the world, comprising approximately 90% to 95% of global scientific and technological knowledge (Chen & al., 2006). How to effectively harness such a vast information resource plays a crucial role in various aspects of research and patent-related activities.

The use of patent information has become increasingly important in various fields, including technology intelligence, innovation management, and scientific research. Patent information contains valuable information about the latest technological developments, the competitive landscape, and emerging trends. However, patent information is often difficult to access and analyze due to the vast amount of data and the complexity of the documents.

In addition to the challenges in accessing and analyzing patent information, attorneys who draft new patents face the additional challenge of ensuring that their patent applications are comprehensive and accurately describe their clients' inventions. This requires a deep understanding of the technical language and concepts related to the invention, as well as the ability to identify and use appropriate terminology to describe the invention in a clear and precise manner.

This study's primary objective is to facilitate the access to the Chinese patent texts for both the patent drafting and the automatic analysis of patents.

## What is Tech-mining?

Tech-mining is short for "text mining of science & technology information resources." First introduced by Alan L. Porter in his book *Tech Mining for Future-oriented Technology Analysis* (2009), the tech-mining is defined as following:

*"Tech-mining is the application of text-mining tools to science and technology information, informed by understanding of technological innovation processes."*

In the past, technology mining relied on conventional approaches such as patent analysis, literature reviews, and expert consultations for gathering insights into technological advancements. These methods were not only time-consuming but also had limited coverage. However, in the era of digital transformation and the availability of extensive datasets, technology mining has undergone a significant transformation.

Today, technology mining harnesses cutting-edge computational techniques, data mining algorithms, Machine Learning, Natural Language Processing, and network analysis to extract valuable insights from vast datasets. It draws information from diverse sources, including scientific publications, patents, technical reports, conference proceedings, funding databases, innovation repositories, and online platforms, enabling the comprehensive tracking and analysis of technology trends.

The applications of technology mining are multifaceted. It empowers organizations to identify emerging technologies and evaluate their potential impacts on markets, industries, and strategic business decisions. It facilitates technology scouting and open innovation initiatives by identifying

external technologies suitable for integration or licensing. Moreover, technology mining plays a pivotal role in shaping policy development and public decision-making processes, aiding in technology roadmapping, research funding prioritization, and regulatory framework guidance.

Among various types of technical and scientific texts, patents stand out as one of the most valuable resources. When technology mining focuses on patent analysis, it is commonly referred to as ‘patent mining’. This specialized approach delves deep into patent datasets to uncover critical insights into technological innovations and intellectual property landscapes.

Furthermore, technology mining serves as an indispensable tool in supporting research and development (R&D) endeavours. It assists researchers in pinpointing pertinent literature, identifying research gaps, and uncovering potential collaborative opportunities. Additionally, it streamlines technology transfer and commercialization efforts by identifying licensing prospects and potential industry collaborators.

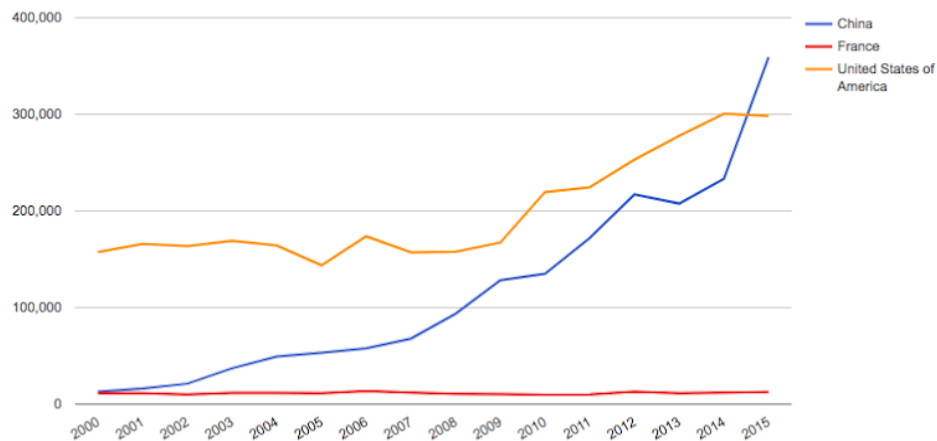
In this thesis, our focus shifts from the conventional task of predicting technology trends in tech-mining to a more oriented approach involving augmented inventing (Lee et Hsiang, 2019) and its fundamental processes.

In this context, we consider augmented invention, which by definition involves humans utilizing AI to comprehend vast amounts of data and interpret suggestions provided by the AI, as a specialized application of tech-mining. Augmented invention may also be seen as an approach to bridge technological gaps through human decision-making, utilizing existing technological datasets.

## Motivation: Why tech-mining on Chinese patents?

In the context of the rapid advancement of science, technology, and global business strategies, patents continue to hold a significant role even after more than five centuries since their systematic granting began in 1450 in Venice. In the 21st century, often referred to as the information age, intellectual property protection has become increasingly crucial. With the proliferation of inventions and the emergence of new technologies in recent decades, the field of patents has gained unprecedented importance. China, experiencing an economic and technological boom, has witnessed a remarkable surge in patent applications across all technological domains since the turn of the century. As a result, there is an urgent need within the Chinese industrial landscape to enhance the efficiency of patent application and maintenance procedures.

## Introduction



**Figure - The Number of Patent Applications Over the Years in China, France, and the United States of America**

In the Figure above we can see a dramatic growth of the blue line that presents the number of granted patents in China between 2000 and 2015<sup>1</sup>.

Traditional methods of patent drafting, like all other legal textual work, often require considerable manual force. In recent years, under the impact of Machine Learning, several attempts have been made to take advantage of this high-potential business, from the Cloem.com<sup>2</sup> to the collaboration of the World Intellectual Property Organization (WIPO)<sup>3</sup> with Google on the automatic translation of patents. At the same time, in China, some companies have emerged in the field such as WenXian Technology<sup>4</sup>, one of the pioneers of automatic patent composition, which seeks to accelerate patent drafting work by employing Machine Learning.

In spite of this rise of interest and these few pioneers, the field of automatic processing of Chinese patent texts remains generally underdiscovered. This is especially true for the theoretical modeling of patent texts (which can serve as a base for text simplification, syntactic parsing, or the generation of texts), whereas the research on machine translation and text mining has always been hot topic ever since the popularization of the computer. From this lack of studies comes the interest of this thesis at the crosslines of law, linguistics, Natural Language Processing, and Machine Learning.

---

<sup>1</sup> Data from the official site of WIPO

<sup>2</sup> Cloem is a company based in Cannes, France, which applies natural language processing (NLP) technologies to assist patent applicants in creating variants of patent claims, called "cloems". (<https://www.cloem.com/>)

<sup>3</sup> <http://www.wipo.int>

<sup>4</sup> <http://www.wxcip.com/>

## Objective and Plan of the Thesis

In this study, we are interested in terms of Chinese patents, especially in the claims. The analysis is focused on their linguistic characteristics, especially morphological characteristics, their syntactic functions and their variation (scope of the claims).

The objective is to describe the internal structure of Chinese technical terms with dependency syntax annotation, and to explore the possibility of the term recognition and the lexical variation feeding the special needs of claim scoping with the help of the construction of a new technical taxonomy.

The organization of the thesis is in five chapters.

We first start with the presentation of the linguistic and technological background of the study by revisiting the morphological and terminological theories in Chinese and we discuss the notions used in this study. At the end of Chapter 1, we will give the principle hypothesis and our methodology.

In Chapter 2, we introduce in detail how we collect and preprocess the patent application data.

Moving to Chapter 3, which is the central interest of this study, we annotate the first Chinese patent treebank, and also the first treebank at character-level, which contains rich morphological information. The parallelism between compound word structure and syntactic structure in Chinese is the theoretical base of the annotation. The final results are published in both the Universal Dependency (UD) project and the Surface-Syntactic Universal Dependency (SUD) project.

Chapter 4 and Chapter 5 are two explorations of the tech-mining on Chinese patents by using the character-level dependency parser as a base. Chapter 4 introduces the automatic term extraction (ATE) based on dependency relations instead of the conventional methods based on the POS tag pattern. Chapter 5, which is based on the results of Chapter 4, presents the construction of a patent-related taxonomy that can serve as a base of lexical variation.

# Chapter 1 - Linguistic and technical background of the research

In the first chapter of this thesis, we provide an overview of the research orientations and main interests of scholars in the field of patent-mining.

Section 1.1 focuses on the patent as the corpus of analysis, including its general structure and scope in intellectual property protection. We also discuss the previous research on Natural Language Processing and tech-mining in the patent domain. In Section 1.2, we take a theoretical and linguistic perspective to introduce the domains of study, specifically Chinese word formation theories, syntax, and terminology. Section 1.3 defines the central notion of “terms” in our study and clarifies the fundamental concepts that will be employed throughout the following sections. Building on this background, Section 1.4 presents a list of hypotheses of linguistic and technical interest, along with the difficulties encountered during the research process and our methodology.

## 1.1 Patents as Cutting-Edge Technological Innovations

### 1.1.1 What Is a Patent?

In WIPO's definition<sup>5</sup>, a patent is “an exclusive right granted for an invention, which is a product or a process that provides, in general, a new way of doing something, or offers a new technical solution to a problem. To get a patent, technical information about the invention must be disclosed to the public in a patent application. In principle, the patent owner has the exclusive right to prevent or stop others from commercially exploiting the patented invention. In other words, patent protection means that the invention cannot be commercially made, used, distributed, imported or sold by others without the patent owner's consent.”

Although patents have scope only in the jurisdiction where they are granted, their definition is pretty much the same internationally, which allows to demand protection for the same invention around the world, if this is needed - and affordable - to the inventor.

The intellectual property law in China recognizes three distinct types of innovations: inventions, utility models, and industrial designs.

According to Article 2 of the Patent Law of the People's Republic of China (2020)<sup>6</sup>:

*The term “invention” as used in this Law refers to inventions, utility models, and designs.*

*An invention refers to a new technical solution proposed for a product, method, or its improvement.*

*A utility model refers to a new technical solution proposed for the shape, structure, or combination of a product that is suitable for practical use.*

*A design refers to a new design that is aesthetically appealing and suitable for industrial application, encompassing the overall or partial shape, pattern, or their combination, as well as the combination of color and shape or pattern.*

In this study, our focus is solely on invention patents, and we do not consider utility model patents or design patents.

Among all types of texts in the field of technology, patents protecting inventions play a particular role because they legally delimit a privatized technological domain, but also because of their very particular style of legal sub-language, in which the authors attempt to cover as much technological “space” as possible in its scope.

---

<sup>5</sup> <https://www.wipo.int/patents/en/>

<sup>6</sup> <https://www.wipo.int/wipolex/en/legislation/details/21027>

### 1.1.1.1 The General Structure and Components of a Patent

A patent must consist of three textual parts: the abstract, the description and the claims.

- Abstract

The abstract is a concise summary of the invention, highlighting its key aspects and potential benefits. It should provide a clear overview of the invention's technical features and its significance. It is legally limited in size to around 150 words.

- Description

The description provides a detailed explanation of the invention, including its technical aspects, how it works, and its practical implementation. It should be thorough enough for someone skilled in the relevant field to understand and replicate the invention. The length of a patent application's description can vary widely based on the technology, the complexity of the invention, the jurisdiction, the strategy of the patent attorney, and other factors. A patent's description averages at around 20 to 30 pages, but some patents might be just a few pages long, while others could be several hundred pages. Patents in fields like software or mechanical devices might generally be shorter, while patents in fields like pharmaceuticals or complex electronics tend to be longer due to the intricate details and data that need to be provided.

- Claims

The claims section of a patent outlines the specific features and elements that define the scope of protection granted to the invention. These claims serve as the basis for legal protection against infringement and unauthorized use.

A patent claim is typically a single sentence, often exceeding 100 words, and can fall into two main categories: product claims and method claims. Product claims describe the objects, arrangements, compounds, or other tangible inventions, while method claims focus on manufacturing methods, utilities, or other processes.

Patent claims play a pivotal role in defining the rights granted to the patent owner and determining the scope of protection for the invention. They outline the specific features of the patented product or process and serve as the reference point for future legal actions against potential infringement or unauthorized use by a third party. Crafting well-drafted claims is essential, as they must strike a balance between being broad enough to cover various use cases and specific enough to satisfy the requirements of the patent office.

In a patent application, claims can be categorized as either independent or dependent. Independent claims are self-standing and do not explicitly reference any prior claims, whereas dependent claims incorporate and refer back to one or more previous claims.

The protection scope of an independent claim in a patent application is the broadest for a particular invention. If a claim encompasses all the technical features of another claim of the same type and further narrows down the technical solution of that other claim, it becomes a dependent claim. Dependent claims, by introducing additional technical features or further limiting the technical



features of the referenced claims, fall within the protection scope of the claims they refer to. The additional technical features in dependent claims can either provide further limitations on the technical features of the referenced claims or introduce entirely new technical elements. In a patent application, there should be at least one independent claim. When there are two or more independent claims, the one placed at the beginning is referred to as the first independent claim.

According to their nature, patent claims can be classified into two basic types: claims for things and claims for activities, often referred to as product claims and method claims. The first basic type of claims includes objects produced by human technology (products, equipment), while the second basic type of claims includes activities that involve elements of time processes (methods, uses). Claims for things typically cover items, substances, materials, tools, devices, and similar items. On the other hand, claims for activities encompass manufacturing methods, usage methods, communication methods, processing methods, and methods for using a product for specific purposes, among others.

In China, the application for a patent is governed by the Patent Law of the People's Republic of China and its implementing regulations. Another important document for patent drafting is the Patent Examination Guidelines issued by the China National Intellectual Property Administration (CNIPA), the Chinese patent office.

During the examination process of a patent application, the features and novelty of the invention described in the application serve as a reference for the decision-making process of the patent office. These features are considered to assess the uniqueness and inventiveness of the proposed invention, and they play a crucial role in determining whether the application meets the requirements for patentability.

Once a patent is granted, the claims in the patent document become the basis for defining the scope of protection conferred by the patent. The description only serves to clarify the claims but does not provide protection for ideas that are not "claimed".

According to Article 64 of the Patent Law of the People's Republic of China (2020):

*"The protection scope of an invention or utility model patent is determined by the content of its claims, while the description and drawings in the patent specification can be used to interpret the content of the claims."*

The claims specify the precise features and elements that are protected by the patent and establish the boundaries within which the patent owner has exclusive rights. The claims serve as a legal tool that helps clarify and define the extent of protection afforded to the patented invention.

To ensure the clarity and precision of the claims, they must be supported by the description and drawings provided in the patent application. Every word in the claims must be carefully chosen to accurately and comprehensively define the invention. The drafting of claims follows specific criteria and guidelines dictated by the patent laws of the respective country or region. This process is necessary to protect the inventor's rights and provide a solid foundation for any legal actions taken against potential infringement.

### 1.1.1.2 Hypernymy in the Patent Drafting

Hypernymy, a linguistic phenomenon denoting a hierarchical relationship in which one term (the hypernym) represents a broader category encompassing another term (the hyponym), holds significant potential within the realm of patent drafting. Its strategic utilization can significantly contribute to augmenting the quality, precision, and scope of patent documents, particularly in the context of broad claim coverage and robustness for further prosecution.

Within the context of patent Case Law, where legal certainty is a cornerstone, the articulation of technical features constituting a patent's claim for protection demands meticulous precision, devoid of ambiguity. This adherence to precision leaves little room for interpretation. A patent application is examined by a specialist, called the examiner, in the patent office. Their role is to check the patentability of the patent application. In general, they try to reduce the scope of the application, and the patent attorney wants to extend the scope to the maximum, in order to protect all imaginable and future configurations.

Suppose we claim “*a mouse with a screen*”. Suppose a mouse is a hyponym of “computing device” and “screen” is a hyponym of “display”. Then it is better to use in the description the more general expressions or variants resulting in “*a computing device with a screen*”, “*a mouse with a display*” or “*a computing device with a display*”, for later reactions against alleged collisions in the prior art, that might be detected by the patent examiner in the patent office. In such a case, if support<sup>7</sup> permits, it is allowed to reshape the pending claims, in the example “*a mouse with a display*”. If claim variants haven't been anticipated, it is not allowed to add subject-matter after filing is done.

As another example, consider a scenario in which a patent application explicitly employs the term an “armoured vehicle.” In such a case, attempting to later excise a feature, for example in a desired claim, a “vehicle” can be questionable as the omission of the modifier “armoured” introduces lack of support and inherent uncertainties. Similarly, the assertion of a claim for a “bulletproof vehicle” is precluded, if the term “bulletproof” is absent from the original description.

As another example, consider the chain “tricycle”, “bicycle”, “vehicle” and “transportation system”. Mastering the paths between words is critical to be able to generalize or specify a given entry expression.

Thus a good practice for drafting is to anticipate as much as possible different combinations of words (with some words levelling up, while others are levelled down; these are called “intermediate generalizations”). This practice allows extracting the appropriate combinations of different abstraction levels.

A patent application essentially serves as a **repository of linguistic constructs**, providing the foundation for claim amendments through procedural mechanisms akin to a “copy and paste” operation (e.g. Article 123 EPC<sup>8</sup>). Consequently, it is incumbent upon the applicant to diversify textual perspectives and encompass a spectrum of linguistic formulations. This strategic approach bolsters the application's resilience during the examination phase, where comparisons with antecedent

---

<sup>7</sup> Support refers to the rest of the patent application where these more general terms have to be mentioned and explained in the context of the invention.

<sup>8</sup> European Patent Convention (<https://www.epo.org/en/legal/epc/2020/a123.html>)

references from the prior art prompt the applicant to innovate in linguistic expression, unveiling novel combinations of verbiage.

Broad claim coverage, a fundamental goal in patent drafting, can be effectively achieved by strategically incorporating hypernyms into the language of patent claims. Hypernyms, as expansive, higher-order terms, possess the capacity to encapsulate multiple specific concepts or embodiments within a single claim. This approach significantly amplifies the scope of protection conferred by the patent claims without necessitating the exhaustive enumeration of every potential variation.

The drafting and generalization of patent claims involve determining the technical features that constitute the technology solution for which patent protection is sought.

In conclusion, the astute application of hypernymy in patent drafting not only enhances precision and clarity but also expands the protective ambit of patent claims. By embracing broader linguistic categories, patent applicants can future-proof their intellectual property, aligning it with the dynamic landscape of evolving technologies and legal interpretations. This strategic integration empowers patent holders to fortify their positions, ensuring the longevity and adaptability of their patents in a rapidly changing intellectual property landscape.

And considering the importance of the hyperonymy/hyponymy relations in patent drafting and examination, we build the patent-related technical taxonomy in Chapter 5.

## 1.1.2 Patent Classification Systems

There are several patent classification systems that are used in different countries and regions. In this section, we will focus on two significant classification schemas, namely the International Patent Classification (IPC) and the Cooperative Patent Classification (CPC), providing a detailed introduction to each of them.

The International Patent Classification (IPC) is a standardized hierarchical system that is employed by more than 100 countries worldwide to ensure consistent classification of patent documents. Its establishment dates back to 1971 under the Strasbourg Agreement, and it operates under the auspices of the World Intellectual Property Organization (WIPO). Regular updates to the IPC are facilitated through the collaboration of a committee comprising experts from participating countries and observers from organizations such as the European Patent Office.

The IPC's hierarchical structure<sup>9</sup>, consisting of eight sections, facilitates the systematic organization of terminological expressions in fields characterized by innovation and enhances the understanding of interrelationships among technological concepts within specific knowledge domains. The eight sections of the IPC are A. Human necessities; B. Performing operations; transporting; C. Chemistry; metallurgy; D. Textiles; paper; E. Fixed constructions; F. Mechanical engineering; lighting; heating; weapons; blasting engines or pumps; G. Physics; H. Electricity. Each section further incorporates

---

<sup>9</sup>

<https://ipcpub.wipo.int/?notion=scheme&version=20230101&symbol=none&menulang=en&lang=en&viewmode=f&fipcpc=no&showdeleted=yes&indexes=no&headings=yes&notes=yes&direction=o2n&initial=A&cwid=none&tree=no&searchmode=smart>

## Chapter 1 - Linguistic and technical background of the research

classes, subclasses, groups, and subgroups, which are identified by alphanumeric codes (e.g., A01, A01B, A01B 1/00, A01B 1/22). Shown in Figure 1.1 and Figure 1.2.

-		A	<b>HUMAN NECESSITIES</b>
			<b>AGRICULTURE</b>
D	-	A01	AGRICULTURE; FORESTRY; ANIMAL HUSBANDRY; HUNTING; TRAPPING; FISHING
D	⚠ +	A01B	SOIL WORKING IN AGRICULTURE OR FORESTRY; PARTS, DETAILS, OR ACCESSORIES OF AGRICULTURAL MACHINES OR IMPLEMENTS, IN GENERAL (making or covering furrows or holes for sowing, planting or manuring A01C 5/00; machines for harvesting root crops A01D; mowers convertible to soil working apparatus or capable of soil working A01D 42/04; mowers combined with soil working implements A01D 43/12; soil working for engineering purposes E01, E02, E21)
D	+	A01C	PLANTING; SOWING; FERTILISING (parts, details or accessories of agricultural machines or implements, in general A01B 51/00-A01B 75/00)
D	+	A01D	HARVESTING; MOWING
D	+	A01F	THRESHING (combines A01D 41/00); BALING OF STRAW, HAY OR THE LIKE; STATIONARY APPARATUS OR HAND TOOLS FOR FORMING OR BINDING STRAW, HAY OR THE LIKE INTO BUNDLES; CUTTING OF STRAW, HAY OR THE LIKE; STORING AGRICULTURAL OR HORTICULTURAL PRODUCE (arrangements for making or setting stacks in connection with harvesting A01D 85/00)
D	+	A01G	HORTICULTURE; CULTIVATION OF VEGETABLES, FLOWERS, RICE, FRUIT, VINES, HOPS OR SEAWEED; FORESTRY; WATERING (picking of fruits, vegetables, hops or the like A01D 46/00; propagating unicellular algae C12N 1/12)
D	+	A01H	NEW PLANTS OR PROCESSES FOR OBTAINING THEM; PLANT REPRODUCTION BY TISSUE CULTURE TECHNIQUES [5]
D	+	A01J	MANUFACTURE OF DAIRY PRODUCTS (for chemical matters, see subclass A23C)
D	+	A01K	ANIMAL HUSBANDRY; AVICULTURE; APICULTURE; PISCICULTURE; FISHING; REARING OR BREEDING ANIMALS, NOT OTHERWISE PROVIDED FOR; NEW BREEDS OF ANIMALS

Figure 1.1 - The first classes and subclasses of IPC Section A

-		A	<b>HUMAN NECESSITIES</b>
			<b>AGRICULTURE</b>
D	-	A01	AGRICULTURE; FORESTRY; ANIMAL HUSBANDRY; HUNTING; TRAPPING; FISHING
D	⚠ -	A01B	SOIL WORKING IN AGRICULTURE OR FORESTRY; PARTS, DETAILS, OR ACCESSORIES OF AGRICULTURAL MACHINES OR IMPLEMENTS, IN GENERAL (making or covering furrows or holes for sowing, planting or manuring A01C 5/00; machines for harvesting root crops A01D; mowers convertible to soil working apparatus or capable of soil working A01D 42/04; mowers combined with soil working implements A01D 43/12; soil working for engineering purposes E01, E02, E21)
	-	A01B 1/00	Hand tools (edge trimmers for lawns A01G 3/06) [2006.01]
	-	A01B 1/02	• Spades; Shovels [2006.01]
		A01B 1/04	• • with teeth [2006.01]
	-	A01B 1/06	• Hoes; Hand cultivators [2006.01]
		A01B 1/08	• • with a single blade [2006.01]
		A01B 1/10	• • with two or more blades [2006.01]
		A01B 1/12	• • with blades provided with teeth [2006.01]
		A01B 1/14	• • with teeth only [2006.01]
	-	A01B 1/16	• Tools for uprooting weeds [2006.01]
		A01B 1/18	• • Tong-like tools [2006.01]
		A01B 1/20	• Combinations of different kinds of hand tools [2006.01]
		A01B 1/22	• Attaching the blades or the like to handles (handles for tools, or their attachment, in general B25G); Interchangeable or adjustable blades [2006.01]
		A01B 1/24	• for treating meadows or lawns [2006.01]

Figure 1.2 - The first groups and subgroups of IPC subclass A01B

These classifications are accompanied by descriptive titles expressed as noun phrases (e.g. A01B 1/00 in Figure 1.2), participle phrases (e.g. A01B 1/24 in Figure 1.2) and prepositional phrases (e.g. A01B

1/04 in Figure 1.2). The IPC version of 2016<sup>10</sup> includes approximately 70,000 classification entries distributed among these subgroups.

Another more recent and widely applied classification schema is the Cooperative Patent Classification (CPC)<sup>11</sup>.

The Cooperative Patent Classification (CPC), hereinafter referred to as CPC, represents a collaborative patent classification initiative involving both the European Patent Office (EPO) and the United States Patent and Trademark Office. (USPTO). The CPC project was initiated on October 25, 2010, with David Kappos, Deputy Secretary of Commerce for Intellectual Property and Director of the USPTO, and Benoît Battistelli, President of the European Patent Office, signing a joint statement. The objective of this collaboration was to harmonize international classification systems and enhance search efficiency. To achieve this, both USPTO and EPO agreed to promote a common classification system based on the European Classification (ECLA).

The composition of the CPC classification system is as follows:

- Sections A to H, corresponding to the eight sections of the International Patent Classification (IPC).
- The addition of Section Y, which covers emerging fields. For example, Y02 pertains to technologies for mitigating climate change, Y04 focuses on smart grids, and it includes cross-domain and cross-referencing art collections and digests derived from the USPC.

The China National Intellectual Property Administration (CNIPA) has commenced the simultaneous classification of all newly filed Chinese invention patent applications under both the International Patent Classification (IPC) and the Cooperative Patent Classification (CPC) systems since January 1, 2016.

Regrettably, an official translated version of the Cooperative Patent Classification (CPC) system has not been made available by the CNIPA. Consequently, we are still reliant on the International Patent Classification (IPC) in our study.

---

<sup>10</sup> We work with this version because it is the only one available with an official translation into Chinese at the moment.

<sup>11</sup> <https://www.epo.org/en/searching-for-patents/helpful-resources/first-time-here/classification/cpc>

## 1.2 Domain of the Study

This section delves into previous research on syntactic analysis and terminological studies, specifically those related to patents. By examining existing works in these areas, we aim to build upon and contribute to the existing knowledge and understanding of Chinese language processing.

This study focuses on two key aspects of patent texts: their syntactic analysis and terminological features. In Section 1.2.1, we provide the research context of syntactic analysis of general Chinese texts, with a specific focus on dependency analysis. Section 1.2.2 lays the groundwork for understanding modern terminology studies.

### 1.2.1 Review of Modern Chinese Morphology and Syntax

In this section, we will present different theories of word formation in Chinese which will guide the annotation of the dependency relations between morphemes/characters that I will present in Section 3.1.

Introduced only after 1907 by Zhang Shizhao, the distinction between characters and a larger word-like units in Chinese is a comparatively unfamiliar and confusing concept even nowadays. With neither natural delimiters nor inflection marks, two main indicators of wordhood in today's languages using Latin letters (Magistry et Sagot, 2012), Chinese script is a continuous chain of characters, only separated by punctuation that is used similarly to European languages, but without anything indicating intermediate units such as words or phrases. Although the notion of "word" is fuzzy in all languages, with the absence of writing conventions on wordhood, the definition of words is particularly unclear and unnatural to the common Chinese speaker. One experiment (Sproat et al, 1996) shows that the rate of agreement on wordhood is only 76% among Chinese native speakers. Due to the fact that linguists have no common agreement on the definition of words in modern Chinese, word segmentation has always been a challenging task in Chinese Natural Language Processing.

In Chi et Lin (2019)'s article "Reconsidering the Parallelism of Chinese Compound Words Structure and Syntactic Structure", it is argued that the fundamental reason for this difficulty is that "*the formation of Chinese compound words mostly adopts syntactic means*" - contrary to languages with inflection where morphological indicators help to define wordhood.

Prominent linguists such as Lu Zhiwei (1957) and Lü Shuxiang and Zhu Dexi (1979) have suggested that "*the formation of disyllabic words is similar to phrases.*" Lu Zhiwei (1957) emphasized that "*one type, in terms of the relationships between its components, can be common to both word formation and sentence formation.*" Later, Zhu Dexi (1982) explicitly stated, "*the structure of compound words is parallel to syntactic structure.*" Some scholars even argue that "*the similarity and consistency between compound word structure and syntactic structure are widely accepted in academia.*" This viewpoint is also shared by scholars like Li Xingjian (1982), Wang Hongjun (1998), Ge Benyi (2001), Dong Xiufang (2011), Shao Jingmin (2016), among others. We adopt this point of view in our character-based syntactic analysis of Chinese to be presented in Section 3.1.

In this section, firstly, we will present the types of Chinese phrases (Section 1.2.1.1). Then, we will present the phenomenon of parallelism between Chinese morphology and syntax (Section 1.2.1.2). And we will introduce some studies on the concept of wordhood or the distinction between words and phrases in modern Chinese (Section 1.2.1.3).

After discussing Chinese wordhood, we will provide a brief introduction to dependency syntax theories and the existing treebanks in Chinese (Section 1.2.1.4), and will present recent works on segmentation and syntax parsing on patents (Section 1.2.1.5).

### 1.2.1.1 The Phrasal Structures in Modern Chinese

In the more recent work of Huang Borong and Liao Xudong (2007), Chinese phrases are categorized based on their structure. We will describe their classes in greater detail. Among the various phrase types in Chinese, five basic structures are the most commonly used and serve as the foundation for sentence syntax analysis. These five basic types of phrases are classified as follows:

- Subject-predicate phrases (主谓短语, zhǔ wèi duǎn yǔ): They consists of two components with a declarative relationship, where the preceding component serves as the subject, and the following component serves as the predicate. It corresponds to the syntactic relation “subj” of the SUD<sup>12</sup> treebank annotation schema.

- (1)     他                    说  
tā                        shuō  
‘He\_PRON’            ‘say\_V’  
‘He says’
- (2)     粮食                丰                收  
liáng shí            fēng                shōu  
‘grain\_N’            ‘abundant\_ADJ’    ‘harvest\_V/N’  
‘The food comes from an abundant harvest’

- Predicate-object phrases (动宾短语, dòng bīn duǎn yǔ): They consists of two components with a governing relationship, where the preceding component, typically a verb representing an action or behaviour, functions as the governing element, and the subsequent part serves as the object. It corresponds to the syntactic relation “comp:obj” of SUD treebank annotation schema.

- (3)     盖                    被子  
gài                     bèi zi  
‘cover\_V’            ‘covers\_N’  
‘Put the covers on’

- Attributive phrases (偏正短语, piān zhèng duǎn yǔ): They comprises two parts with a modifier relationship, where the modifier comes first, and the word being modified, known as the headword, follows. It corresponds to the syntactic relation “mod” of SUD treebank annotation schema.

<sup>12</sup> <https://surfacesyntacticud.github.io/guidelines/u/>

- Adjective attributive phrases (定中短语, dìng zhōng duǎn yǔ)

(4) 坏 话  
 huài huà  
 ‘bad\_ADJ’ ‘speech\_N’  
 ‘Negative comments’

- Adverbial attributive phrases (状中短语, zhuàng zhōng duǎn yǔ)

(5) 慢慢 说  
 màn màn shuō  
 ‘slowly\_ADV’ ‘speak\_V’  
 ‘Speak slowly’

- Predicate-complement phrases (述补短语, shù bǔ duǎn yǔ): They consists of two components in a complementary relationship. The first part is the predicate or headword, which is supplemented by the second part, the complement. It corresponds to the syntactic relation “comp” of SUD treebank annotation schema.

(6) 说 完  
 shuō wán  
 ‘speak\_V’ ‘finished\_ADJ’  
 ‘Done talking’

- Coordinative phrases (联合短语, lián hé duǎn yǔ): They consists of two or more components with equal grammatical status. These components are connected by coordinating relationships, which can include coordination, selection, etc. Sometimes, conjunctions like “and (和 hé, 并 bìng)”, “or”, etc., are used to connect them. It corresponds to the syntactic relation “conj” of SUD treebank annotation schema.

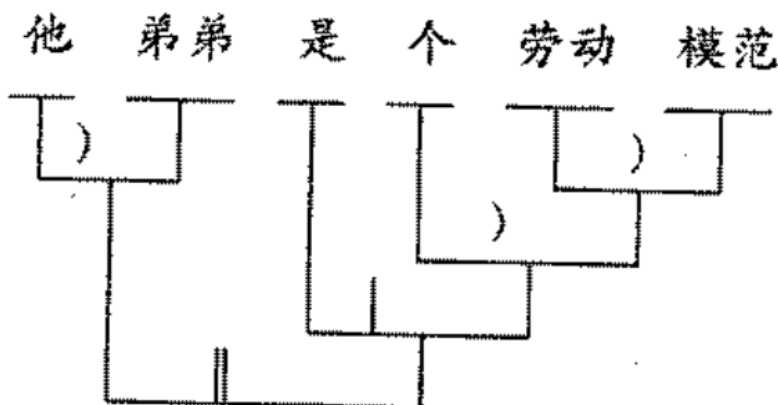
(7) 你 和 我  
 nǐ hé wǒ  
 ‘you\_PRON’ ‘and\_CCONJ’ ‘I/me\_PRON’  
 ‘You and me’

(8) 继承 并 发展  
 jì chéng bìng fā zhǎn  
 ‘inherit\_V’ ‘and\_CCONJ’ ‘develop\_V’  
 ‘Inherit and develop’

In Chinese, sentences are formed through the nested arrangement of various phrase structures as mentioned above. In Huang and Liao’s work (2007), there is an example sentence showing the nested arrangement.



- (9) 他 弟弟 是 个 劳动 模范  
 tā dì dì shì gè láo dòng mó fàn  
 'he' 'brother' 'is' 'a' 'work' 'model'



*'His brother is a model worker.'*

**Figure 1.3 - Example of Nested Sentence Structure (Huang et Liao, 2007)**

In the example in Figure 1.3, the first layer of character combination are attributive structures (tā dìdì, 他 弟弟; láodòng mófàn, 劳动 模范), the second layer is also an attributive structure (gè láodòng mófàn, 个 劳动模范), the third layer is a verb-object structure (shì gè láodòngmófàn, 是个 劳动模范), and the final fourth layer is a subject-predicate structure (tā dìdì shì gè láodòngmófàn, 他弟弟 是个 劳动模范).

In Indo-European languages, sentences and phrases are in contrast: a sentence must have a finite verb, whereas a phrase must not be headed by a finite verb (定式动词, *dìng shì dòng cí*), except if it is a verb phrase. A sentence must necessarily involve a subject-verb relationship, while a phrase must not involve such a relationship (except in embedded subordinate clauses of course). (Lu Jianming, 2003)

Unlike Indo-European languages, Chinese lacks the distinction between finite and non-finite verbs because of its absence of morphological changes.

In the field of Chinese linguistics, sentence construction assumes a deep complexity, marked by the intricate nesting of a variety of phrase structures. These phrase structures, in turn, nest both single words and more complex ones. This results in a hierarchical relationship among characters, words, phrases, and sentences, forming a progressively layered structure.

Within the realm of Chinese linguistics, an intricate structure of grammatical units is recognized, comprising four fundamental categories: sentences, phrases, words, and morphemes. These units form the hierarchical framework upon which the Chinese language is built, allowing for the expression of meaning and communication. As shown in Figure 1.4 below, at the lowest level of this hierarchy are morphemes, which are the smallest units of meaning in the language. Morphemes can be assembled into words, which represent more concrete and contextually significant units. Further up the hierarchy, words can be combined to create more complex linguistic structures known as phrases. And finally, both words and phrases can become sentences with a speech rhythm.

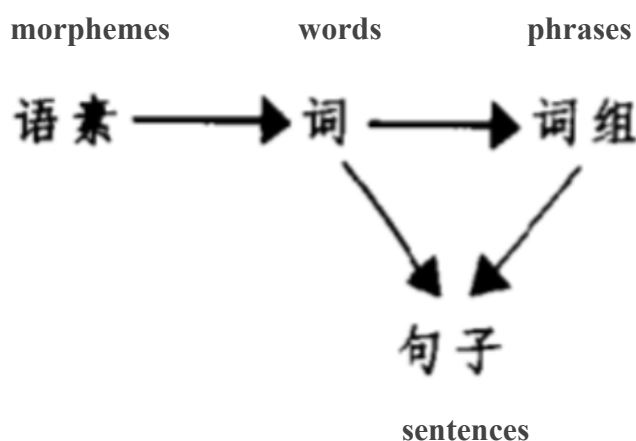


Figure 1.4 - The Granularity of Linguistic Units in Chinese (Lu Jianming, 2003)

In our study, the analysis of sentence structures only focused on two levels of granularity: the indivisible units and divisible syntactic units. Intermediate levels, such as compounds, multi-word expressions (MWE), and phrases, were not taken into consideration. As will be detailedly explained in Chapter 3, we adapt in our annotation guidelines the mark “@m” to represent that a relation is a word-internal relation, distinct from the normal syntactic relations in conventional UD/SUD treebanks. Normal syntactic relations were assigned to words within complex terms. For a more detailed explanation of the labelling process, please refer to Section 3.1.1.

#### 1.2.1.2 The Parallelism Between Compound Word Structure and Syntactic Structure in Chinese

The Modern Chinese Dictionary (《现代汉语词典》(Xiàndài hànyǔ cídiǎn), fifth Edition, 2005) defines “compounds” (复合词, fù hé cí) as:

*“words composed of two or more word elements. Compounds can be divided into two categories: words composed of two or more word roots are called compound words, and those formed by adding affixes to word roots are called derived words.”*

Through the history of Chinese morphological study, “compounds” (复合词, fù hé cí) have been classified into different categories in various ways based on the internal morphological relationships.

Zhao Yuanren in 1948 initially classified compounds into six categories: “subject-verb” (主谓, zhǔ wèi), “parallel” (并列, bìng liè), “main-subordinate” (主从, zhǔ cóng), “verb-object” (动词宾语, dòng cí bīn yǔ), “verb-complement” (动词补足语, dòng cí bǔ zú yǔ), and “lexical compounds” (单性复合词, dān xìng fù hé cí). In 1968, he further simplified this classification into five categories: “subject-verb” (主谓, zhǔ wèi), “parallel” (并列, bìng liè), “main-subordinate” (主从, zhǔ cóng), “verb-object” (动词宾语, dòng cí bīn yǔ) and “verb-complement” (动词补足语, dòng cí bǔ zú yǔ).

Lu Zhiwei in 1957 categorized compounds into nine structural types: “polyphonic root words” (多音的根词, *duō yīn de gēn cí*), “parallel” (并立, *bìng lì*), “repetition” (重叠, *chóng dié*), “centrally modified” (向心修饰, *xiàng xīn xiū shì*), “post-complement” (后补, *hòu bǔ*), “verb-object” (动宾, *dòng bīn*), “subject-verb” (主谓, *zhǔ wèi*), “prepositional element” (前置成分, *qián zhì chéng fèn*), and “postpositional element” (后置成分, *hòu zhì chéng fèn*).

Zhou Zumo in 1959 classified compounds into six types: “attributive type” (偏正式, *piān zhèng shì*), “coordinative type” (联合式, *lián hé shì*), “controlling type” (支配式, *zhī pèi shì*), “complementary type” (补充式, *bǔ chōng shì*), “expressive type” (表述式, *biǎo shù shì*), and “repetitive type” (重叠式, *chóng dié shì*).

Ding Shengshu in 1961 categorized compounds into six types based on their word formation: “parallel type” (并列式, *bìng liè shì*), “attributive type” (偏正式, *piān zhèng shì*), “verb-object type” (动宾式, *dòng bīn shì*), “verb-complement type” (动补式, *dòng bǔ shì*), “subject-verb type” (主谓式, *zhǔ wèi shì*), and “additional type” (附加式, *fù jiā shì*).

Li Jinxi in 1962 established a “Compound Word Category Composition Table” and divided Chinese compounds into three main classes: “integrated” (合体, *hé tǐ*), “parallel” (并行, *bìng xíng*), and “related” (相属, *xiāng zhǔ*). Integrated compounds are further subdivided into four subcategories: “double sound” (双声, *shuāng shēng*), “double rhyme” (叠韵, *dié yùn*), “other”, and “special nouns”. Parallel compounds are categorized into three subtypes: “synonymous”, “opposite”, and “repetition”. Related compounds are divided into eight subcategories: “noun-noun”, “verb-noun”, “adjective-noun”, “verb-verb”, “verb-adverb”, “adjective-adverb”, “adverb-adverb”, and “with affixes”.

Consequently, the current academic consensus generally categorizes compounds into three main classes: compound words, overlapping words and appending words<sup>13</sup>. And compound words into five basic structural patterns: “attributive” (偏正式, *piān zhèng shì*), “coordinative” (联合式, *lián hé shì*), “predicate-object” (述宾式, *shù bīn shì*), “predicate-complement” (述补式, *shù bǔ shì*), and “subject-predicate” (主谓式, *zhǔ wèi shì*). Although the terminology used may vary, these five categories are commonly accepted.

Here we mainly discuss the five basic types of compound words, which largely correspond to the five basic phrase structures presented in Section 1.2.1.1.

The distinction between compounds (复合词, *fù hé cí*) and phrases (短语, *duǎn yǔ*) in the Chinese language has been a long-standing challenge in the fields of lexicology and grammar. Scholars from previous generations have expressed various views on this matter.

Wang Li (1980), for instance, once remarked that “*there is no absolute boundary between words and phrases.*” Similarly, Lü Shuxiang (2001) explicitly stated that “*due to the lack of developed morphology in Chinese, many grammatical phenomena are gradual rather than abrupt, making it easy to encounter various ‘intermediate states’ in grammatical analysis. The boundaries between words and non-words (things smaller or larger than words), and the boundaries between word classes, are difficult to delineate universally. This is an objective fact that cannot be eliminated and should not be concealed.*”

<sup>13</sup> The two last can also be regarded as derived words.

The fundamental reason for this difficulty is often believed to be that “*the construction of Chinese compounds mostly employs syntactic means.*” (Chi et Lin, 2019)

As emphasized by distinguished linguists, including Zhang Shilu (1956), Lu Zhiwei (1957), Lü Shuxiang (1979), Zhu Dexi (1979; 1982), Li Xingjian (1982), Wang Hongjun (1998), Ge Benyi (2001), Dong Xiufang (2011), Shao Jingmin (2016), Lu Jianming (2016), and others, there is a notable parallelism between the formation of compounds, phrases, and syntactic structures in the Chinese language.

According to Dong Xiufang (2011), the parallelism between lexicology and syntax in Chinese has deep historical roots. Givén (1971) introduced a famous viewpoint that today’s lexicology was yesterday’s syntax. Selkirk (1982) argued that lexical structures share the same general formal requirements as syntactic structures, derived from the same regular system. Baker (1985) introduced the Mirror Principle to suggest that the derivation of word structure must reflect the derivation of syntactic structure, and vice versa.

Tang Tingchi, from a generative grammar perspective, discussed the consistency between Chinese lexicology and syntax on multiple occasions in his work (1991; 1992; 1994). He believed that the syntactic structure and word structure in Chinese are highly similar in terms of hierarchical organization and linear order, governed by the same principles and parameters.

The origin of the phenomena of parallelism between syntax and morphology can be found in the bi-syllabilization of Chinese Words. The historical development of Chinese vocabulary confirms this point: Many disyllabic words in Chinese have originated from frozen phrases. The transformation from phrases to disyllabic words<sup>14</sup> is a major way of producing disyllabic words. These disyllabic words often have their roots in phrases and have undergone a process of fusion, turning syntactic structures into lexical structures (Dong Xiufang, 2011).

Dong (2011) introduced three types of lexicalization in Chinese evaluation:

1. from phrases

马	车
mǎ	chē
‘horse’	‘vehicle’
‘carriage’	

2. from grammatical structures

所	得
suǒ	dé
‘marker turning the verb into a more nominal (noun-like) form’	‘gain’
‘income’	

---

<sup>14</sup> The disyllabic words here is equal to the bigrams or the words composed of two characters in Chinese.

3. from cross-layer structures<sup>15</sup>

可	以
kě	yǐ
‘can’	‘by’
‘can’	

These three perspectives are vital to identifying the word-internal structure in Modern Chinese. The theory of bi-syllabilization will be discussed in detail in Section 3.1.

Most bisyllabic terms are lexicalized from phrases. In Dong’s work (2011), following the conventional categorization, Chinese phrases are classified into five basic structural patterns as presented in Section 1.2.1.1: coordination, modifier-head, verb-object, subject-predicate, and verb-complement. Her prolonged examination reveals that all five types of phrase structures can be reduced to compound words. This historical process demonstrates the close relationship between syntactic and lexical levels in the Chinese language.

Expanding beyond phrases, Huang Borong and Liao Xudong (2007) delve into a more detailed analysis of the categorization of words (including both monomorphemic words and compounds), elucidating how these linguistic entities contribute to the broader syntactic structure of Modern Chinese. They propose several examples for each type of compound in Chinese (a detailed analysis of technical terms in patent will be presented in Sections 3.1.1 and 3.3.2):

- Monomorphemic words (单纯词, *dān chún cí*), i.e. words that contain only one morpheme.
  - Consecutive words (连绵词, *lián mián cí*): Two different syllables are concatenated together to express a single meaning, and they cannot be separated; in other words, the two characters form a morphemic word.

(1)	蜘蛛	哆嗦	芙蓉 <sup>16</sup>
	zhī zhū	duō suō	fú róng
	‘spider’	‘shiver’	‘cotton rose’

- Reduplicated characters (叠音词, *dié yīn cí*): Monomorphemic words formed by the reduplication of non-morphemic characters, i.e. characters that can never appear as a morpheme on their own.

(2)	猩猩	姥姥
	xīng xīng	lǎo lǎo
	‘gorilla’	‘grandma (mother’s mother)’

<sup>15</sup> The term “跨层结构” can be translated to “cross-layer structure” in English. It refers to a structure composed of two elements that do not form a direct constituent pair but belong to different syntactic layers and are adjacent in a linear sequence (Hao Jingcun & Liang Boshu, 1992). Some disyllabic words, primarily function words, are derived from cross-layer structures.

<sup>16</sup> These and the following examples are taken from Huang Borong and Liao Xudong (2007).

○ Transliterated loanwords

(3) 巧 克 力  
qiǎo kè lì  
'chocolate'

● Compounds (合成词, hé chéng cí): The words that contains more than one morpheme.

○ Compound words (复合式, fù hé shì): Words composed of two different root<sup>17</sup> morphemes. The five basic types of compound words correspond to the five basic phrase structures.

■ Coordinative type (联合型, lián hé xíng; 并列式, bìng liè shì): Composed of two morphemes with similar, related, or opposite meanings.

A. The meaning of the whole term is the combination of its two component characters, which often have similar meanings.

(4) 价 值  
jià zhí  
'price\_N' 'value\_N'  
'price; value\_N/V'

(5) 美 好  
měi hǎo  
'beautiful\_ADJ' 'good\_ADJ'  
'wonderful\_ADJ'

(6) 收 获  
shōu huò  
'collect\_V' 'gain\_V'  
'harvest\_V'

B. The meaning of the whole term is not equal to either of its component characters.

(7) 骨 肉  
gǔ ròu  
'bone\_N' 'flesh\_N'  
'Blood relation\_N'

(8) 买 卖  
mǎi mǎi  
'buying\_V' 'selling\_V'  
'Business\_N'

(9) 开 关  
kāi guān

---

<sup>17</sup> Meaningful morphemes that can be both inside compound words and as standalone words, with a variable position within compound words.

‘open\_V’      ‘close\_V’  
‘Switch\_N’

C. The meaning of the whole term inclines to one of its component characters, which often have contradictory meanings.

(10) 国                  家  
guó                  jiā  
‘country\_N’      ‘home\_N’  
‘Country, Nation’

(11) 质                  量  
zhì                  liàng  
‘quality\_N’      ‘quantity\_N’  
‘Quality’

(12) 忘                  记  
wàng                jì  
‘forget\_V’        ‘remember\_V’  
‘Forget\_V’

■ Attributive type (偏正型, piān zhèng xíng): The preceding morpheme restrictively modifies the following morpheme.

A. The head is a NOUN.

(13) 中                  国  
zhōng                guó  
‘central\_ADJ’    ‘country\_N’  
‘China\_N’

(14) 冰                  箱  
bīng                  xiāng  
‘ice\_N’            ‘box\_N’  
‘Refrigerator\_N’

(15) 飞                  机  
fēi                  jī  
‘fly\_V’            ‘Machine\_N’  
‘Airplane\_N’

B. The head is an ADJ.

(16) 深                  蓝  
shēn                 lán  
‘deep\_ADJ’      ‘blue\_ADJ’  
‘Deep blue\_ADJ/N’

- (17) 滚                      烫  
gǔn                      tàng  
'boil\_V' 'hot\_ADJ'  
'Boiling hot\_ADJ'
- (18) 透                      明  
tòu                      míng  
'transparent\_ADJ'      'clear\_ADJ'  
'Transparent\_ADJ'

C. The head is VERB.

- (19) 热                      爱  
rè                      ài  
'hot\_ADJ'              'love\_V'  
'Love, Passionate about\_V'
- (20) 迟                      到  
chí                      dào  
'late\_ADJ'              'arrive\_V'  
'Be late\_V'
- (21) 稍                      息  
shāo                      xī  
'slightly\_ADV'              'break\_V'  
'Stand at ease\_V'

There are also some terms where the modifier of the verbal or adjective character is a nominal character. The modification can be divided into four different cases: analogy, means, location and time.

a. Analogy: V like N.

- (22) 火                      红  
huǒ                      hóng  
'fire\_N' 'red\_ADJ'  
'Red like fire\_ADJ'
- (23) 鼠                      窜  
shǔ                      cuàn  
'rat\_N'                  'flee\_V'  
'Scamper off like a rat\_V'
- (24) 粉                      碎  
fěn                      suì  
'powder\_N'              'broken\_ADJ'  
'Smash, shatter\_ADJ'

b. By means of: Using N to V.

- (25) 笔                      谈  
bǐ                      tán  
'pen\_N' 'talk\_V'  
'Written conversation\_V/N'



- (26) 法                      治  
fǎ                              zhì  
'law\_N' 'rule\_V'  
'Rule of law\_V'
- (27) 目                      送  
mù                              sòng  
'eye\_N' 'see off\_V'  
'Watching someone or something leave\_V'

c. Location: At/To/From N to V.

- (28) 左                      倾  
zuǒ                              qīng  
'left\_N' 'incline\_V'  
'Left leaning\_V'
- (29) 上                      传  
shàng                              chuán  
'up\_ADV' 'transport\_V'  
'Upload\_V'
- (30) 空                      袭  
kōng                              xí  
'sky\_N' 'attack\_V'  
'Air raid\_V'

d. Time: At the time of N, V.

- (31) 夜                      游  
yè                              yóu  
'night\_N' 'tour\_V'  
'Noctivagation\_N/to be noctivagant\_V'
- (32) 春                      耕  
chūn                              gēng  
'spring\_N' 'plowing\_V'  
'Spring plowing\_N/to plow in spring\_V'
- (33) 午                      休  
wǔ                              xiū  
'noon\_N' 'nap\_V'  
'Afternoon nap\_N/have an afternoon nap\_V'

- Predicate-complement type (补充型, bǔ chōng xíng): The subsequent morpheme provides supplementary information to the preceding morpheme.

A. VERB + Complements (resultitive and directional)

- (34) 提                      高  
tí                              gāo  
'hold\_V' 'high\_ADJ'  
'raise\_V'

- (35) 延                      长  
yán                      cháng  
‘extend\_V’              ‘long\_ADJ’  
‘extend\_V’

B. NOUN + Classifier/ Units of Things (事物单位)<sup>18</sup>

- (36) 人                      口  
rén                      kǒu  
‘person\_N’              classifier<sup>19</sup>  
‘population\_N’

- (37) 稿                      件  
gǎo                      jiàn  
‘manuscript\_N’        classifier  
‘manuscript\_N’

- Predicate-object type (动宾型, dòng bīn xíng; 支配式, zhī pèi shì): The first morpheme represents an action or behavior, while the second morpheme represents the entity or object associated with that action or behavior. These cases could be annotated as collocations (Mel’cuk 1998), and thus as separate words. We prefer to annotate them as a morphological phenomenon because they cannot be separated by other words without losing their meaning and thus modifiers are placed outside of these words. (see Section 3.1)

- (38) 司                      令  
sī                      lìng  
‘take charge of\_V’      ‘command\_N’  
‘commander\_N’

- (39) 注                      意  
zhù                      yì  
‘pour\_V’                      ‘conscience\_N’  
‘pay attention\_V’

- (40) 有                      限  
yǒu                      xiàn  
‘have\_V’                      ‘limit\_N’  
‘limited\_ADJ’

- (41) 失                      业  
shī                      yè  
‘lose\_V’                      ‘work\_N’  
‘unemployed\_ADJ; unemploy\_V’

<sup>18</sup> Should be regarded as being of attributive type according to Zhou Jian (2016).

<sup>19</sup> Classifiers, also known as measure words in Chinese, are an integral part of the language, used in conjunction with numerals to specify the quantity of nouns. Unlike English, where "a piece of" or "a cup of" can often be omitted, in Chinese, classifiers are obligatory when quantifying nouns. The principle behind the use of classifiers is closely tied to the noun being referred to, with the choice of classifier depending on the characteristics of that noun, such as its shape, category, or other physical attributes.

- Subject-predicate type (主谓型, zhǔ wèi xíng): The latter morpheme predicates the thing or object described by the former morpheme.

(42)	地 dì 'earth_N' 'earthquake_NV'	震 zhèn 'tremble_V'
(43)	月 yuè 'moon_N' 'moon_N'	亮 liàng 'bright_ADJ'
(44)	年 nián 'age_N' 'young_ADJ'	轻 qīng 'light_ADJ'
(45)	自 zì 'self_PRON' 'automatic_ADJ'	动 dòng 'move_V'

- Overlapping type (重叠式): Words composed of repeated identical root morphemes. We can use one of the repeated morphemes instead of its overlapping form, which makes them distinct from the reduplicated monosyllabic words.

(46)	哥 哥 gē gē 'brother_N'	哥 gē 'brother_N'
(47)	刚 刚 gāng gāng 'just_ADV'	刚 gāng 'just_ADV'

- Appending type (附加式): Words composed of root morphemes and affixes<sup>20</sup>.

- Prefixing type (前加式, qián jiā shì): **Prefix** + Root

(48)	老 lǎo old_ADJ 'Tiger_N'	虎 hǔ tiger_N	老 乡 lǎo xiāng old hometown 'sb from the same hometown_N'
(49)	阿 姨 ā yí 'Aunt_N'	阿 伯 ā bó 'Uncle_N'	
(50)	第 五 dì wǔ 'Fifth'	第 一 dì yī 'First'	

<sup>20</sup> Affixes are morphemes with a fixed position and an abstract meaning within compound words.

■ Suffixing type (后加式, hòu jiā shì): Root + **Suffix**

(51)	刀	子	瓶	子	
	dāo	zi	píng	zi	
	'Knife_N'		'Bottle_N'		
(52)	木	头	苦	头	
	mù	tóu	kǔ	tóu	
	'Wood_N'		'Sufferings_N'		
(53)	硬	性	创造	性	
	yìng	xìng	chuàng zào	xìng	
	'Rigid_ADJ'		'Creativity_N'		
(54)	绿	化	自动	化	
	lǜ	huà	zì dòng	huà	
	'Greening_V'		'Automation_V'		
(55)	在	于	勇	于	用
	zài	yú	yǒng	yú	yòng
	'Consist in_V'		'Be brave enough to_V'		'Use for_V'

Furthermore, apart from the five basic structures of compound words introduced by Huang et Liao (2007), in Zhou Jian's "On Vocabulary" (2016), two additional compound word structures are introduced: the "successive type" and the "associative type".

- Successive Type (递续式, dì xù shì): Words in the successive type are all verbs. The two direct constituents that make up such words both govern or relate to some object or action. Unlike the attributive type, neither of the two direct constituents specifies the object of governance. The successive type also differs from the predicate-complement type, as neither of its direct constituents provides supplementary information about the governing action.

(56)	听	信	查	封
	tīng	xìn	chá	fēng
	'hear_V'	'believe_V'	'examine_V'	'seal_V'
	'Hear and believe_V'		'Sequestration_N/V'	

It might be tempting to analyze these constructions as serial verbs. However, their meaning is not as transparent as serial verb constructions, and we will analyze them as serial verbs but with the morphological relation marker "@m".

- Idiomatic Type (意合式, yì hé shì)<sup>21</sup>: In compound words of the associative type, the two direct constituents are not combined in the usual sense of association as commonly understood. The internal relations of these words are not transparent for the speaker of contemporary Mandarin and can only be explained etymologically.

<sup>21</sup> Some ideographic compounds, when studied by experts, can shed light on their historical development and evolution. For example, the word "感冒 (gǎnmào)", meaning "cold" or "common cold", is one such term. An article titled "The Origin of the Term '感冒 (Common Cold)' from Qing Dynasty Official Terminology" was published on a Sina blog on October 14, 2008. This article provides a detailed examination of the reasons and historical context for the transition of "感冒" from a term related to illness to a colloquial expression with a different connotation. (Zhou Jian, 2016)

(57)	收 shōu 'put away_N' 'Wind up_V'	尾 wěi 'tail_N'	熨 yùn 'iron_V' 'Iron_N'	斗 dòu 'measuring instrument for food_N'
------	--	----------------------	----------------------------------	---

The parallelism between compound words, phrases, and syntactic structures in Modern Chinese unveils the systematic nature of the language's construction. This linguistic phenomenon not only highlights the intricacies of the language but also provides linguists with a comprehensive framework for understanding the grammatical rules that govern its use.

Compound words	Phrases	Relation
Coordinative type (联合型, lián hé xíng; 并列式, bìng liè shì)	Coordinative phrases (联合短语, lián hé duǎn yǔ)	Coordinative heads
Attributive type (偏正型, piān zhèng xíng)	Attributive phrases (偏正短语, piān zhèng duǎn yǔ)	Modifier-Head
Predicate-complement type (补充型, bǔ chōng xíng)	Predicative phrases (中补短语, zhōng bǔ duǎn yǔ)	Head-Complement
Verb-object type (动宾型, dòng bīn xíng; 支配式, zhī pèi shì)	Verb-object phrases (动宾短语, dòng bīn duǎn yǔ)	Head-Object
Subject-predicate type (主谓型, zhǔ wèi xíng)	Subject-predicate phrases (主谓短语, zhǔ wèi duǎn yǔ)	Subject-Head

**Table 1.1 - The five fundamental structures of compounds and phrases that share the same relationship in Chinese**

In Table 1.1 above, we have summarized and compared the five fundamental structures of compounds and phrases that share the same relationship in Chinese. In Section 3.1, we will introduce their correspondence with the SUD (Surface-Syntactic Universal Dependencies) annotation schema, providing examples from our patent claim treebank to illustrate these relationships.

Some phrase types are more conducive to word formation than others, with “attributive” phrases accounting for over half of the proportion, while “predicate-complement” and “subject-predicate” phrases making up a very small proportion.

From the statistics of disyllabic words with different structural patterns in modern Chinese, it's evident that the proportion of disyllabic words constructed from the five main structural patterns varies significantly. According to Zhou Jian's statistics on the 1983 edition of the “Modern Chinese Dictionary” (《现代汉语词典》), the proportions are as follows:

- Attributive (偏正式): 50.72%

## Chapter 1 - Linguistic and technical background of the research

- Coordinative (并列式): 25.7%
- Predicate-Object (动宾式): 15.6%
- Subject-Predicate (主谓式): 1.17%
- Predicate-Complement (述补式): 0.93%

Similarly, Bian Chenglin's statistics on the 1990 edition of the "Modern Chinese Dictionary" (《现代汉语词典》) yielded these proportions:

- Attributive (偏正式): 52.75%
- Predicate-Object (动宾式): 20.18%
- Coordinative (并列式): 19.31%
- Predicate-Complement (述补式): 2.62%
- Subject-Predicate (主谓式): 1.39%

These statistics demonstrate a significant imbalance in the numbers of different types of disyllabic words. This imbalance to some extent reflects the varying degrees of difficulty in lexicalizing different types of phrases; some phrase types are more challenging to convert into words than others.

### 1.2.1.3 Wordhood in Chinese

Bloomfield (1933) pointed out that the differences in interrogative forms among different languages are greater than syntactic differences. The grammatical characteristics of compound words also vary depending on the language, so the boundary between compound words and phrases is not an easily definable issue.

One significant reason for the difficulty in delineating this boundary is that most disyllabic words have evolved from phrases. Some disyllabic words may still bear the imprint of phrases, whether deep or shallow, and they are still in the process of transitioning from phrases to words, not yet fully solidified. Therefore, it is challenging to separate them entirely from phrases. Wang Hongjun (1994) proposed that in Chinese, one should start with the combination of characters within a word group, first identify the rules of free combinations between characters, and then use a process of elimination to determine words. According to Dong (2011), in cases where it is more difficult to directly define words, this is a more feasible approach.

Despite differing opinions among scholars regarding whether compound words and phrases are parallel, efforts to identify the differences between the two and their manifestations have continued unabated.

From a linguistic perspective, the Lexical Integrity Hypothesis (LIH) is a famous theory in word-phrase distinction (Jackendoff, 1972; Selkirk, 1984; Huang, 1984), which states that no phrase-level rule should be applicable to a word whose internal structures are no longer accessible<sup>22</sup>.

A methodology first introduced by Kratochvíl and developed by Duanmu and N'Guyen based more on syntactic definitions and proposed a set of widely applied linguistic criteria to decide what is a

---

<sup>22</sup> However, as a general rule, the LIH can be challenged by resultative (V-R) and verb-object (V-O) structures (Huang, 1984), such as in example (5) below.

word, including (1) Conjunction Reduction, (2) Freedom of Parts, (3) Semantic Composition, (4) Exocentric Structure, (5) Adverbial Modification, (6) XP Substitution, (7) Productivity Criterion, (8) Syllable Count and (9) Insertion. (Kratochvíl, 1966; Huang, 1984; Dai, 1992; Duanmu, 1998; Packard, 2000; Nguyen, 2006; Magistry, 2013)

In China, Liu Shuxin (1990) summarized the differences between these two as five aspects: (1) whether there is an internal phonetic pause, (2) whether the meaning is simple, (3) whether the component relationship is additive or combinatorial, (4) whether the direct components can be independent, and (5) the structural stability. Hu Yushu (1995) summarized it into three aspects: (1) semantic fusion, (2) fixed sound form, and (3) whether it is the smallest independent unit in terms of grammar. Huang Yue Yuan (1996) summarized a list of criteria to distinguish compound words and phrases, from (1) phonetic, (2) semantic and (3) grammatical perspectives separately. Cao Wei (2004) summarized it into four points: (1) whether the structure is complete and standardized, (2) whether there is a sense of entirety in meaning, (3) whether it is expandable, and (4) syllable length limitations. Cheng Xiangqing (2008) believes that there can also be four criteria in Classical Chinese: (1) structure, (2) word meaning, (3) frequency, and (4) rhetoric. Shao Jingmin (2016) summarizes it into three points: (1) words have semantic integrity, (2) there are no internal pauses in terms of pronunciation, and (3) there is no grammatical expansion.

Feng (2004a) proposed five criteria for distinguishing linguistic units in Chinese, which are (1) has meaning; (2) is the smallest unit; (3) can be used independently; (4) the number of morphemes included in the unit; and (5) the number of words included in the unit. He also summarized several non-grammatical factors of word segmentation, including four semantic approaches, two prosodic approaches, and three non-linguistic principles. Among these, the domain-oriented principle is particularly important in Chinese Word Segmentation (CWS), which is a crucial step in conventional Natural Language Processing (NLP) tasks for Chinese.

In “Modern Chinese” (Huang et Liao, 2007), “phrases” are defined as linguistic units that are composed of words in a hierarchical manner, layer by layer, while “words” are “the smallest units in language that can function independently with both sound and meaning”. The “expansion method” or “insertion method” are proposed to distinguish between words and phrases, in which we try to insert a character into a word to see if there is a change in the meaning. This is an important criterion mentioned in a series of morphology works on Modern Chinese.

As in the following four examples,

- |     |                         |                                |         |
|-----|-------------------------|--------------------------------|---------|
| (1) | 冰                       | 箱                              |         |
|     | bīng                    | xiāng                          |         |
|     | ‘ice’                   | ‘box’                          |         |
|     | ‘refrigerator’          |                                |         |
|     | 冰                       | 的                              | 箱子      |
|     | bīng                    | de                             | xiāngzi |
|     | ‘ice’                   | ‘possessive particle’          | ‘box’   |
|     | ‘a box of ice’          |                                |         |
| (2) | 改                       | 进                              |         |
|     | gǎi                     | jìn                            |         |
|     | ‘to change; to improve’ | ‘to advance; to make progress’ |         |
|     | ‘improve; improvement’  |                                |         |

## Chapter 1 - Linguistic and technical background of the research

改 gǎi 'to change; to improve' 'capable of being improved'	得 dé 'can; may'	进 jìn 'to advance; to make progress'	
(3) 子 zǐ 'son(s)' 'children'	女 nǚ 'daughter(s)'		
子 zǐ 'son(s)' 'sons and daughters'	和 hé 'and'	女 nǚ 'daughter(s)'	
(4) 买 mǎi 'to buy'	卖 mài 'to sell'		
又 yòu 'both'	买 mǎi 'to buy'	又 yòu 'both'	卖 mài 'to sell'
'both buying and selling'			
(5) 洗 xǐ 'to wash'	澡 zǎo 'bath'		
'take a bath'			
洗 xǐ 'to wash'	了 <sup>23</sup> le 'aspect marker'	个 gè 'a'	澡 zǎo 'bath'
'took a bath'			

The nominal word ‘冰箱 (bīng xiāng, refrigerator)’ is not identical to ‘冰的箱子 (bīng de xiāngzi, a box of ice)’ with the insertion of the character “的 (DE, possessive particle)” in example (1). And it also comes true for the other three examples with the insertion of “得 (dé, can; may)”, “和 (hé, and)”, “又...又... (yòu ... yòu ..., both)”. One exception is the separable verbs (离合词, lí hé cí), which is a specific type of verb in Chinese that allows insertion, such as in example (5). However, the authors do not give an explanation of how to distinguish separable verbs in this work.

The theory of word formation in Modern Chinese pertains to the process of bi-syllabization of Chinese words. In his chronological research on bi-syllabication in the Chinese lexicon, Dong (2011) offers an evaluation of various structure types of words, along with a discussion on how they differ from phrases with similar structures for each type. These tests include permutation, category

<sup>23</sup> When “了” appears immediately after a verb, like in this example, it indicates the completion of an action, essentially functioning as a perfective aspect marker.



conversion, accent position, insertion, and the non-specificity of nouns, among others. A detailed presentation of Dong’s study will be provided in Section 3.1.2, accompanied by examples of the six principal word internal structures in Chinese morphology.

Dong (2016) believes that “*Chinese compound words and syntactic structures exhibit obvious isomorphism*”, but she does not support the integration of lexicology into syntax because “*the generality and universality of lexicon cannot compare with syntax; lexicon only has weaker generality and regularity, while syntax has strong generality and regularity.*”

According to Zhou Jian’s statistics (2014) on the “Modern Chinese Dictionary” (《现代汉语词典》), complex words can be divided into two categories: syntactic words and morphological words. Syntactic words make up approximately 96.57% of the total, while morphological words account for only about 3.43%. This data suggests that syntactic words seem to have an overwhelming advantage in the dictionary.

However, Zhou did not provide a specific explanation for the term “morphological words”. Based on the context, “morphological words” likely refer to words that are “*difficult or impossible to explain using syntactic structural patterns*”. These words may have been extracted from ancient texts, purely based on phonetic considerations, difficult to interpret literally, or constructed using function words with unclear meanings. They do not fit into the categories of phrase structures such as coordination, modification, complementation, assertion, combination, overlap, or continuation, as mentioned by the author. Examples of such “morphological words” could include terms like “弱冠” (ruò guān, weak crown, ‘reaching adulthood’), “皮傅” (pí fù, skin teacher, ‘making associations based on a superficial and partial understanding’), “驴骡” (lǘ luó, donkey mule, ‘hinny (the offspring of a male horse (a stallion) and a female donkey (a jenny))’), “马骡” (mǎ luó, horse mule, ‘the inverse of the hinny’), “雷同” (léi tóng, thunder same ‘identical’), “天牛” (tiān niú, sky ox, ‘longhorn beetle’), and so on.

These tests offer a means to examine wordhood from a syntactic perspective, considering the relation between morphemes, or characters, as specific links (“morph” or “in-word”) in the syntactic analysis and an integral part of the syntactic parsing process. Notably, recent neural parsers on the character level are also built from this point of view (Zhongguo Li, 2011; Li et al., 2018; Yan et al., 2019; Hang Yan et al., 2020).

Apart from linguistic studies, one of the most widely applied standards for Chinese word segmentation is the Segmentation Guidelines for the Penn Chinese Treebank (3.0) (Xia, 2000a). These guidelines are based on wordhood tests summarized by Dai (1992), which include the following criteria while rejecting the productivity test and the frequency test<sup>24</sup>:

- Bound morpheme: a bound morpheme should be attached to its neighbouring morpheme to form a word when possible.
- Productivity: if a rule that combines the expression X-Y does not apply generally (i.e., it is not productive), then X-Y is likely to be a word.
- Frequency of co-occurrence: if the expression X-Y occurs very often, it is likely to be a word.
- Complex internal structure: strings with complex internal structures should be segmented when possible.
- Compositionality: if the meaning of X-Y is not compositional, it is likely to be a word.

<sup>24</sup> Assuming the string that we are trying to segment is X-Y, where X and Y are two morphemes.

## Chapter 1 - Linguistic and technical background of the research

- Insertion: if another morpheme can be inserted between X and Y, then X-Y is unlikely to be a word.
- XP-substitution: if a morpheme cannot be replaced by a phrase of the same type, then it is likely to be part of a word.

The notion of “word” proposed in these guidelines is based on a minimal syntactic unit, and this segmentation standard has been adopted in later developments such as the UD Chinese HK treebank (Leung et al., 2016).

In practice, as highlighted by Duanmu (1998), determining phrasal rules can be challenging, and various test criteria may yield conflicting outcomes. This discrepancy between theoretical standards and practical requirements has led researchers to tailor the concept of “words” to suit the specific needs of Natural Language Processing. As asserted by Sproat and Shih (2002), this potential mismatch may not pose a serious problem in computational linguistics, where the definition of words can be flexible and contingent upon the specific usage and processing of words in computer applications.

Chinese is a unique language in many ways, including its writing system, grammar, and syntax. One of the most significant challenges in Chinese Natural Language Processing (NLP) is word segmentation, the process of identifying word boundaries in written text. The issue of word segmentation has long been a complicated problem for the Chinese language, resembling a chicken-and-egg situation. The concept of words in Latin languages, separated by natural boundary marks like spaces, has never gained wide acceptance in Chinese. Due to the absence of natural delimiters or inflectional marks, the two primary indicators of wordhood (Magistry et al., 2012), the distinction between characters and larger word-like units in Chinese has been an unfamiliar and confusing notion since it was introduced in 1907 by Zhang Shizhao. This results in a low rate of agreement on wordhood, only 76%, among Chinese native speakers (Sproat et al., 1997).

As a practical attempt, the “Contemporary Chinese Language Word Segmentation Specification for Information Processing” (Liu et al., 1994) aims to establish the standard for Chinese word segmentation. Released in 1994, it is the first and only official guideline for Chinese word segmentation. The specification defines the segmentation unit as “the basic unit used for Chinese information processing with a specific semantic or grammatical function” and includes a referent vocabulary list with 40,000 terms. It categorizes all Chinese terms into 13 types and provides detailed segmentation rules for each type. However, this work heavily relies on lexical categories and bases its standards on examples instead of applicable tests.

	size	words		characters	
		tokens	types	tokens	types
Academia Sinica (AS)	516 KB	5 449 698	141 340	8 368 050	6 117
City University of Hong Kong (CITYU)	154 KB	1 455 629	69 085	2 403 355	4 923

Peking University (PKU)	177 KB	1 109 947	55 303	1 826 448	4 698
Microsoft Research (MSR)	41 KB	2 368 391	88 119	4 050 469	5 167

**Table - 1.2 The Size of Each Bakeoff Corpus**

Besides the Contemporary Chinese Language Word Segmentation Specification for Information Processing, another valuable resource is the conference “International Chinese Word Segmentation Bakeoffs”, held five times from 2003 to 2010. These “bakeoffs” provide four annotated evaluation corpora<sup>25</sup>, each following different segmentation standards. Below is the table containing the size of each Bakeoff corpora (Table 1.2). Similar to the “Contemporary Chinese Language Word Segmentation Specification for Information Processing”, the standards offered by the competition are based on four independently manually annotated corpora, without any official guidelines that provide general tests.

From a statistical and unsupervised perspective, Magistry (2013) provides a comprehensive overview of various segmentation guidelines in Chinese to explore the notion of wordhood. Magistry defines a word as “minimal autonomous sequences of characters to which we can attribute at least one wordclass” and aims to develop an entropy-based word segmenter.

In computational linguistics, other researchers (Wu, 2003; Gao et al., 2005) define Chinese words from a different viewpoint. They propose a categorization of Chinese words into five types: lexicon words, morphologically derived words, factoids, named entities, and new words. These types of words have distinct computational properties and are processed differently in their system. Among these types, morphologically derived words are often the most ambiguous and are further subdivided into three ambiguity types: (1) Reduplication and Merging/Splitting, (2) Affixation, and (3) Directional and Resultative Compounds.

The idea of customizable segmentation presented by these researchers serves as a practical-oriented definition and is in line with the statement made by Sproat and Shih (2002). Building on this work, Yixuan Li and Kim Gerdes (2019) analyze the Chinese word segmentation (CWS) on patent claims at a multi-level granularity by adopting the classification of segmentation ambiguity. They segment the Chinese patent claim corpus based on six factors: (1) Type of compounding; (2) Word length; (3) Frequency; (4) Mutual information; (5) Resultative verbs; and (6) Insertion of “得 (dé, ‘to be able to’) 不 (bù, ‘not’)” into verbal terms.

Despite this challenge, all existing treebanks and syntactic annotation schemes for Mandarin Chinese adopt word segmentation as the first annotation step. However, their segmentation criteria differ significantly without a clear unified standard. As a result, dependency parsers trained on these treebanks yield inconsistent results, particularly on corpora containing a large number of domain-specific new terms like patent texts (Li et al., 2019).

<sup>25</sup> <http://sighan.cs.uchicago.edu/bakeoff2005/>

In this context, we have decided to explore the idea of developing a character-based annotation schema for Chinese. The treebank annotated with extra inter-characters relations can serve as resources to train a joint segmenter-parser that combines two steps into one. Moreover, we (Li et al., 2019) have also shown that character-level annotation, even a rough one, helps to improve the dependency parsing result on Chinese text of different genres. The annotation of such a treebank is presented in detail in Chapter 3.

### 1.2.1.4 Dependency Syntax Analysis and Existing Chinese Treebanks

Dependency syntactic analysis constitutes a significant branch of syntactic analysis, characterized by its focus on the relationships between words in a sentence, expressed through directed links known as dependencies. In the realm of Chinese linguistic studies, dependency syntax has not been extensively explored, but led to the development of several Chinese treebanks. These treebanks serve as annotated resources that play a crucial role in the training and evaluation of Natural Language Processing (NLP) systems, including parsers and machine translators. Dependency syntax has achieved mainstream status in the field of NLP, serving as a foundational component for various language-related tasks.

The foundational principles of Dependency syntax, as we comprehend it today, were pioneered by the French linguist L. Tesnière in his seminal work “*Eléments de syntaxe structurale*” (1959). This theory has significantly shaped the landscape of linguistic research and has particularly influenced computational linguistics, although dependency parsing did not become mainstream before the year 2000.

In the context of China, the exploration of dependency grammar commenced in the 1980s. Key proponents of this theory include Feng Zhiwei (2010) and Liu Haitao (2018), who have notably contributed to the domain of Machine Translation. Owing to the growing demand for Natural Language Processing and Artificial Intelligence, both nationally and globally, dependency syntax has garnered increasing attention in recent years and has established itself as a dominant paradigm within the realm of NLP.

Syntactic analysis, being the foundational pillar of Natural Language Processing, has remained a central focus of linguistic inquiry. Despite the distinctive morphological traits of the Chinese language, the methodologies applied in Chinese NLP bear resemblance to those used in Western languages like English and French. This similarity, however, results in comparatively lower performance scores for Chinese across a range of syntactic analysis tasks, in comparison to other well-resourced languages. This discrepancy is exemplified by the CoNLL shared tasks conducted between 2007 and 2017, as noted by Zeman et al. (2017).

Concurrently, an empirical study (Qiu & al. 2015) highlights the intricacies of the patent domain. Comparing medical, oral, Weibo, and patent texts, the patent domain exhibits the lowest parsing accuracy. This observation underscores the linguistic complexities inherent in the patent genre. This challenges syntactic analysis not only inter-linguistically but also intra-linguistically within the Chinese language itself. The unique structural traits of patent texts, coupled with the inherent characteristics of the Chinese language, necessitate a comprehensive and meticulous examination.

In light of these observations, it becomes evident that dependency syntactic analysis is an essential endeavour, particularly when applied to intricate and domain-specific textual contexts such as patents.

Its utilization holds the promise of enhancing our understanding of the intricate syntactic structures in Chinese patent documents, contributing to more effective Natural Language Processing and enabling improved automated linguistic analysis within this complex domain.

Several scholarly works have delved into the adaptation of syntactic dependency parsing techniques for English patent documents. These endeavours have unveiled a multitude of challenges posed by patent language phenomena, which significantly complicate the parsing process. Notably, some of these complications encompass (Burga & al. 2013):

- Long Sentences: English patent texts are often characterized by lengthy and convoluted sentence structures. These extensive sentences pose a challenge to syntactic dependency parsers, which may struggle to accurately identify the relationships between various components.
- Complex Syntactic Structures: The intricate nature of patent documents frequently gives rise to intricate syntactic structures. These complexities can confound parsing algorithms, potentially leading to errors in capturing the intended linguistic relationships.
- Peculiar Multi-Word Expressions and Terminology: Patent language is replete with domain-specific multi-word expressions and technical terminology. These linguistic entities are often unconventional and resist straightforward parsing due to their unique compositional and semantic characteristics.
- Different PoS Distribution of Characteristic Tokens: Patent language exhibits distinct part-of-speech (PoS) distribution patterns compared to general language corpora. This variance in PoS distribution can undermine the effectiveness of parsing algorithms that rely on typical linguistic patterns.

These identified challenges correspond closely to the issues revealed in our preliminary experiments on syntactic patent analysis. The results of our investigations align with prior research, underscoring the inherent difficulties posed by the aforementioned linguistic phenomena in the patent context.

Furthermore, in the context of the Chinese language, the process of word segmentation presents a significant hurdle. This issue becomes particularly pronounced when dealing with unknown or domain-specific terms. Presently, there exists no Chinese segmenter specifically tailored to patent texts, exacerbating the difficulty. The accuracy and performance of existing linguistic tools can be severely compromised when confronted with terminology unfamiliar to tools trained on more general linguistic corpora.

In light of these challenges, it becomes apparent that the syntactic analysis of patent documents, both in English and Chinese, requires specialized techniques that can accommodate the intricate language patterns and unique terminological landscape. Addressing these challenges holds the potential to enhance the accuracy and efficacy of syntactic analysis, enabling more robust and insightful linguistic insights within the patent domain.

In the realm of existing syntactic treebanks for Chinese, since the 1990s, there has been significant development with the creation of numerous treebanks (as listed below) that vary in terms of maturity, influence, and scale. These treebanks primarily focus on phrase structure trees, with the Penn Chinese Treebank standing out as a prominent representative.

- The Sinica Chinese Sentence Treebank, developed by (Chen et al., 2003), is a valuable resource for the study of Chinese syntax and language structure. Its development began in 1986 under the supervision of the Academia Sinica Lexicon project (CKIP) and was derived from the Academia Sinica Balanced Corpus (Sinica Corpus), a modern Chinese balanced corpus.

This treebank follows the principles of Information-based Case Grammar (ICG) as its foundational framework for structural representation. The process involves the automatic parsing of sentences using computer algorithms to generate structure trees. Subsequently, human experts manually review, correct, and verify the results to ensure accuracy.

The Chinese Sentence Treebank has evolved and currently stands at version 3.0, comprising six files, 61,087 Chinese tree diagrams, and a total of 361,834 words.

- The Chinese Treebank (CTB) project, which focuses on constituent syntactic annotation, has a rich history. It was initiated in 1998 at the Institute for Research in Cognitive Science (IRCS) of the University of Pennsylvania, later continued at the University of Colorado, and eventually relocated to Brandeis University.

The primary objective of the Chinese Treebank project is to compile a substantial corpus of Chinese text that is not only part-of-speech tagged but also fully bracketed for syntactic analysis. The initial release, Chinese Treebank 1.0, encompassed 100,000 words with syntactic annotations sourced from the Xinhua News Agency newswire.

Chinese Treebank 8.0, released in 2013, represents a significant expansion, including newly annotated data extracted from various sources like newswire articles, magazine texts, and government documents. This release comprises 3,007 text files, encompassing 71,369 sentences, 1,620,561 words, and 2,589,848 characters, which may include Chinese characters (hanzi) and foreign characters.

- The Peking University Treebank, initiated in 1993, utilizes the Head-Driven Phrase Structure Grammar (HPSG) approach. It encompasses 55,611 sentences, 882,326 words, and 1,281,169 characters. This extensive corpus covers a wide range of language usage domains, including literature, academia, news, and applied texts.
- The Tsinghua Chinese Treebank (TCT) is another crucial resource. It extracts a corpus of one million Chinese characters from a balanced corpus of two million characters, spanning literature, academia, news, and applications. These texts undergo automatic sentence

segmentation and syntactic analysis, followed by manual proofreading, resulting in a corpus with complete syntactic structure trees.

- Stanford Dependencies (SD) for Chinese is a linguistic framework that was initiated in 2005 by Huihsen Tseng and Pi-Chuan Chang. This framework is specifically tailored for the analysis of Mandarin Chinese sentences. It is part of the broader Stanford Dependencies project, which originated with the development of a linguistically sound, surface-syntax-oriented dependency representation for English.

In the realm of Mandarin Chinese, Stanford Dependencies for Chinese, often referred to as Stanford Chinese, has gained widespread recognition and use. It adopts its part-of-speech (POS) tagset directly from the Chinese Treebank (CTB), which is currently maintained at Brandeis University and was formerly known as the Penn Chinese Treebank (Penn Chinese).

It is worth noting that the success and acceptance of Stanford Dependencies in Chinese led to collaborative efforts to propose Universal Dependencies in 2013, aiming to create a similar dependency representation suitable for a wide range of languages.

- The Harbin Institute of Technology's Chinese Dependency Treebank, known as HIT-CDT, is a valuable resource for linguistic analysis and research. Published in 2013, this treebank primarily focuses on syntactic relationships while incorporating supplementary semantic information.

This treebank, known as Chinese Dependency Treebank 1.0, contains nearly 50,000 Chinese sentences, equivalent to over 900,000 words. These sentences were randomly selected from People's Daily newswire stories published between 1992 and 1996, and each sentence is annotated with syntactic dependency structures.

- The PKU Chinese Dependency Treebank (PKU-CDT) was created by the Institute of Computational Linguistics at Peking University in 2015. This treebank employs dependency grammar as its core annotation framework and utilizes a multi-view annotation system. It encompasses a wide range of text types, including news, medical, patent, and more, with a total of 1.4 million words.
- Universal Dependencies (UD), originally developed in 2005 as a tool for the Stanford parser, started as a backend to aid in Recognizing Textual Entailment systems. However, since its inception in 2013, it has grown into a significant annotation project. UD is a framework designed for the consistent annotation of grammar elements, such as parts of speech, morphological features, and syntactic dependencies, across various human languages.

The Universal Dependencies (UD) project has transformed into an open, community-driven initiative, boasting a community of over 500 contributors. Together, they are actively engaged in developing treebanks for over 200 languages, spanning a vast linguistic landscape. UD's

overarching goal is to promote the development of multilingual parsers, facilitate cross-lingual learning, and advance parsing research, all while considering language typology as a guiding perspective. UD's annotation scheme draws its inspiration from various sources, including universal Stanford dependencies, Google's universal part-of-speech tags, and the Intersect interlingua for morphosyntactic tagsets. This approach aligns with the project's fundamental philosophy, which seeks to establish a universal framework of categories and guidelines. These standardized elements ensure consistent annotation practices across diverse languages, with the flexibility to incorporate language-specific extensions when deemed necessary. UD's remarkable success can be attributed to its ability to strike a delicate balance between these multiple objectives.

At the moment, There are 6 Mandarin treebanks available on the [universaldependencies.org](http://universaldependencies.org) site:

- PUD: This is a part of the Parallel Universal Dependencies (PUD) treebanks created for the [CoNLL 2017 shared task on Multilingual Parsing from Raw Text to Universal Dependencies](<http://universaldependencies.org/conll17/>).
  - HK: A Traditional Chinese treebank of film subtitles and of legislative proceedings of Hong Kong, parallel with the Cantonese-HK treebank.
  - CFL: The Chinese-CFL UD treebank is manually annotated by Keying Li with minor manual revisions by Herman Leung and John Lee at City University of Hong Kong, based on essays written by learners of Mandarin Chinese as a foreign language. The data is in Simplified Chinese.
  - GSD: Traditional Chinese Universal Dependencies Treebank annotated and converted by Google.
  - GSDSimp: Simplified Chinese Universal Dependencies dataset converted from the GSD (traditional) dataset with manual corrections.
  - PatentChar<sup>26</sup>: A treebank of Chinese patent application texts collected from the Chinese patent office's website CNIPA. The sentences are randomly selected from the patent claims of the IPC section "G" from November 2017 to September 2018.
- In addition to UD, Chinese is also included in the surface-syntactic Universal Dependencies (SUD) project<sup>27</sup>, which provides a complementary set of syntactic annotations that focus on surface-level syntax rather than deep syntax.

SUD, a dependency-based annotation scheme, relies on surface-syntactic distributional criteria for its annotation choices while aiming to maintain compatibility with the UD annotation scheme. It serves as an alternative, rather than a competitor to UD, designed to convey the same informational content. SUD and UD enjoy a high degree of two-way convertibility, allowing conversions between the two without significant information loss. The correspondences between them are typically regular and predictable. Notably, SUD

---

<sup>26</sup> This is our treebank built based on this thesis.

<sup>27</sup> <https://surfacesyntacticud.github.io/data/>



annotations are more concise than UD, replacing a subset of 17 UD relations with three major relations: subj, comp, and mod, with occasional use of udep. Unlike UD, SUD doesn't systematically label content words as heads. Instead, it designates adpositions, subordinating conjunctions, auxiliaries, and copulas as heads based on their role in determining the distribution of the associated syntactic units, leading to certain syntactic relationship directions being reversed between SUD and UD, such as aux, cop, case, and mark relations.

The SUD project currently includes three Chinese treebanks: Chinese-CFL, Chinese-GSD, Chinese-PUD, Chinese-HK, and Chinese-PatentChar. Each of these treebanks was annotated following the SUD guidelines, which define a set of syntactic categories and dependency relations that are specifically designed for surface-level syntax.

In Chapter 3, we will modify the SUD annotation schema to suit our character-level treebank, allowing us to take advantage of its distributional criteria. This adaptation will help us illustrate the similarity of structures at both morphological and syntactic levels in Chinese.

#### 1.2.1.5 Segmentation and Syntactic Parsing of Chinese Patents

In the research titled “Chinese Word Segmentation Techniques for Patent Documents” (Zhang Gui Ping et al., 2010), an analysis of distinctive characteristics pertaining to terms within patent documents is presented. These characteristics encompass the following facets:

1. **Diverse Domains and Technical Vocabulary:** Patent documents span a wide array of domains and inherently encompass an extensive collection of technical terms. This repository of technical terms continuously evolves in tandem with technological advancements.
2. **Structured Language Style:** The linguistic structure of patent documents adheres rigorously to established regulations. Sentences are characterized by formalization, and the selection of terms conforms to standardized norms. Notably, a cluster of terms frequently recurs within the texts.
3. **Proliferation of Affixes:** Technical terms within patents frequently manifest affixes, accentuating the prevalence of morphological variations.
4. **Contextually Bound Named Entities:** The occurrence of named entities is relatively sparse, often bounded by contextual constraints.
5. **Nesting Phenomena:** Technical terminologies exhibit a notable tendency toward employing nesting structures, adding layers of complexity to their composition.
6. **Optimal Length:** Words within patent documents typically span a character length ranging between two and six characters.

Their work utilizes explicit and implicit segmentation markers from the literature as rules for text segmentation and processing. The approach involves employing suffix arrays and the longest common prefix to extract repeated substrings and word frequencies. Additionally, a credibility formula is applied to filter the optional word set, resulting in the extraction of contextual information from the segmented text. Based on this contextual information, a coarse segmentation is conducted, followed by maximum probability word segmentation. Finally, the study employs prefix and suffix rules for

further processing. And got 96.41% and 94.48% as F-scores for close and open test sets, better than other conventional segmenters.

Following their research, Zhai Dongsheng and Ma Wenshan (2011), further investigations were conducted into the word segmentation algorithm for Chinese patent texts. Subsequently, a comprehensive analysis of numerous patent claims revealed that these claims could be broadly categorized into three primary types: feature claims, composition claims, and method claims.

The study involved the utilization of dictionaries, including general dictionaries and domain-specific dictionaries, to establish a segmentation word list for classifying feature words in patent claims. Through this approach, F-scores were achieved at an impressive rate of 95.48%.

In a related study, by Zhang Jie, Zhang Haichao, and Zhai Dongsheng (2014), the authors introduced a term extraction method, which resulted in F-scores of 92%.

Indeed, there is a limited body of work focused on Chinese patent dependency parsing. One notable study is “Research on Semantic and Syntactic Analysis of Patent Literature”, conducted in 2016. This research delved into the realms of syntactic and semantic analysis within patent texts.

In their analysis, the researchers employed dependency parsing tools such as MSTParser and MaltParser. They applied Fine-grained Semantic Code (FSC) to their parsing efforts, achieving a LAS (Labeled Attachment Score) of 76.03%.

### 1.2.2 Review of Terminology Studies

In this section, after introducing the Origin and Development of Terminology Studies (Section 1.2.2.1), we present a brief history of Terminology Studies in Chinese (Section 1.2.2.2). And section 1.2.2.3 presents a list of Chinese terminology works.

#### 1.2.2.1 The Origin and Development

The establishment of modern terminology as an independent discipline, distinct from lexicology, can be traced back to the 1930s when the Austrian scientist Eugen Wüster (1898–1977) pioneered its development. Wüster, widely regarded as the father of modern terminology, laid the groundwork for the German-Austrian school by publishing a series of articles on the General Theory of Terminology, which have been revisited and modified over the years by Wüster himself, other terminologists, and European terminology organizations. He emphasized the significance of concepts in the study of terminology and made substantial contributions to the standardization of terminology under ISO (Wüster 1931; Trojar, 2017).

In contrast to the German-Austrian school, which sought to distance itself from a purely linguistic perspective, the Russian school and Czech school believed that linguistic theories could serve as a starting point for the study of terminology. These schools considered linguistic principles and frameworks as valuable resources in their exploration of terminological studies. Lastly, the

Canadian-Quebec school emerged as another noteworthy branch in the field, critically building upon Wüster's work and incorporating their own insights and perspectives (Feng, 2001).

In recent years, a myriad of innovative approaches have emerged in the field of terminology. These approaches consider the communicative, social, diachronic, and sociocognitive dimensions of terminology, in addition to its linguistic aspects. One such approach, known as textual terminology as presented by Bourigault and Slodzian (1999), aligns well with a lexical-semantic perspective. In textual terminology, the analysis begins with the text itself, and terms are viewed as "constructed" entities resulting from the terminographer's analysis, which takes into account factors such as the term's position in a corpus, validation by experts, and the specific objectives of the terminological description.

It is worth noting that while both lexicology and terminology share an interest in words, terminology is distinguished by its grounding in practical applications, whereas lexicology is a linguistic discipline focused on the study of vocabulary within a language (Marie-Claude L'Homme, 2004).

The practical nature of terminology necessitates its utilization across various domains, requiring knowledge and expertise for effective implementation. As a language for special purposes (LSP), terminology is intricately tied to specific fields, playing a crucial role in facilitating the communication of complex concepts and tasks in scientific research and engineering. In contrast to everyday communication, even slight deviations in meaning can lead to significant discrepancies in outcomes. Therefore, utmost precision is demanded in the selection of terms to ensure their accuracy.

China, along with other countries and regions globally, acknowledges the importance of standardizing its terminology. Liu and Huang (2003) have compiled a list of several characteristics that technical terms in the field of science and technology should possess. They argue that these terms should accurately convey the nature and specific attributes of scientific concepts by carefully selecting the constituent characters. Furthermore, they should strive to avoid ambiguity arising from synonymy, where a single concept may be translated into multiple Chinese terms, such as "overland flow" being rendered as “坡面水流 (pō miàn shuǐ liú)”, “坡面漫流 (pō miàn màn liú)”, “陆面水流 (lù miàn shuǐ liú)”, “地面径流 (dì miàn jìng liú)”, “表面水流 (biǎo miàn shuǐ liú)”, and so on. Additionally, the terms should steer clear of polysemy, where different English terms, such as "mass" and "quality," are both expressed as “质量 (zhì liàng)” in Chinese. As individual terms form an integral part of the scientific and technological system, they should also align harmoniously with the surrounding symbolic framework. Compatibility between disciplines and different languages also serves as an important criterion for selection, facilitating the development of interdisciplinary permeation, communication, and international cooperation in the era of information. Furthermore, considerations such as conciseness and customary usage should also be taken into account for pragmatic ease. The determination of a standard for terminology is precisely the decision-making process that involves weighing the various criteria mentioned above.

Zheng (2010) introduced three main methodologies in theoretical terminology studies:

1. the systems approach, which views concepts as interconnected systems with interacting components, aiming to reveal their wholeness, identify relationships, and create a unified theory, with macro-systems (e.g. chemistry) that encompass millions of terms, and micro-systems (e.g. organic compounds) that consist of smaller subsystems;

2. the semiotic approach, where characteristics of terms represent the primary and core issues in the study, examined from three perspectives, semantics, syntax<sup>28</sup>, and pragmatics;
3. and the linguistic approach, in which the analysis of the formal structure and semantic structure of individual terms, or from a specific terminology system or collection of terms, in order to reveal their shared characteristics and various aspects of the interplay between form and content.

In our research, we primarily focused on the linguistic approach when dealing with the formal structure of individual terms during the treebank annotation process. However, in Section 5, where we constructed taxonomies, we can also view this as an application of the systems approach.

### 1.2.2.2 A Very Short History of Terminology in China

In China, although certain linguistic thoughts related to terminology existed in ancient times, such as the Han Dynasty's "Erya," which can be regarded as an ancient terminology dictionary, modern terminology research had a relatively late start in our country. Subsequently, works like Ge Hong's "Baopuzi"<sup>29</sup>, Zu Chongzhi's "Zhui Shu"<sup>30</sup>, Nadao Yuan's "Shui Jing Zhu"<sup>31</sup>, Shen Kuo's "Mengxi Bitan"<sup>32</sup>, Xu Guangqi's "Nongzheng Quanshu"<sup>33</sup>, Song Yingxing's "Tiangong Kaiwu"<sup>34</sup>, and Li Shizhen's "Bencao Gangmu"<sup>35</sup> made notable contributions. However, it was not until the 1980s that projects focused on the translation and introduction of foreign works, such as Liu Gang's "Introduction to Terminology", Zhang Yide's "Applied Terminology", and Zou Shuming's "Modern Terminology and Dictionary Compilation". (Zheng, 2010)

And Feng's Introduction to Modern Terminology (1997) is widely recognized as the first comprehensive work solely dedicated to modern terminology. In this influential work, he systematically examines the fundamental theories and principles of terminology study, providing a thorough analysis of the types and structures of Chinese terminology and introducing two groundbreaking new theories: the Potential Ambiguity Theory and the Economical Law in Term Formation. Additionally, he explores the theories of terminology compilation, issues related to terminology storage and exchange, and the emerging field of computational terminology. He also takes the lead in the development of the first Chinese technical term database - GLOT-C<sup>36</sup>, which

---

<sup>28</sup> Syntactically, core terms should have the ability to generate new terms, particularly through derivation, such as compound terms.

<sup>29</sup> 葛洪的《抱朴子》

<sup>30</sup> 祖冲之的《缀术》

<sup>31</sup> 哪道元的《水经注》

<sup>32</sup> 沈括的《梦溪笔谈》

<sup>33</sup> 徐光启的《农政全书》

<sup>34</sup> 宋应星的《天工开物》

<sup>35</sup> 李时珍的《本草纲目》

<sup>36</sup> <https://www.zgbk.com/ecph/words?SiteID=1&ID=91786&Type=bkzyb&SubID=44691>

contains all terms of data processing in ISO-2382<sup>37</sup> since 1975. In the database, each term entry contains its English translation, its Chinese synonyms, its structural information and its type of ambiguity.

In Feng's works (Feng, 1989a; Feng, 2004a; Feng, 2004b), a detailed examination of single terms and compound terms is provided. Feng highlights that the majority of Chinese technical terms are compounds consisting of at least two words. These compounds typically exhibit a more rigorous structure and convey a relatively straightforward meaning.

In his works (Feng, 1989 b; Feng, 1989 c; Feng, 1995), Feng explores the compound structure and the potential ambiguity of Chinese technical terms. He approaches the analysis of terms from three distinct perspectives: the PT-structure (phrase-type structure)<sup>38</sup>, which views compound terms as a collection of words or phrases of different types; the SF-structure (syntactic-functional structure)<sup>39</sup>, which describes compound terms based on the syntactic relations between their constituent parts; and the LS-structure (logical-semantic structure), which focuses on the semantic roles of each component. Feng's research specifically addresses the issue of potentially ambiguous structures that arise due to the incompatibility of these three structure types.

For this study, the PT-structure and SF-structure are most relevant to our objective, as they can serve as a base for the annotation of the internal structure of Chinese technical terms. The former is about the delivery of Part-of-Speech (POS)<sup>40</sup>, and the latter is about the labelling of dependency relations in treebanks. In Section 3.3 below, we will also apply his theories as a reference for the Chinese patent treebank annotation.

As previously mentioned, the parallelism between the internal structures of Chinese compound words and syntactic structures is also observed in Chinese technical terms. This means that the composition and arrangement of words within compound terms align with the syntactic rules and structures of the Chinese language. In his work (Feng, 1989c), Feng cites the discussion by renowned Chinese linguist Zhu Dexi on the characteristics of the Chinese language. Zhu Dexi states, "*If we describe the structure of various phrases in sufficient detail, then the structure of sentences is essentially described as well. After all, sentences are nothing more than independent phrases.* (my translation)" This is a remarkable view of syntax, saying that the construction of a sentence follows the same principles as the ones that can be observed inside phrases: A sentence is nothing but a phrase that can be uttered independently; a view that shows influences of Generative Grammar of its time. We agree to the extent that a good description of phrases such as terms is certainly a cornerstone of the syntax of any language.

The influence of Generative thinking in Feng's work can also be seen in the PT (phrase-type) structures that he considers: V + V and VP + VP are two combinations that he calls "Predicative-Complement". The distinction between V and VP seems ad-hoc and irrelevant to us from

---

<sup>37</sup> <https://www.chinesestandard.net/Default.aspx?StdID=GB%2fT+5271>

<sup>38</sup> Corresponding to constituent syntax.

<sup>39</sup> Corresponding to dependency syntax.

<sup>40</sup> Also known as word class or grammatical category, the part-of-speech is a category of words that have similar grammatical properties. There are 14 lexical classes in his work, which are Adjective (A), Adverb (AD), Noun (N), Verb (V), Quantitative Adjective (QA), directional or locative words (FN), Noun Adjective (NA), Noun-Verb (NV), Adjective Verb (AV), Noun-Quantitative Adjective (NQA), Preposition (PR), Noun-Verb Phrase (NVP), Adjective Phrase (AP) and Prepositional Phrase (PP).

a perspective of today's syntax, in particular for an isolating language<sup>41</sup> such as Chinese. We will discuss some problems of this analysis below when discussing different types of ambiguities.

All technical terms in GLOT-C are classified into seven different types of SF-structure: (1) Subject-Predicate (SP), (2) Predicate-Object (PO), (3) Predicate-Complement (PC), (4) Adjective-Head (AH), (5) Adverb-Head (DH)<sup>42</sup>, (6) Repetitive-Verbs (RV), and (7) Repetitive-Nouns (RN)<sup>43</sup>. By combining (4) and (5) into Modifier-Head and (6) and (7) into Conjunction, we get the exact same number of types in the conventional analysis of Chinese phrases presented in Section 1.2.1.1. Examples are shown in Table 1.3.

SF-structure	Example	Possible PT-structures
SP	机器                  学习 jī qì                  xué xí machine learning 'Machine learning'	N+V, N+VP, NP+VP, NP+NVP, N+NV, N+NVP, NQA+NV, NVP+NV, C+V, C+NV, NP+NV, NP+NVP
PO	编制                  程序 biān zhì              chéng xù Preparation          program 'Preparation of the program'	V+N, V+NQA, AV+N, AV+NP, NV+N, VP+N, NVP+N, V+NP, NV+NP, NV+NVP
PC	读                      出 dú                      chū read                    out/leave 'Readout'	V+V, VP+VP
AH	数据                  媒体 shù jù                  méi tǐ data                    media 'Data media'	V+N, V+NQA, AV+N, AV+NP, NV+N, VP+N, NVP+N, V+NP, NV+NP, NV+NVP, QA+NV, NA+NV, A+NV, VP+NV, AP+NV, QA+NP, AV+NVP, VQA+NV, VP+NVP, V+VP, NP+NV, NP+NVP
DH	再                      启动 zài                      qǐ dòng again                   start up 'Reboot'	N+V, N+VP, NP+VP, C+V, C+NV, QA+NV, NA+NV, A+NV, VP+NV, AP+NV, QA+NP, AV+NVP, VQA+NV, VP+NVP, V+VP, VP+VP
RV	读                      写 dú                      xiě read                    write 'Fill out or in'	V+V, VP+VP
RN	字母                  数字 zì mǔ                  shù zì letters                  numbers 'Alphanumerics'	N+N, NP+NP

**Table 1.3 - Examples of the Different Types of SF-structures and Their Possible PT-structures**

<sup>41</sup> An isolating language is characterized by its one-to-one ratio of morphemes to words and a complete absence of inflectional morphology.

<sup>42</sup> This category also includes the structure of auxiliary verbs and verbs.

<sup>43</sup> This category also includes the structure of conjuncted quantifiers (e.g. “吨/公里 (tonne/kilometer)”, “千瓦/小时 (kilowatt/hour)”).

From this perspective, he asserts that “*the syntactic and semantic analysis of Chinese term structures, which are characterized by word-group patterns, serves as a breakthrough for the automatic parsing of Chinese sentences and forms a fundamental aspect of Chinese information processing. Essentially, this represents a study of restricted grammar limited to Chinese technical terms, providing a concise linguistic model for computational linguistics in Chinese.*” (Feng 1988, my translation).

The other notable characteristic of Chinese grammar, as discussed by Zhu Dexi (1985), is the absence of a one-to-one correspondence between word categories and syntactic components, which is referred to in this paper as the PT-SF non-correspondence.

The potential ambiguity emerges when there are several interpretations of SF-structure for a given PT-structure. One example given by the author is “分割字符 (fēn gē zì fú, ‘split character’)”, while the PT-structure is “V N”, the SF-structure can be both “PO” and “AH”, which signifies “segmenting characters” and “characters as segmentation markers” separately.

- |     |                 |   |             |
|-----|-----------------|---|-------------|
| (1) | 直接              | 插入  | 子 程序        |
|     | zhí jiē         | chā rù                                      | zǐ chéng xù |
|     | direct(ly)      | insert(ion)                                 | subroutines |
|     | a.              | ‘direct insertion of subroutines’           |             |
|     | b.              | ‘directly insert into the subroutines’      |             |
|     |                 |   |             |
| (2) | 自动              | 数据  | 处理          |
|     | zì dòng         | shù jù                                      | chǔ lǐ      |
|     | automatic(ally) | data  | processing  |
|     | a.              | ‘automated data processing’                 |             |
|     | b.              | ‘perform the data processing automatically’ |             |

There are two types of ambiguity (Feng, 1989 b). In the first case, namely “true ambiguity structure”, the ambiguity would bring in changes in the semantic interpretation, such as in example (1). In contrast, the second type of ambiguity would not introduce changes in meaning, such as in example (2), and is called “quasi-ambiguity structure”. In Western linguistic terminology, we talk about “spurious” or “structural” ambiguity. Note also that the two examples provided by Feng are not really technical terms in our point of view, because they are completely compositional and do not designate any typical item of a technical field.

In a theoretical sense, the distinction between compounds and phrases in terminology remains unresolved. Feng emphasizes the significance of automatic segmentation in term extraction and other related tasks, like automatic knowledge mining and extraction, particularly given the rapid advancement of technology and the increasing number of technical terms. Yet, the tools at his disposal were the phrase structure grammars of his time.

In his opinion, it is evident that phrases require segmentation, while compounds do not. However, we believe that this distinction is not actually operational, and in this study, we would like to explore the possibility of annotating compounds with syntactic relations. Further discussion is presented in Section 3.3.

Furthermore, Feng's approach leaves unresolved issues concerning complex multi-word phrases composed of three or more words.

For the length of technical terms, in another study "A Preliminary Investigation of the Number of Characters in Chinese Scientific and Technical Terms and Related Issues", conducted by Liang Jixiang (1991), the number of characters in technical terms from the "Physics Terminology" (《物理学名词》) and "Electronics Terminology" (《电子学名词》) lists was counted. The findings reveal that a significant portion of the terms in both lists consists of four characters, followed by terms with three or five characters.

Unfortunately, the study does not provide specific information regarding the number of words. However, considering that the majority of Chinese words are bisyllabic or trisyllabic, we roughly estimate that complex technical terms composed of three or more words account for approximately 10% of these two domains.

### 1.2.2.3 Resources for Chinese Technical Terms

In the previous section, we introduced the GLOT-C, which was the first database for Chinese technical terms. It addresses the challenges related to the structure and semantic ambiguity of technical terms through a comprehensive examination of potential ambiguity.

In this section, we discuss various recent developments in lexicons, dictionaries, thesauri, ontologies, and knowledge bases that encompass Chinese technical terms. While these resources contain valuable information, none of them are openly accessible even for research purposes, although some propose free online access. Furthermore, unlike the GLOT-C, none of these resources focus on the morphological and syntactic analysis of terms.

- **Chinese Thesaurus**《汉语主题词表》<sup>44 45</sup>

The Chinese Thesaurus is a large-scale compilation of subject terms jointly compiled by the Chinese Academy of Sciences' Institute of Scientific and Technical Information (中国信息情报研究所) and the National Library of China (北京图书馆), which commenced in the 1970s. It encompasses a total of 108,568 entries.

In 2009, the Institute of Scientific and Technical Information (中国科学技术信息研究所) embarked on the recompilation of the Chinese Thesaurus, which was divided into separate volumes for Engineering and Technology, Natural Sciences, Life Sciences, and Social Sciences. The Engineering and Technology volume was completed in September 2014, comprising 13 volumes and encompassing 196,000 concepts and 360,000 vocabulary entries. The Natural Sciences volume covers disciplines such as mathematics, physics, chemistry, astronomy, and earth sciences, while the Life Sciences volume includes biological sciences, agricultural sciences, and medicine.

---

<sup>44</sup> <https://ct.istic.ac.cn/site/organize/word>

<sup>45</sup> <https://www.zgbk.com/ecph/words?SiteID=1&ID=78324&Type=bkzyb&SubID=46544>



- **Chinese Classified Thesaurus (CCT)**《中国分类主题词表》<sup>46</sup>

The “Chinese Classified Thesaurus” is the largest integrated subject classification and indexing tool in China, published in 1994. It is based on the Chinese Library Classification (CLC, 《中图法》) system and covers 22 major categories, including philosophy, social sciences, natural sciences, and engineering and technology. The thesaurus includes 51,873 categories according to the CLC system, 120,818 preferred subject terms, 46,434 non-preferred subject terms (entry terms pointing to preferred terms), and 66,373 subject concept phrases (entry phrases corresponding to CLC categories) (Zhang, 2004)

It was developed by the Committee of the Chinese Library Classification (《中图法》编委会) and built upon the foundation of the Chinese Thesaurus (《汉语主题词表》), aiming to achieve integrated subject classification and indexing, simplify indexing work, and enhance retrieval efficiency. The web version of the thesaurus was officially launched in 2010.

- **WIPO Pearl** <sup>47</sup>

WIPO Pearl gives access to scientific and technical terms in ten languages (Arabic, Chinese, English, French, German, Japanese, Korean, Portuguese, Russian, and Spanish), extracted from patent documents. These terms have been meticulously curated by a team of skilled WIPO language experts and terminologists.

One of the key features of WIPO Pearl is the establishment of distinct concept maps that illustrate the relationships and connections between the terms. In cases where an equivalent term is not available in the target language, WIPO Pearl incorporates machine translation capabilities provided by WIPO. Additionally, users can utilize the platform to search for terms and their equivalents across languages within the entire PATENTSCOPE corpus, in which 43 million local or international patent documents are accessible.

- **The Standardized Terminology Database** 规范术语数据库 <sup>48</sup>

The Standardized Terminology Database is a knowledge resource database established based on the approval and publication of disciplinary terms by the National Committee for the Standardization of Scientific and Technical Terminology (全国科学技术名词审定委员会) over the past 30 years. It consists of over 400,000 entries and covers various fields such as basic sciences, engineering and technology, agricultural sciences, medicine, humanities and social sciences, and military sciences. The database's data structure includes discipline, Chinese name, English name, Taiwanese name, abbreviation, full name, alternative names, former names, colloquial names, names in other languages, and definitions.

---

<sup>46</sup> <https://www.zgbk.com/ecph/words?SiteID=1&ID=78324&Type=bkzyb&SubID=46544>

<sup>47</sup> <https://www.wipo.int/reference/en/wipopearl/>

<sup>48</sup> [http://www.cnterm.cn/syfw/sysjk/index\\_25470.html](http://www.cnterm.cn/syfw/sysjk/index_25470.html)

- **Termonline.cn** 术语在线 <sup>49</sup>

Termonline.cn, launched in 2016, is a terminology knowledge service platform managed by the National Committee for the Standardization of Scientific and Technical Terminology. It features a collection of over 450,000 standardized terms, covering more than 100 disciplines in various fields, including basic sciences, engineering and technology, agricultural sciences, medicine, humanities and social sciences, and military sciences. The platform provides free functions such as terminology retrieval, terminology sharing, terminology correction, terminology collection, and terminology solicitation (Du, 2021).

- **Hownet** 知网 <sup>50 51</sup>

Hownet is a language knowledge base for Chinese, which is characterized by its application of “sememe” as indivisible semantic units of concepts. In Hownet, tens of thousands of Chinese and English words are annotated with 2,000 sememes. (Dong, 2003)

- **OpenKG.CN** 中文开放知识图谱 <sup>52</sup>

OpenKG is an open knowledge graph project launched by the Language and Knowledge Computing Professional Committee of the Chinese Information Processing Society of China. It comprises 122 graph sets and 54 tool sets, aiming to provide a platform for sharing and utilizing knowledge in an open and collaborative manner. Two projects of special interest for tech-mining are the SciKG by Tsinghua University and the ai-patent by Patsnap.

- SciKG<sup>53</sup> is a knowledge graph project of scientific publication in the computer science domain, containing entities of concepts, definitions of concepts, authors and articles, and relations between them.
- ai-patent<sup>54</sup> is a knowledge base of patent information, which consists of each patent, its information on abstract, application date, application number, assignee, claim, CPC, etc.

---

<sup>49</sup> <https://www.termonline.cn/index>

<sup>50</sup> <https://openhownet.thunlp.org/>

<sup>51</sup> <https://github.com/thunlp/OpenHowNet>

<sup>52</sup> <http://openkg.cn/>

<sup>53</sup> <http://openkg.cn/dataset/scikg>

<sup>54</sup> <http://openkg.cn/dataset/ai-patent>

### 1.3 Defining Notions

Prior to delving into the specific issues addressed in this study, it is essential to establish a clear understanding of the definitions and applications of certain significant concepts that revolve around the central idea of this work, namely, the “term” or “technical term”, and the surrounding concepts that may introduce ambiguity.

In this study, we refrain from employing the notion “word” in the context of Chinese, due to its inherent ambiguity when compared to characters, and phrases, which may introduce uncertainty within the study’s scope (see Section 1.3.1), and also considering the technical nature of the patent corpus. Instead, we employ the notion of “term”.

As the central topic of terminology studies, many terminologists have given definitions of “term” in their works.

- Eugen Wüster's 1931 paper, "Terminologie als angewandte Sprachwissenschaft" (Terminology as Applied Linguistics), is considered to be a seminal work in the field of terminology, see Section 1.2.2.1. In this paper, Wüster proposed a general theory of terminology that is still influential today. Wüster views “terms” as labels assigned to concepts, and concepts as mental constructs arising from the perception of real-world objects and phenomena. Ideally, there should be a one-to-one correspondence between terms and concepts within a specific domain. However, this ideal is often unattainable due to the richness and diversity of concepts and terminology.
- Sager’s Perspective (1990): Sager distinguishes between “terms” and “words”. Terms are characterized by special reference within a discipline, while words have general reference. This differentiation highlights the role of “terminology” as distinct from “vocabulary”.
- Kageura and Umino’s Perspective (1996): According to Kageura and Umino, terms are linguistic units that characterize specialized domains. They are lexical units whose meaning is considered in relation to a specific domain of expertise.
- Cabré’s Perspective (1998): Cabré emphasizes that terms, like words, are distinctive and meaningful signs used in specialized language discourse. They have both a systematic side (formal, semantic, and functional) and a pragmatic side, serving as units in specialized communication.
- ISO Definition (2011)<sup>55</sup>: The International Organization for Standardization (ISO) 10241-1:2011 defines a “term” as a verbal designation of a general concept in a specific subject field. This designation may contain symbols and variants.

To draw a conclusion, the characteristic mentioned by all three definitions is that terms express a concept and are related to specific fields or domains, or, as the Wiktionary puts it a technical term is “a word that has a specific meaning within a specific field of expertise.”<sup>56</sup>

<sup>55</sup> <https://www.iso.org/obp/ui/#iso:std:iso:10241:-1:ed-1:v1:en>

<sup>56</sup> [https://en.wiktionary.org/wiki/technical\\_term](https://en.wiktionary.org/wiki/technical_term)

Apart from the conventional terminology studies from a semantic theoretical perspective, most definitions of “term” come from a statistical perspective in the Natural Language Processing studies, especially the Automatic Term Extraction (ATE) task.

In the context of term extraction, there are two crucial factors referring to the qualities of terms to consider: *unithood* and *termhood*, that Nakagawa (2001) calls “two essential aspects of the nature of terms”.

*Unithood* pertains to the strength and stability of syntagmatic combinations and collocations, particularly in the case of multi-word terms like noun phrases that refer to a single conceptual unit. It measures how likely it is for the words in a phrase to appear together as a fixed combination, rather than as independent words. Unithood can be assessed by a collocation analysis, which involves examining how frequently words co-occur in a specific order. Tools and statistical methods like mutual information or t-score (Manning and Schütze, 1999) can help quantify the strength of these collocations. English examples of high unithood include “Climate Change”, “Heart Attack”, and “Deep Learning”. Each of these terms are composed of words that frequently appear together and form a concept that is distinct from the individual meanings.

On the other hand, *termhood* focuses on the degree of association between a linguistic unit and domain-specific concepts. Termhood is often calculated based on term frequency and frequency bias, and higher termhood values indicate the term’s greater ability to distinguish different domains. Formal descriptions of these concepts are provided by Kageura and Umino (1996). English examples of high termhood include “Neural Network” in Computer Science, “Jurisprudence” in Law, and “Mitosis” in Cell Biology. These terms are heavily associated with studies in their respective fields and have a high termhood in domain-specific contexts, as opposed to its general meaning or usage.

To sum up, *unithood* deals with the stability of syntactic units, while *termhood* measures a lexical unit’s relevance to domain-specific concepts. Term extraction involves the identification of both unithood and termhood, with unithood being relevant to complex terms and termhood being applicable to both simple and complex terms.

In the context of patent claims, it’s crucial to recognize that, apart from standardized phrases such as “其特征在于” (qí tè zhēng zài yú, “characterized by”) and 权利要求 (quán lì yāo qiú, ‘claim’), the majority of the content is comprised of specialized technical terminology, in the sense that these terms do not appear frequently in general texts or have a different meaning there. Note that the hierarchical structure of the patent classification could define a “termhood” relative to the IPC classification: A term could be specific to only a subclass (such as G06A) as opposed to another subclass (G06B) or be used in the whole section G, as opposed to H. We leave this to future work. The primary focus here is on identifying the targeted terms of lexical substitution with their hypernyms within patent texts. In this specific context, evaluating the termhood of these technical phrases only has to distinguish terms that are of technical nature, relating to the technical domain, from those that are specific to patent domain, i.e. they appear in many different patent domains. However, the concept of unithood remains a pivotal factor for assessment.

### 1.3.1 Unithood in the context of patents

For the purposes of this study, the notion of “term” is used to denote a lexical unit identified in the patent corpus and also a target for lexical substitution later. This terminology deviates somehow from the conventional interpretation of domain-specific and fixed technical concepts. Essentially, these are viewed as potential terms that are subject to extraction and substitution by their hypernyms. In this subsection, we are going to conclude the criterion of such a term, by comparing with other important notions in the tasks, such as “word”, “noun phrase” and “multi-word expressions”, therefore, by determining the unit boundaries based on the intrinsic strength between components of a term.

In this work, we are primarily interested in the technical nominal lexical units, including single nouns and noun phrases of specific technical domains. This term can vary from a smallest syntactic units (the usual definition we give to “word”) to a sequence of words.

In the Chinese language, when conveying complex technical concepts, terms typically deviate from the most common disyllabic structure found in general words. Instead, they tend to consist of three to five characters, as noted by Liang Jixiang (1991).

Additionally, Tsou and colleagues (2020) contend that technical terms exhibit distinctive characteristics due to their incorporation of elements from Classical Chinese. The autonomous nature of these classical characters contributes to the stability and fixed nature of terms, setting them apart from simple character sequences.

On the other hand, it’s also essential to note that not all “terms” are synonymous with multi-word expressions (MWEs).

Savary (2008) succinctly outlined three key criteria for defining multi-word expressions: 1. They consist of two or more graphical words; 2. They exhibit varying degrees of non-compositionality in terms of morphology, syntax, distribution, or semantics; 3. They maintain unique and consistent references. However, a term can be composed of one single word. And one hypothesis suggests that, due to their deliberate construction, technical terms tend to be more compositional and possess transparent internal structures compared to multi-word expressions, which are typically less consciously constructed and may exhibit less clear internal relationships.

Furthermore, in alignment with the practical objectives of patent attorneys in drafting, we do not strictly adhere to an indivisible criterion for the notion of “terms”. The specific criteria utilized for segmentation depending on the term’s internal structure type are comprehensively discussed in Sections 3.1.1 and 4.3.2.

### 1.3.2 Termhood: distinguishing patent-domain specific terms from technical domain specific terms

In order to distinguish “terms” from general words and phrases, especially noun phrases, we examine a list of desirable and undesirable term-like expressions in the patent claims.

Desirable Expressions	Undesirable Expressions
呼吸训练器 ‘Respiratory Training Device’ 复配乳化增稠剂 ‘Complex Emulsifying Thickening Agent’ 壳体上壁 ‘Upper Shell Body’ 第二转动螺栓 ‘Second Rotating Bolt’ 压力传感器 ‘Pressure Sensor’ 灌装系统 ‘Filling System’ 槽轮 ‘Slot Wheel’	特征 ‘Feature’ 权利要求 ‘Claim’ 上端 ‘Upper End’ 生产加工工艺 ‘Production and Processing Technology’ 重量配比 ‘Weight Ratio’ 如下步骤/以下步骤 ‘The Following Steps’ 所述温度 ‘Said Temperature’

**Table 1.4 - Examples of Desirable and Undesirable Term-like Expressions**

By contrasting the desirable and undesirable term-like expressions presented in Table 1.4, it becomes apparent that the desirable terms are closely linked to specific techniques and system components in an individual patent. In contrast, other term-like expressions are commonly encountered across various patents, such as “特征 ‘Feature’” and “权利要求 ‘Claim’”, which are nearly ubiquitous in most patents.

In the timeframe of this thesis, we do not attempt to provide a detailed analysis of the extent to which the desired terms can be identified by their frequency across IPC classes. We will simply make use of the reference numerals, and make the bold assumption that the list of numbered terms equals the desired term-like expressions.

The following table gives an idea what reference numerals are and how they appear in the patent application.

Chinese	English
CN114028779A	
图中:1、训练床;2、防落护栏;3、放置板;4、呼吸训练箱;5、雾化箱;6、吸痰泵;7、呼吸管;8、雾化管;9、吸尿管;10、四通管;11、吸嘴;12、显示控制屏;13、检测圆盘;14、碰撞检测板;15、钢珠;16、转动齿轮;17、伺服电机;18、通气横管;19、第一训练管;20、第二训练管;21、连接管;22、电动滑台;23、滑块;24、磁铁块;25、清理环;26、排气通道;27、进气通道;28、第一单向活瓣;29、光线接收板;30、紫外线灯;31、阻抗机构;3101、活塞;3102、阻力塞;3103、空气腔;3104、凸块;32、第二单向活瓣;33、第三单向活瓣;34、调节筒;35、电磁铁;36、液压杆;37、支撑板;38、压力传感器;39、矩形槽;40、推块;41、弹簧;42、连接件;43、限制杆;44、双层缓冲垫;45、安装杆;46、呼吸肌训练罩;47、定位贴;48、水平定位槽;49、移动块;50、垂直定位槽;51、滑动板;52、连接杆;53、空腔;54、充气模块;55、抽气模块;56、第一导气管;57、第二导气管。	In the figure: 1. a training bed; 2. a falling-preventing guardrail; 3. placing the plate; 4. a breath training box; 5. an atomization box; 6. a sputum suction pump; 7. a breathing tube; 8. an atomizing tube; 9. an extraction tube; 10. a four-way pipe; 11. a suction nozzle; 12. displaying a control screen; 13. detecting the disc; 14. a collision detection plate; 15. steel balls; 16. a rotating gear; 17. a servo motor; 18. a horizontal ventilation pipe; 19. a first training tube; 20. a second training tube; 21. a connecting pipe; 22. an electric sliding table; 23. a slider; 24. a magnet block; 25. cleaning a ring; 26. an exhaust passage; 27. an air intake passage; 28. a first one-way valve; 29. a light receiving plate; 30. an ultraviolet lamp; 31. an impedance mechanism; 3101. a piston; 3102. a resistance plug; 3103. an air chamber; 3104. a bump; 32. a second one-way valve; 33. a third one-way valve; 34. an adjusting cylinder; 35. an electromagnet; 36. a hydraulic lever; 37. a support plate; 38. a pressure sensor; 39. a rectangular groove; 40. a push block; 41. a spring; 42. a connecting member; 43. a restraining bar; 44. a double-layer cushion pad; 45. mounting a rod; 46. a respiratory muscle training mask; 47. positioning paste; 48. a horizontal positioning groove; 49. a moving block; 50. a vertical positioning groove; 51. a sliding plate; 52. a connecting rod; 53. a cavity; 54. an inflation module; 55. an air extraction module; 56. a first

	air duct; 57. a second air duct.
CN107411005A	
<p>图中:4、清洗池;5、洗瓶机;6、烘干机;7、垂直标签机;8、水平标签机;9、输送系统;11、槽体;12、过滤装置;13、出料管;14、蒸汽管;15、搅拌装置;111、进料口;21、送料系统;22、浓缩系统;23、收集系统;24、灌装系统;211、中转罐;212、一级过滤器;213、二级过滤器;222、蒸发器;223、真空浓缩机;224、真空气管一;2221、蒸汽管一;2222、蒸汽管二;231、三级过滤器;232、收集装置;233、进料管;234、出料管;2321、成品罐;2322、搅拌装置;2323、蒸汽管二;2324、冷却水管;2325、热处理管;241、四级过滤器;242、灌装罐;243、真空气管二;244、保温进水管;245、保温出水管;246、进蜜管;247、出蜜管;51、机箱;52、旋转盘;54、供水系统;55、控制系统;541、进水管;542、出水管;543、喷水管;544、淋水管;71、基板;72、标签卷;73、胶带辊;75、过渡辊;76、张力机构;77、驱动装置;78、分离板;79、挡板;710、抚平板。</p>	<p>In figure:4th, service sink;5th, bottle washing machin;6th, dryer;7th, vertical price labeling;8th, horizontal price labeling;9th, induction system;11st, cell body;12nd, filter;13rd, discharge nozzle;14th, steam pipe;15th, agitating device;111st, charging aperture;21st, feed system;22nd, concentration systems;23rd, collection system;24th, bulking system;211st, transfer tank;212nd, grade one filter;213rd, secondary filter;222nd, evaporator;223rd, vacuum decker;224th, vacuum tracheae one;2221st, steam pipe one;2222nd, steam pipe two;231st, three-level Filter;232nd, collection device;233rd, feed pipe;234th, discharge nozzle;2321st, finished pot;2322nd, agitating device;2323rd, steam Pipe two;2324th, cooling water pipe;2325th, heat treatment pipe;241st, level Four filter;242nd, filling tank;243rd, vacuum tracheae two;244th, it is incubated water inlet pipe;245th, it is incubated outlet pipe;246th, honey tube is entered;247th, honey tube is gone out;51st, cabinet;52nd, rotating disk;54th, supply water System;55th, control system;541st, water inlet pipe;542nd, outlet pipe;543rd, sparge pipe;544th, spray pipe;71st, substrate;72nd, label Volume;73rd, adhesive tape roller;75th, transition roller;76th, tension mechanism;77th, drive device;78th, separating plate;79th, baffle plate;710th, plate is smoothed.</p>
CN105143092B (Description)	
<p>图1示出了包括第一可旋转牵引筒2和第二可旋转牵引筒3的本发明的牵引绞盘1的示意图,其中,以轴向平行的方式设置第一和第二可旋转牵引筒2、3。在每个牵引筒2、3的轴向周缘周围,设置了多个槽轮或带轮4-15,其中,槽轮4-15中的每个具有与缆索或绳索16互补的凹槽。要注意的是,槽轮应解释为单独的盘(与在图1中的槽轮4-6和13的情况一样)或作为物体的部分或整体组成部分的盘(与在图1中的槽轮7-12和14-15的情况一样)。绳索16在图1中显示为以轴向并排关系在牵引筒2、3的凹槽之上实施多次缠绕绳索16,绳索16的端部离开与槽轮4轴向相对的第二筒3上的槽轮15,该端部在该槽轮4上进入第一筒2内。在绳索16在高负荷侧17(即,用于拉入或降低讨论中的负荷的侧)上进入第一筒2时,该绳索围绕第一筒的第一可旋转槽轮4的一部分17弯曲。在这个实施方式中,第一可旋转槽轮4主要用作导向盘,这是因为根据特定的设置,该第一槽轮的旋转/弯曲通常等于或小于90°。在通过期望的弯曲通过第一槽轮4之后,绳索16继续进入与第一槽轮4一样位于第二筒3的高负荷侧17'上的第二槽轮5内,然后,继续至与第一槽轮4相邻的位于第一筒2上的第三槽轮6内。重复这个布置,直到绳索16在位于低负荷侧18'上的最后一个槽轮15上离开牵引绞盘。在这个实施方式中,最后一个槽轮15是在第二筒3上</p>	<p>Fig. 1, which is shown, includes the towing capstan 1 of the invention of the first rotatable pull cylinder 2 and second rotatable pull cylinder 3 of schematic diagram, wherein, the first and second rotatable pull cylinders 2,3 are set in axially in parallel mode. In each pull cylinder 2,3 Around axial periphery, there is provided multiple sheaves or belt wheel 4-15, wherein, each having and cable or rope 16 in sheave 4-15 Complementary groove. It should be noted that sheave should be interpreted that single disk (as the situation of sheave 4-6 and 13 in Fig. 1) Or the disk of a part or whole part part as object (as the situation of sheave 7-12 and 14-15 in Fig. 1). Rope 16 are shown as implementing repeatedly winding rope 16, rope 16 in axial side by side relationship on the groove of pull cylinder 2,3 in Fig. 1 End leave sheave 15 on second 3 axially opposing with sheave 4, the end enters in first 2 on the sheave 4. When rope 16 enters first 2 in high load side 17 (that is, the side for pulling in or reducing the load in discussing), the rope A part 17 around the first rotatable sheave 4 of first is bent. In this embodiment, the first rotatable sheave 4 is led To be used as positioning disk, because according to specific setting, rotation/bending of first sheave is usually equal to or less than 90°. By it is desired curve through the first sheave 4 after, rope 16 goes successively to be located at as the first sheave 4 height of second 3 In the second sheave 5 on load side 17', then, threeth sheave 6 that is located at first tin 2 on adjacent with the first sheave 4 is proceeded to It is interior. This arrangement is repeated, until rope 16 leaves towing capstan on last sheave 15 on the 18' of underload side. In this embodiment, last sheave 15 is (axial direction) end sheave on second 3. ...</p>

<p>的(轴向)端部槽轮。 ...</p>	
<p>CN105143092B (Claims)</p>	
<p>1.一种牵引绞盘(1), 所述牵引绞盘用于绞吊细长物品(16), 所述细长物品具有能连接至负荷的高张力端以及能连接至储存装置的低张力端, 所述牵引绞盘(1)包括: 两个或更多个能旋转的筒(2、3), 这些筒设置为彼此相邻, 并且这些筒的旋转轴线大体上平行, 每个所述筒(2、3)具有带凹槽的多个平行的周向槽轮(4-15), 所述槽轮(4-15)相对于彼此轴向地偏移, 以允许以螺旋的方式围绕这两个筒(2、3)的槽轮(4-15)缠绕所述细长物品(16), 其中, 多个所述槽轮(4-15)包括: 固定槽轮(7-12、14-15), 这些固定槽轮相对于它们下面的筒(2、3)是固定的; 以及 能旋转槽轮(4-6), 这些能旋转槽轮相对于它们下面的筒(2、3)是能旋转的, 所述筒(2、3)中的至少一个筒的所述能旋转槽轮(4-6)的大部分被设置为在所述绞盘的高负荷支撑侧上彼此相邻, 其特征在于, 所述能旋转槽轮(4-6)中的至少一个的旋转速度能通过至少一个制动装置(20)而降低, 其中, 所述至少一个制动装置(20)通过朝向所述能旋转槽轮的底侧施加压力来制动这些槽轮, 该压力明显地降低这些槽轮的旋转速度。</p>	<p>1. a kind of towing capstan (1), the towing capstan is used for twisted hanging elongate articles (16), the elongate articles, which have, to be connected To the high-tension end of load and it can be attached to the low-tension end of storage device, The towing capstan (1) includes: Two or more revolvable cylinders (2,3), these cylinders are disposed adjacent to, and these rotation axis is substantially It is upper parallel, Each cylinder (2,3) has multiple parallel circumferential sheaves (4-15) with groove, the sheave (4-15) relative to Axially offset each other, to allow the sheave (4-15) in a helical pattern around the two cylinders (2,3) to wind the elongated thing Product (16), Wherein, Multiple sheaves (4-15) include: Fixed sheave (7-12,14-15), these fixed sheaves are fixed relative to the cylinder (2,3) below them; And Sheave (4-6) can be rotated, these can rotate sheave be relative to the cylinder (2,3) below them it is revolvable, The major part that can rotate sheave (4-6) of at least one in the cylinder (2,3) is arranged in the capstan winch High load capacity support-side on it is adjacent to each other, Characterized in that, The rotary speed of at least one that can be rotated in sheave (4-6) can be dropped by least one brake apparatus (20) It is low, Wherein, at least one described brake apparatus (20) brakes this by applying pressure towards the bottom side that can rotate sheave A little sheaves, the pressure significantly reduces the rotary speed of these sheaves.</p>

**Table 1.5 - Examples of Indexing of Components in Patent Description and Claims (Reference Numeral Lists)**

In highly specialized text such as patents, identifying terms that are suitable for extraction and substitution to assist and inspire patent drafting can be a challenging task without expert manual examination. In the absence of expert annotations, an alternative approach is to utilize the indexing of components found within the description section of certain patents, the so-called “reference numerals”. This component indexing commonly refers to representations of component symbols in the accompanying drawings and manifests within the “description of the drawings” section or in the “embodiment” section of the patent description, with the corresponding numbers appearing in the drawing itself. In some patents the list of numbered terms are listed in an independent section called “Reference Numeral List” The first two examples in Table 1.5 demonstrate component indexing in the "description of the drawings," while the third example showcases indexing in the “embodiment” section. The first type of indexing is a more concise format. Additionally, indexations can also appear



in patent claims, as illustrated by the fourth example in Table 1.5 (the third and fourth examples are from the same patent). Note that in that latter case, the reference numerals appear in parenthesis).

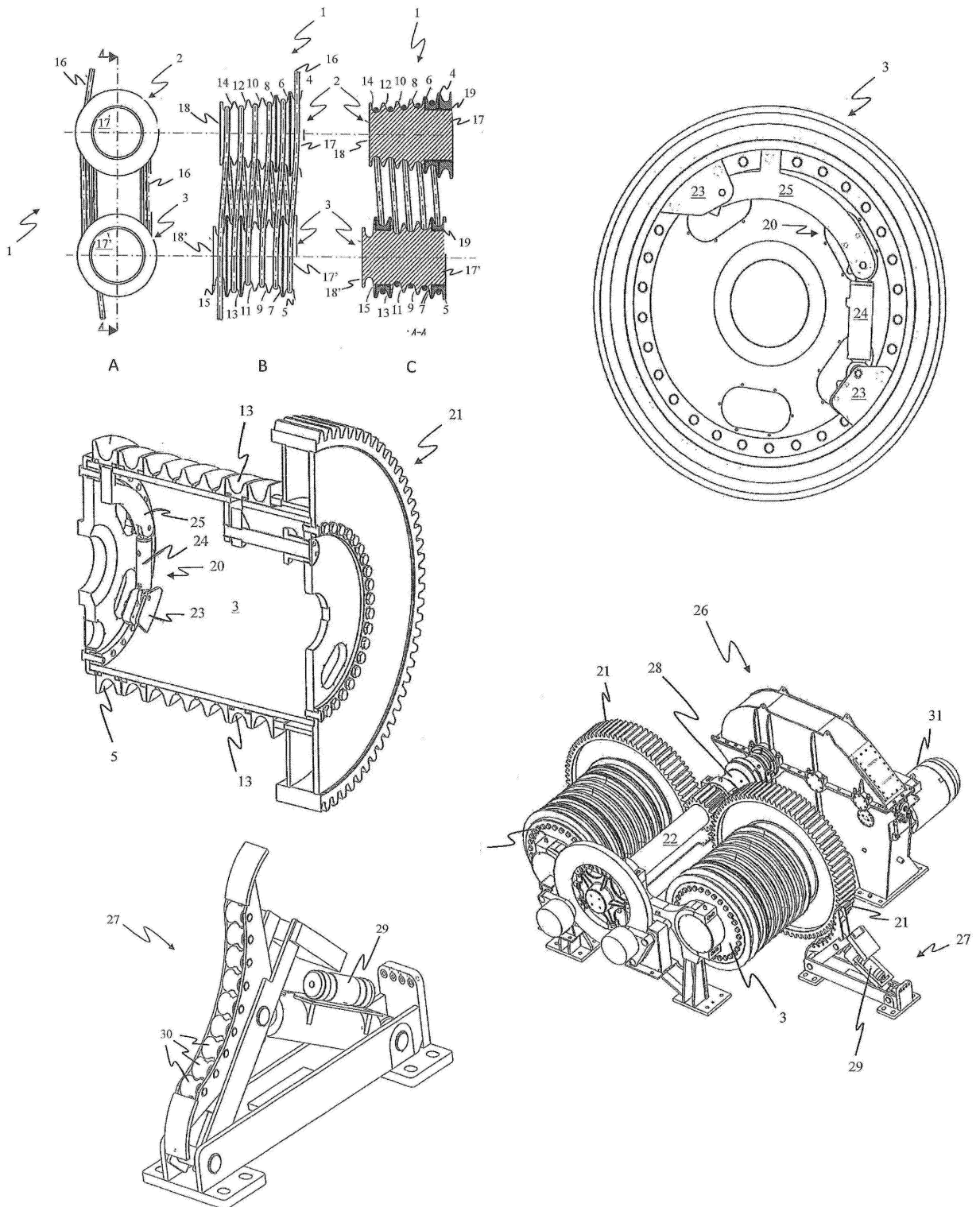


Figure 1.5 - Drawing from patent CN105143092B shown in Table 1.5

## Chapter 1 - Linguistic and technical background of the research

Further details on the use of reference numeral lists in the term recognition and the lexical variation is presented in Section 4.3.2 and 5.3.2.

## 1.4 Hypothesis, Difficulties and Methodology

In this research, as described in the introduction, our primary focus lies in examining terms used within Chinese patents, particularly in the claims section. Our analysis centres on understanding their linguistic attributes, their roles in sentence structures, and their variability, specifically in the context of claim scope.

Our primary objective is to provide an in-depth analysis of the internal structure of Chinese technical terms through dependency syntax annotation. Additionally, we aim to investigate the potential for lexical variation to cater to the specialized requirements of claim scoping by developing a novel technical taxonomy.

As already described in Introduction, the study is divided into two main segments: the first part entails syntactic analysis (Chapter 3), while the second part involves recognizing terms based on the character-level parser and constructing technical taxonomies relevant to patents for facilitating the recognition and the substitution of technical terms by their hypernyms (Chapter 4 and Chapter 5).

The whole pipeline is as in the following Figure 1.5.

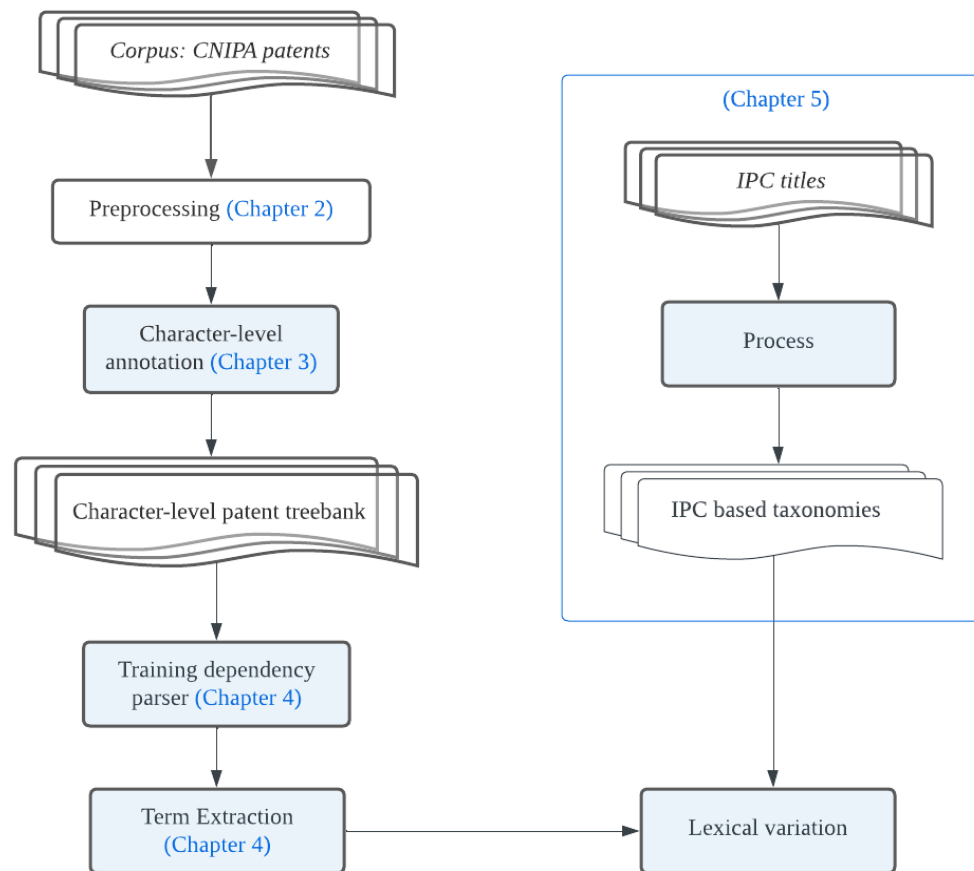


Figure 1.6 - The Entire Processing Pipeline of this Study

The challenges we face in this research include the limited availability of Chinese resources such as treebanks, models, wordnets, and terminology databases. Additionally, there is a lack of coherent Universal Dependencies (UD) or Surface-Syntactic Universal Dependencies (SUD) treebank guidelines specific to Chinese. The Chinese annotation guidelines developed for the UD Mandarin HK treebank (Leung et al. 2016), discussed in Section 1.2.1, only apply to this specific treebank, which moreover annotates translations of spoken Cantonese, a very different genre from patent texts. In collaboration with Wu Qishen (PhD student from the laboratory MoDyCo), we are currently developing an annotation guideline based on the work of this thesis that aims to enable coherent surface syntactic (SUD) annotation of Mandarin Chinese, to be published soon on the SUD website<sup>57</sup>. Also note that commonly used NLP tools like SpaCy do not provide native support for noun chunk extraction in Chinese.

Moreover, as our research involves very specific tasks, we encounter a scarcity of existing benchmarks to measure our progress. In response to these challenges, we have developed our own evaluation methods tailored to the unique nature of our research objectives.

Throughout this thesis, given consideration of the difficulties above, we delve into several critical aspects and address key questions related to Chinese technical terms and their syntactic structures. The principal inquiries we explore are as follows:

- Can we coherently describe the internal structure of Chinese technical terms using standard syntactic links?
- Do Chinese character-based parsers (which consider the internal structure of terms) yield better results of syntactic parsing than word-based approaches?
- Does syntax aid in terminology recognition, especially for new terms?
- How can we evaluate a taxonomy?

These interests and inquiries serve as the central focal points of this thesis. The overarching goal is to offer comprehensive insights and analyses regarding the composition and identification of Chinese technical terms. Additionally, we seek to explore the role of syntax in terminology recognition and conduct evaluations and comparisons related to taxonomies, including their distance and other semantic relations.

---

<sup>57</sup> <https://surfacesyntacticud.github.io/>

## Chapter 2 - Collection and Preprocessing of Patent Texts

This chapter is dedicated to the preprocessing of the corpus, commencing with the procurement of the initial published patent data and culminating in the systematic arrangement of the unprocessed texts in a structured manner, demarcating distinct sections such as the abstract, description, and claim portions of each patent. In Section 2.1, the discourse delves into the intricacies of patent collection, followed by the elucidation of the fundamental structure of Chinese patent applications in Section 2.2. The subsequent Section 2.3 intricately explores the process of XML parsing along with the application of various cleansing techniques to the parsed files. Section 2.4 comprehensively outlines the procedure of patent categorization across the eight IPC domains. Lastly, Section 2.5 provides an analysis of certain linguistic attributes inherent in Chinese patent texts, particularly those pertaining to claim sentences.

## 2.1 Collecting patent application data from CNIPA

In the context of extensive data mining endeavours, there arises the imperative to amass a substantial corpus of patent texts in the Chinese language. To fulfil this requisition, an acquisition was undertaken encompassing all patent application data spanning a span of four years. This data was sourced from the official repository of the China National Intellectual Property Administration (CNIPA; Chinese: 国家知识产权局<sup>58</sup>), commonly known as the Chinese Patent Office. The CNIPA, functioning as the patent office of the People's Republic of China (PRC), is vested with the responsibility of overseeing patent-related affairs and coordinating international engagements within the realm of intellectual property.

The data collection process encompassed the procurement of all downloadable application files corresponding to published Chinese patents within the timeframe spanning November 2017 to March 2021. These files are made accessible through the official CNIPA platform and are subjected to updates every three to four days. With each update, a collection of ZIP files is disseminated, each housing a multitude of patent directories. The structural arrangement of these directories follows a hierarchical framework, as depicted below.

Consequently, the outcome of our efforts yielded an aggregated accumulation of XML patent files amounting to an excess of 300 gigabytes. This compilation is demarcated by a yearly distribution, encompassing the entirety of patent applications across the span of 12 months for the years 2018, 2019, and 2020. However, for the years 2017<sup>59</sup> and 2021<sup>60</sup>, the availability is relatively constrained, encompassing only 2 and 3 months' worth of data respectively.

	<b>Number of month</b>
<b>2017</b>	2
<b>2018</b>	12
<b>2019</b>	12
<b>2020</b>	12
<b>2021</b>	3
<b>TOTAL</b>	41

**Table 2.1 - The Collection of Published Patent Applications from 2017 to 2021**

<sup>58</sup> CNIPA (China National Intellectual Property Administration), the former SIPO (State Intellectual Property Office), changed its name on September 3, 2018.

<sup>59</sup> The data only remains 6 months on the site, the application files of November 2017 are the most ancient when we started the collection.

<sup>60</sup> We have downloaded all applications in 2021 but some zip files are damaged and can not be used for analysis.

The fundamental structure of the file is visually depicted in Figure 2.1 below. Within each month, a notable pattern emerges wherein four or five updates occur, and correspondingly, each update is characterized by a collection of zipped documents. These zipped documents individually encapsulate one patent per document.



**Figure 2.1 - The Structure of CNIPA Monthly Updated Documents**

## 2.2 The general structure of a patent application

In the distributed bulk format, a standard patent application consists of a single XML file accompanied by several JPG files. These JPG files encompass diverse elements such as mathematical and chemical formulas, tables, flowcharts, component diagrams, deployment diagrams, and more. As an illustrative instance, consider the patent with the code 'CN102020000987007', which is provided below. Within its directory, this specific patent includes a singular XML file and a total of 23 JPG files.



**Figure 2.2 - The Content inside the ZIP File**

The subsequent Figure 2.3 presents the textual arrangement of a patent. Positioned below the `<?xml>` tag and `<!DOCTYPE>` tag, the entirety of the patent application resides within the `business:PatentDocumentAndRelated` tag, which encompasses five subordinate tags: `business:BibliographicData`, `business:Abstract`, `business:Description`, `business:Drawings`, and `business:Claims`.



```

<?xml version="1.0" encoding="utf-8" standalone="no" ?>
<!DOCTYPE business:PatentDocumentAndRelated SYSTEM "/DTDS/ExternalStandards/ippbdb-entities.dtd" []>
<business:PatentDocumentAndRelated xmlns:base="http://www.sipo.gov.cn/XMLSchema/base"
xmlns:business="http://www.sipo.gov.cn/XMLSchema/business" xmlns:m="http://www.w3.org/1998/Math/MathML"
xmlns:tbl="http://oasis-open.org/specs/soextblx" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.sipo.gov.cn/XMLSchema/business
/DTDS/PatentDocument/Elements/OtherElements.xsd" xsdVersion="V2.2.1"
file="CN102020000987007CN00001121831480AFULZH20210105CN00M.XML" dateProduced="20210101"
status="C" lang="zh" country="CN" docNumber="112183148" kind="A" datePublication="20210105">
  <business:BibliographicData>
    </business:BibliographicData>
  <business:Abstract dataFormat="original" lang="zh" sourceDB="national office" processingType="original"
creator="03">
    </business:Abstract>
  <business:Description lang="zh" dataFormat="original" sourceDB="national office" creator="03">
    </business:Description>
  <business:Claims lang="zh" dataFormat="original" sourceDB="national office" creator="03">
    </business:Claims>
</business:PatentDocumentAndRelated>

```

Figure 2.3 - Example of a Part of the Content One XLM File of Chinese Patent

<?xml version="1.0" encoding="utf-8" standalone="no" ?>
<!DOCTYPE business:PatentDocumentAndRelated SYSTEM "/DTDS/ExternalStandards/ippbdb-entities.dtd" []>
<business:PatentDocumentAndRelated xmlns:base="http://www.sipo.gov.cn/XMLSchema/base" xmlns:business="http://www.sipo.gov.cn/XMLSchema/business" xmlns:m="http://www.w3.org/1998/Math/MathML" xmlns:tbl="http://oasis-open.org/specs/soextblx" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.sipo.gov.cn/XMLSchema/business /DTDS/PatentDocument/Elements/OtherElements.xsd" xsdVersion="V2.2.1" file="CN102020000987007CN00001121831480AFULZH20210105CN00M.XML" dateProduced="20210101" status="C" lang="zh" country="CN" docNumber="112183148" kind="A" datePublication="20210105">
<business:BibliographicData> </business:BibliographicData>
<business:Abstract dataFormat="original" lang="zh" sourceDB="national office" processingType="original" creator="03"> </business:Abstract>
<business:Description lang="zh" dataFormat="original" sourceDB="national office" creator="03"> </business:Description>
<business:Claims lang="zh" dataFormat="original" sourceDB="national office" creator="03"> </business:Claims>
</business:PatentDocumentAndRelated>

Table 2.4 - The XML Structure of the CNIPA Dataset

Within the “business:BibliographicData” section, there is a compilation of metadata and publication references pertinent to the patent. The remaining four subordinate tags correspondingly represent the four constituents comprising the patent application document: the abstract, description, drawings, and claims.

The Table 2.4 gives the detailed structure under the top tag `<business:Patent DocumentAndRelated>`, there are four principle parts:

- `<business:BibliographicData>` containing the meta-data features of the current patent application including WIPOST3Code, DocNumber, Kind, Date, PublicAvailabilityDate, ClassificationIPCR, InventionTitle, ApplicantDetails, InventorDetails and AgentDetails.
- `<business:Abstract>` that contains the abstract of each patent.
- `<business:Description>` that contains the description of each patent, in which each paragraph is an individual `<base:Paragraphs>` item.
- `<business:Drawings>` that contains the description of attached figures in the application file, each figure is a `</base:Figure>` item.
- `<business:Claims>` is the claims of the current patent, which is the most important part of a patent and also is an essential part of the tech mining in our work. Like in the Table 2.2, each claim is a `<base:ClaimText>` item.

## 2.3 From XML files to less space-consuming raw texts in a unified format

The initial patent applications are stored in an XML format characterized by redundancies and intricate structures, often containing extraneous information. Recognizing the limitations of computer storage capacity and the necessity for streamlined subsequent processing, we will undertake a conversion to a format that is less resource-intensive. This transformation will involve the extraction of pertinent information, which will be organized in a format that is both machine-readable and comprehensible to humans. This restructured format is depicted in Figure 2.5 for reference.

We used the Python package *xml.etree.ElementTree* with which we extract the following information among all this information contained in the XML files:

1. title                      InventionTitle
2. date                        ClassificationIPCR
3. IPC class                 Date
4. abstract                  Abstract
5. description              Description
6. claims                     Claims

In this way, all patent XML files are converted to plain text with a unified format. Take the example of a patent application CN102020000987007.xml in Section 2.2, the goal is CN102020000987007.txt whose structure is shown in Table 2.2 (with English translation).

	Chinese	English translation <sup>61</sup>
<b>Title</b>	_____一种批量条码定位方法及识别系统	_____A batch barcode positioning method and recognition system
<b>Date</b>	_____20210105	_____20210105
<b>IPC class</b>	_____G06K_7/14: G06N_3/04: G06N 3/08	_____G06K_7/14: G06N_3/04: G06N 3/08
<b>Abstract</b>	本发明公开了一种批量条码定位方法及识别系统。本发明的批量条码定位方法,连续获取慢速运动场景中的图像,结合多运动目标跟踪方法实现批量条码的准确定位;本发明的批量条码识别系统使用低像素摄像头模块获取含有目标条码的连续多张低像素图像,并通过无线通信模块快速将其传输至	The present invention discloses a batch barcode positioning method and recognition system. The batch barcode positioning method of the present invention continuously acquires images in slow motion scenes and combines with multi-motion target tracking method to achieve accurate positioning of batch barcodes; the batch barcode recognition system of the present invention uses a low-pixel camera module to acquire multiple consecutive low-pixel images containing target barcodes and quickly transmits them to a positioning server through a wireless

<sup>61</sup> Automatically translated by <https://www.deepl.com/>.

	定位服务器;定位服务器利用低像素的图像,运行本发明的批量条码的定位方法获取相关坐标位置及解码所需旋转角度;本识别系统也同时获取高像素图像,通过高低像素图像的映射关系,进而快速获取高像素图像中批量条码的相关坐标位置;最后基于高像素图像和解码所需旋转角度快速完成批量条码解码,用户体验极大提高。	communication module; the positioning server uses the low-pixel images and runs The positioning server uses the low-pixel image and runs the positioning method of the batch barcode of the present invention to obtain the relevant coordinate position and the rotation angle required for decoding; the recognition system also obtains the high-pixel image at the same time, and then quickly obtains the relevant coordinate position of the batch barcode in the high-pixel image through the mapping relationship between the high and low-pixel images; finally, the decoding of the batch barcode is quickly completed based on the high-pixel image and the rotation angle required for decoding, and the user experience is greatly improved.
<b>Description</b>	_____d: 一种批量条码定位方法及识别系统 技术领域 ..... 背景技术 ..... 发明内容 ..... 附图说明 ..... 具体实施方式 ..... 实施例1 实施例2 .....	_____d: A Batch Barcode Positioning Method and Recognition System Technical Field ... Background Technology ... Content of the Invention ... Illustrated with drawings ... Specific embodiments ... Example 1 Example 2 ...
<b>Claims</b>	_____c: 1. 一种批量条码定位方法,其特征在于,包括如下步骤: ..... 2. .... 3. ....	_____c: 1. a batch barcode positioning method, characterised by comprising the steps of ..... 2. .... 3. ....

**Table 2.5 - One example of the results of the data processing**

Three primary categories of errors have been identified in the transformed texts: Unpaired opening symbols such as "{ " or "(", as illustrated in examples in Figure 2.6 (a); The presence of ", ." at the conclusion of sentences, as depicted in instances in Figure 2.6 (b); Superfluous spaces within established phrases or formulas, showcased in samples in Figure 2.6 (c).

- 从V个图像中任意选取两个图像v和w构成一组,针对关节点p,根据这一组图像获得一组距离值{d}。

Select any two images, v and w, from a set of V images to form a pair. With respect to the keypoint p, derive a set of distance values {d} based on this image pair.

**Figure 2.6 (a) - Examples of Unpaired Opening Symbols**

- 评价模块, 便于用户给出评价, 。

Evaluation module, facilitating users to provide feedback, .

- 3. 根据权利要求1所述的电池箱, 其特征在于, 。

3. The battery pack according to claim 1, characterized by, .

- 所述类别特色字典模型包括如下函数更新步骤, 。

The category-specific lexicon model includes the following function update steps, .

**Figure 2.6 (b) - Examples of Pounctuations**

- 在服务器终端可以时时接收从APP客户端上仓库客户B发送的信息, 并且根据信息进行分配至仓库A, 此时仓库A内的空间剩余为**Tb- B1**;

At the server terminal, real-time reception of information sent from Warehouse Customer B through the APP client is possible. Based on this information, allocation to Warehouse A is performed, with the remaining space in Warehouse A denoted as **Tb- B1**.

- **D= (x1,x2,...,xn)**
- 时间为**90 ~120**

The duration is between 90 and 120.

- 根据公式**Z2= aX + bY +C2**计算出第二加热丝的通电持续率;

Calculate the electrical activation rate of the second heating wire, Z2, using the formula **Z2 = aX + bY + C2**.

- 6. 根据权利要求 2 所述的一种随水位变化智能排水设备, 其特征在于:
- A water-level-responsive intelligent drainage device according to claim 2, characterized by:

**Figure 2.6 (c) - Examples of Superfluous Spaces**

- 所述NR-U UE仅在所述第一指示指示的时间窗(RAR)内检测PDCCH, 所述PDCCH由所述gNB采用相同的RA-RNTI进行加扰。

The NR-U UE only monitors the PDCCH within the time window (RAR) indicated by the first indication, and the PDCCH is scrambled by the gNB using the same RA-RNTI.

- 所述第一指示承载于RAR MAC CE中。

The first indication is carried within the RAR MAC CE.

- 维生素C 15-25份

Vitamin C 15-25 parts

- 所述主控模块为Arduino Mega 2560控制器;

The main control module is an Arduino Mega 2560 controller.

**Figure 2.6 (d) - Examples of Abbreviations**

The final class of errors presents a more intricate challenge, as distinguishing between instances that warrant the inclusion of spaces (as demonstrated in Fig 2.6 (d)) is occasionally difficult to achieve through automated procedures. To address this complexity and aim for optimal rectification, we employed regular expressions to enforce adherence to specific matching rules.

Additionally, we applied the following normalization rules:

- Eliminating spaces at the ends of equal, plus, and minus signs.
- Ensuring “~” is devoid of spaces at both ends if the adjacent characters are numeric<sup>62</sup>.
- Similar to the above, spaces are omitted at both ends of the hyphen character if the adjacent characters on either side are numeric or alphabetic characters.

The objective of compiling a collection of specific Chinese patent expressions in Section 2.5, particularly those prevalent in patent claims, is to streamline word segmentation during the creation of the treebank discussed in the subsequent chapter, as well as to facilitate subsequent syntactic analyses. This standardized approach aims to ensure uniform separation and annotation practices for these specific character sequences across various segmenters and parsers. This harmonized methodology allows for direct performance comparison among different segmenters and parsers.

<b>Year</b>	<b>Size</b>	<b>Number</b>
2017	2.6G	137,092
2018	13G	678,181
2019	29G	1,151,398
2020	21G	859,640
2021	4.9G	237,338
TOTAL	71G	3,063,649

**Table 2.4 - The Sizes and Numbers of the Collected Datasets over Months**

Table 2.4 presents the final results of the collection of the patent corpus after extraction. In the end, we got 3,063,649 patent applications in total, transformed in raw text, 71G in size

---

<sup>62</sup> The errors falling within the this category stem from intricacies encountered during the XML parsing process.

## 2.4 Classification into IPC domains

As elucidated in Section 1.1, the International Patent Classification (IPC), introduced through the Strasbourg Agreement of 1971, institutes a hierarchical framework of symbols that transcends linguistic barriers, enabling the systematic categorization of patents and utility models according to their respective technological domains. Annually, on January 1, a fresh iteration of the IPC comes into effect (in this study, we employ version 2006.01.01, as it is the sole officially translated version available for download from the website of the China Intellectual Property Office at the beginning of the study, despite its datedness).

Various editions of the IPC adhere to a consistent classification structure, underpinned by the subdivision of eight distinct domains. These eight technical domains within the IPC (ranging from Section A to Section H) are delineated below, along with their corresponding titles:

- A. HUMAN NECESSITIES
- B. PERFORMING OPERATIONS; TRANSPORTING
- C. CHEMISTRY; METALLURGY
- D. TEXTILES; PAPER
- E. FIXED CONSTRUCTIONS
- F. MECHANICAL ENGINEERING; LIGHTING; HEATING; WEAPONS; BLASTING
- G. PHYSICS
- H. ELECTRICITY

Each of these domains (or sections) is composed of a number of classes, subclasses, groups and subgroups organized in a hierarchical structure.

Below there is one example of the classification hierarchy of “G06N\_3/021”:

- **Section G** => **PHYSICS**
  - **Class 06** => **COMPUTING; CALCULATING; COUNTING**
    - **Subclass N** => **COMPUTING ARRANGEMENTS BASED ON SPECIFIC COMPUTATIONAL MODELS**
      - **Group 3/00** => **Computing arrangements based on biological models**
        - **Subgroup 021** => **using neural network models**

Considering the fact that a great number of patents belong to more than one IPC classes, such as in the example below (Table 2.7), we simply chose the most frequent<sup>63</sup> IPC class for each of them.

---

<sup>63</sup> If two or more IPC domains have the same highest count, the first one is chosen.

<b>Title</b>	<b>IPC classes</b>	<b>Domain count</b>	<b>Final decision</b>
CN102020000987007	G06K_7/14 G06N_3/04 G06N 3/08	G: 3	<b>G</b>
CN102016000027125	C12N_1/04 A23B_7/155 C12R_1/84	A: 1 C: 2	<b>C</b>
CN102017000515545	B62M_6/90 B62J_99/00 F03D_9/11 F03D_9/32	B: 2 F: 2	<b>B</b>

**Table 2.7 - Example of Classification of Patent Applications**

	<b>IPC domain</b>	<b>number</b>
<b>A</b>	HUMAN NECESSITIES	817,452
<b>B</b>	PERFORMING OPERATIONS; TRANSPORTING	1,836,694
<b>C</b>	CHEMISTRY; METALLURGY	1,471,134
<b>D</b>	TEXTILES; PAPER	189,144
<b>E</b>	FIXED CONSTRUCTIONS	345,459
<b>F</b>	MECHANICAL ENGINEERING; LIGHTING; HEATING; WEAPONS; BLASTING	723,481
<b>G</b>	PHYSICS	1,864,221
<b>H</b>	ELECTRICITY	1,457,240

**Table 2.8 - The Number of Patents of Each IPC Domain**

After the process of classification, the final results on the number of patents by IPC section is shown in Table 2.8. The subsequent Figures 2.9 (a-e) depict the count of patent applications within each IPC section categorized by year.



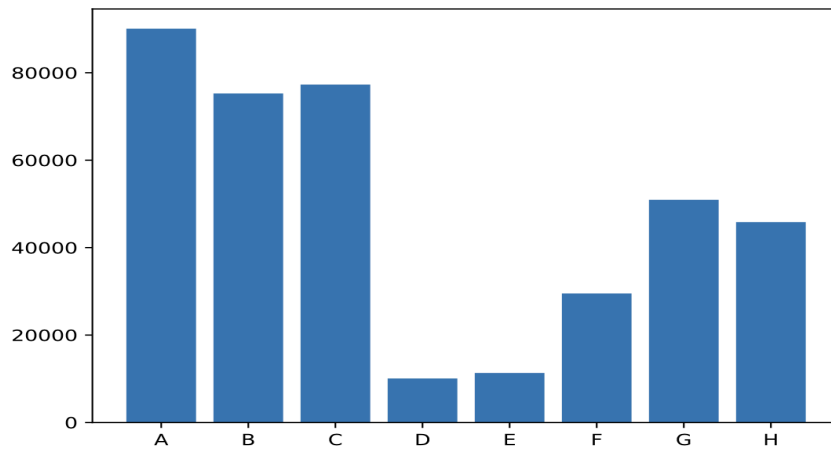


Figure 2.9 (a) - 2017

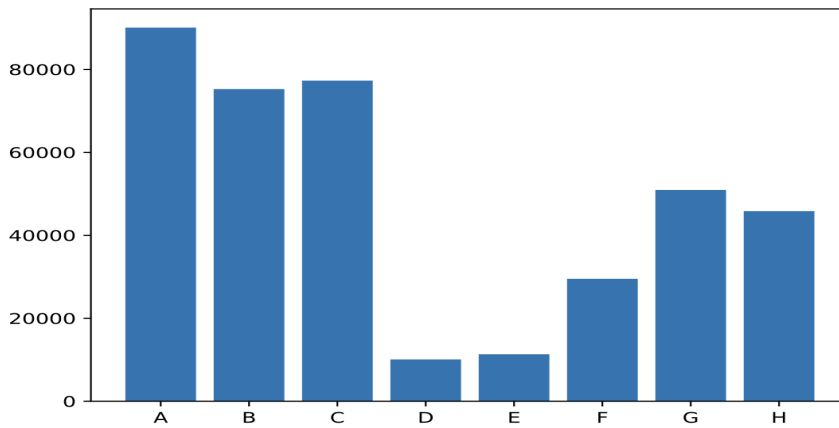


Figure 2.9 (b) - 2018

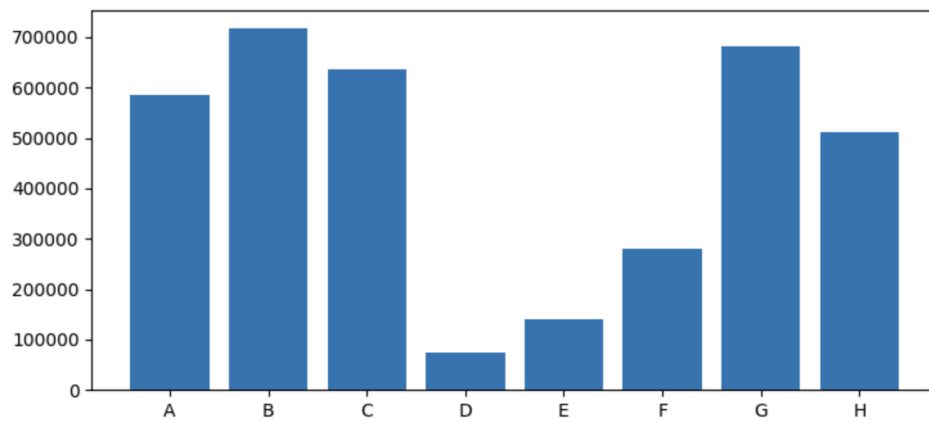


Figure 2.9 (c) - 2019

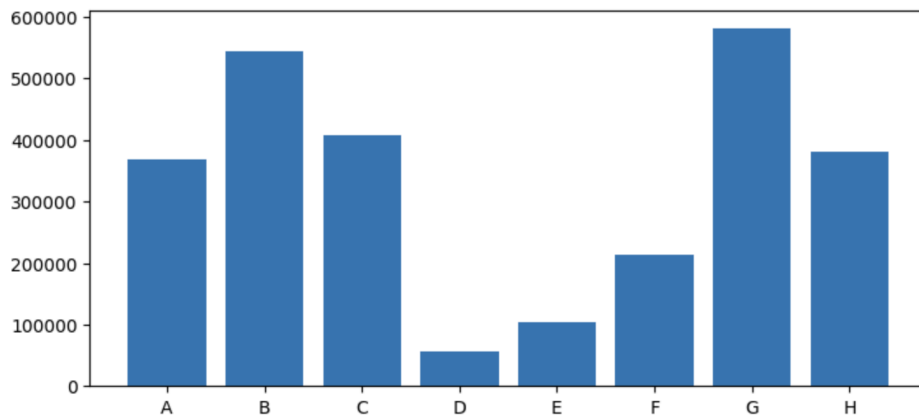


Figure 2.9 (d) - 2020

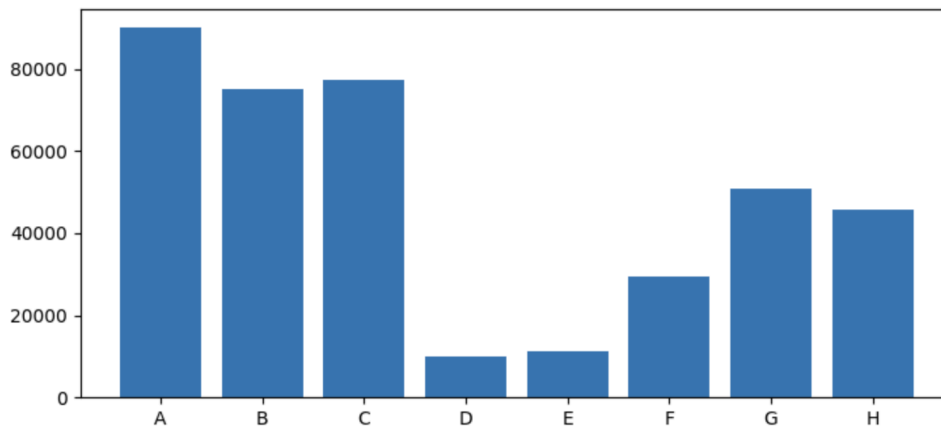


Figure 2.9 (e) - 2021

An evident observation emerges, highlighting that Section D - TEXTILES; PAPER and Section E - FIXED CONSTRUCTIONS comprise notably fewer applications annually in the corpus. Another notable inference is the significant growth in the count of applications within Section G - PHYSICS, particularly since 2019, possibly attributed to the advancement in Artificial Intelligence.

## 2.5 The Writing Style and Linguistic Specificities of Chinese Patents

The requirements for drafting Chinese patent applications are extensively stipulated in the Chinese Patent Law and its implementing regulations.

Another significant document related to the preparation of patent applications is the “Patent Examination Guidelines (2021) (《专利审查指南》)” published by CNIPA (the National Intellectual Property Administration of the People's Republic of China). Within these regulations, specific standards are set for the abstract, description, and claims sections of the patent application.

For the abstract:

*“The abstract text portion should include the title of the invention and the relevant technical field, providing a clear reflection of the technical problem to be solved, the key points of the technical solution to the problem, and the main purposes of the invention. In cases where the invention title is not provided or the key points of the technical solution are not adequately reflected, the applicant should be notified to make corrections. If commercial promotional language is used, the applicant can be notified to remove it, or the examiner may remove it; in cases where the examiner removes it, the applicant should be notified.”*

*The abstract text portion should not use a title, and the textual content (including punctuation) should not exceed 300 words. If the abstract exceeds 300 words, the applicant can be notified to condense it, or the examiner may condense it; in cases where the examiner condenses it, the applicant should be notified.”*

For the description:

*“Article 26, paragraph 3 of the Patent Law, along with Article 17 of the Implementing Regulations of the Patent Law, provide regulations concerning the substantive content and writing style of the specification.*

*The first line of the first page of the specification should contain the invention's title, which must be consistent with the title in the request form and should be centered both horizontally and vertically. The term “Invention Title” or “Title” should not be prefixed to the invention title. There should be a blank line between the invention title and the main body of the specification.*

*The format of the specification should include the following sections, each preceded by a clear title:*

- *Technical Field*
- *Background Art*

- *Summary of the Invention*
- *Brief Description of the Drawings*
- *Detailed Description of the Invention*

*The specification text may include chemical formulas, mathematical equations, or tables, but should not include figures. If the specification text includes a description of the drawings, there should be corresponding drawings included. If there are drawings in the specification, the specification text should provide a detailed explanation of the drawings.”*

And for the Claims:

*“The claims shall be based on the description, clearly and concisely defining the scope of protection sought for the claimed invention or utility model. The claims shall set forth the technical features of the invention or utility model, which can be constituent elements of the technical solution of the invention or utility model or the interrelationships between these elements.*

*Article 26, Paragraph 4 of the Patent Law and Articles 19 to 22 of the Implementing Regulations of the Patent Law provide provisions on the content and drafting of the claims.*

*In one set of claims, there should be at least one independent claim, and it can also include dependent claims.*

*If there are multiple claims in the claims, they shall be numbered in consecutive Arabic numerals, and the numbering shall not be preceded by words such as “claim” or “item.” Chemical or mathematical formulas can be included in the claims, and tables can also be used when necessary, but illustrations are not allowed.*

*The claims shall be numbered consecutively in Arabic numerals.*

*In patent claims, terms with uncertain meanings, such as ‘thick’, ‘thin’, ‘strong’, ‘weak’, ‘high temperature’, ‘high pressure’, ‘a wide range’, etc., should not be used unless such terms have a recognized precise meaning in a specific technical field, such as ‘high frequency’ in amplifiers. For terms without recognized meanings, if possible, more accurate wording from the specification should be selected to replace the aforementioned uncertain terms.*

*Expressions like “for example”, “preferably”, “especially”, “if necessary”, etc., should not appear in the claims. Such expressions can result in different scopes of protection within a single claim, leading to unclear protection. When an upper-level concept appears in a claim followed by a subordinate concept introduced by the above expressions, the applicant should be required to modify the claim. They may be allowed to retain one of them in that claim or limit them separately in two different claims.*

*In general, terms like ‘about’, ‘approximately’, ‘similar to’, ‘or the like’, and similar expressions should not be used in claims, as they often lead to unclear claim scope.*

*However, if such terms appear in a claim, the examiner should assess whether their use leads to lack of clarity based on the specific circumstances. If not, they may be allowed.*

*Except for parentheses used for figure numbering or chemical and mathematical formulas, the use of parentheses in claims should be minimized to avoid unclear claim language, for example, '(concrete) molded brick'. Nevertheless, parentheses with generally acceptable meanings are permissible, such as '(methyl) acrylate', '(comprising 10% to 60% by weight) of A.'*

Below is an example of a patent claim sentence in Chinese (with English translation):

**Example:**

(CN102218144A)

“根据权利要求16至18中任一项所述的方法，其特征在于，所述疾病选自动脉粥样硬化，血管损伤导致的肥胖、高血糖和慢性炎症中的一种或多种。”

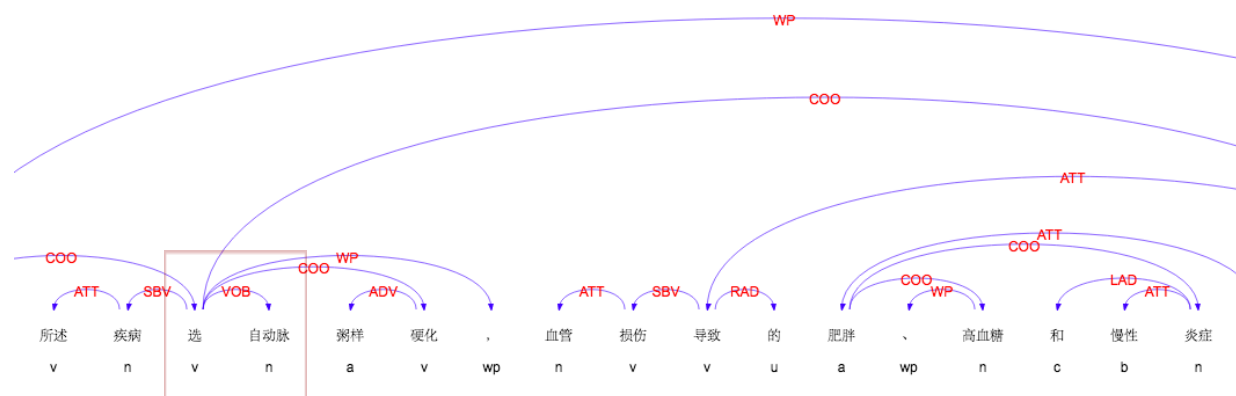
**Translation:**

*“The method according to any one of claims 16 to 18, wherein the disease is selected from one or more of atherosclerosis, obesity due to vascular injury, hyperglycemia, and chronic inflammation.”*

**Figure 2.7 - The Example from Patent CN102218144A with its English Translation by Google**

As elucidated below, the patent text, functioning as a textual representation at the forefront of emerging technologies, employs a substantial volume of specialized terminology that is either uncommon or unfamiliar to automated segmenters, posing considerable challenges for automated processing. This is particularly pronounced in languages that adhere to a “scriptua continua”, lacking word-separating spaces, as exemplified by Chinese. Moreover, characterized by an intricately dense structure aimed at minimizing ambiguity, patent texts adopt a writing style that may appear unfamiliar, replete with distinctive “legalese” expressions like substituting “said” for the definite determiner “the”. Some of these expressions have been transposed to Chinese patents to mirror the style of Western patents. Furthermore, the patent texts exhibit intricate logical connections and notably lengthy sentences, as each claim is obligated to maintain the form of a singular sentence. Collectively, these attributes render patents remarkably intricate to analyze – a challenge applicable both to human readers and existing parsing technologies.

One specificity of claim sentences is the complexity of the syntactic structure. Below is a presentation of the syntactic analysis of the example patent text taken from parts of a claim sentence in Chinese:



**Figure 2.8 - The Claim Sentence Parsed by the LTP Dependency Parser**

This long sentence was passed to the LTP parser<sup>64</sup> of the Harbin Institute of Technology. The comparatively poor quality of parsing is notable in the result, where even common fixed expressions like 权利要求 (quán lì yāo qiú, right request, ‘claim’) have been divided into two parts. Problems also come from out-of-vocabulary (OOV) terminology such as 动脉粥样硬化 (dòngmài zhōuyàngyìng huà, “atherosclerosis”), which degrades the accuracy of the segmentation. The correct segmentation including the preceding verb “选自 (xuǎn zì, ‘choose from’) / 动脉粥样硬化 (dòngmài zhōuyàngyìng huà, ‘atherosclerosis’) ” becomes “选 / 自动脉 / 粥样 / 硬化 (xuǎn/ zì dòngmài/ zhōu yàng/ yìng huà, “chosse/ automatic artery/ sclerosis)”<sup>65</sup>; the dependency analysis is not only built on these wrong tokens (in red retangle) but even has difficulties with the analysis of expressions such as “中所述的 (zhōng suǒ shù de, “as described in”)” annotated as one term, which are frequent in patent claims.

On the technical side, the lack of annotated training corpora limits the straightforward application of Natural Language Processing tools as well as the adaptation of Machine Learning methods. The Cloem.com<sup>66</sup> project has gained some experience in solving these difficulties for English: Interception and rule-based decomposition of long sentences; completion of elliptical claim articles; replacement of certain reserved terms of patent texts in order to facilitate the training process without large annotated out-of-domain corpora.

conveying characteristics	delineating composition	detailing methods
特点在于; 其特征; 其特点; 其特性; 特征是; 特征为; 特征	其复合组分; 以下复合成分; 各复合成分; 主要由; 是由; 包有; 包括; 包含; 含有; 其具有; 组	过程为; 如下过程; 如下工艺; 以下过程; 以下工艺; 以下方法; 工艺步骤为; 工艺为; 步骤; 如下方法; 生产方法; 方法是; 方法为; 方法; 制备工艺

<sup>64</sup> <https://www.ltp-cloud.com/demo/>

<sup>65</sup> The granularity of segmentation is another problem that is dependent on the following tasks.

<sup>66</sup> <https://en.wikipedia.org/wiki/Cloem?oldformat=true>

	成为; 所组成; 组成是; 组成; 其组成; 处方是; 组分为; 其组分; 各组分; 以 下成分; 以下组分; 成分为; 成分; 如下配方; 配方是; 配方为	
is characterized by;	is composed of; include; comprise; contain; possess; consist of; are constitute;	the following process; the following method; steps;

**Table 2.10 - Keywords of the Three Categories of Specific Expressions**

Despite the often intricate syntactic structures and lengths observed in patent claims, the inherent claim structures themselves display a certain degree of simplicity, allowing for their categorization into discernible patterns. Researchers such as Zhai Dongsheng et al. (2011) arrived at a summarization of common claim-specific expressions. They organized these expressions into three distinct categories: those conveying characteristics, those delineating composition, and those detailing methods (Table 2.10).

To examine the other distinctive attributes of Chinese patent claim sentences, we conducted a quantitative analysis on a subset of a group of claim sentences chosen randomly from each IPC domain. For each randomly selected application document, we retained its initial claim and, if present, an additional claim. The claim sentences, chosen for practical reasons, fall within the range of 10 to 100 characters for analysis.

	Number	Length	Modifier “DE”
<b>A</b>	1002	39.42	1.07
<b>B</b>	1004	43.30	1.24
<b>C</b>	1001	38.87	1.10
<b>D</b>	252	46.67	1.32
<b>E</b>	501	42.43	1.22
<b>F</b>	1011	43.35	1.38
<b>G</b>	1019	42.33	1.41
<b>H</b>	1001	42.60	1.35
<b>UD_Chinese-HK</b>	1004	14.01 <sup>67</sup>	0.32
		9.83 <sup>68</sup>	

<sup>67</sup> For characters.

<sup>68</sup> For segmented words.

**Table 2.11 - The sentence number, the average length of claim sentences by character (except UD\_Chinese-HK Treebank<sup>69</sup>), and the average number of the modifier mark “DE<sup>70</sup>” of each sentence.**

In this work, we summarize the frequent structures in claims, the results as examples are shown below.

- (1) 1.一种抗衰老护肤品, 其特征在于, 由以下重量份数的组分制成:  
 1.Yī zhǒng kàng shuāi lǎo hùfū pǐn, qí tèzhēng zàiyú, yóu yǐxià zhòngliàng shù de zǔfèn zhìchéng:  
 1. A type of anti-aging skincare product, characterized by being made up of components in the following weight proportions:
- (2) 1.一种淡水砂梨种质离体保存与恢复方法, 其特征在于, 包括如下步骤:  
 :  
 Yī zhǒng dànshuǐ shā lí zhǒngzhì lí tǐ bǎocún yǔ huīfù fāngfǎ, qí tèzhēng zàiyú, bāokuò rúxià bùzhòu:  
 A method for ex situ preservation and recovery of germplasm of freshwater sand pear, characterized by including the following steps:
- (3) 1.一种妊娠母猪预混合饲料, 按重量份计算, 包括玉米、豆粕, 鱼粉及多种维生素, 其特征是, 玉米650-670克、豆粕150-160、鱼粉150-170、维生素A 25-30万IU、维生素D。  
 Yī zhǒng rěnnshēn mǔzhū yùhé fèi hézhiliáo, àn zhòng fèn jisuàn, bāokuò yùmǐ, dòubǔ, yúfěn jí duō zhǒng wéishēngsù, qí tèzhēng shì, yùmǐ 650-670 gè, dòubǔ 150-160 gè, yúfěn 150-170 gè, wéishēngsù A 25-30 wàn IU, wéishēngsù D.  
 A premixed feed for pregnant sows, calculated in weight portions, comprising corn, soybean meal, fish meal, and various vitamins. The features are as follows: corn 650-670 grams, soybean meal 150-160 grams, fish meal 150-170 grams, vitamin A 250,000-300,000 IU, and vitamin D.
- (4) 2.根据权利要求1所述的抗衰老护肤品, 其特征在于, 由以下重量份数的组分制成:  
 Gēnjù quánlǐ yāoqiú yī suǒ shù de kàng shuāi lǎo hùfū pǐn, qí tèzhēng zàiyú, yóu xià liàng wù fènshù de zǔfèn zhìchéng:  
 2. The anti-aging skincare product according to claim 1, characterized by being composed of components in the following weight proportions:
- (5) 3.根据权利要求1所述的番茄红素微囊粉的工艺生产方法, 其特征在于:

<sup>69</sup> A Traditional Chinese treebank of film subtitles and of legislative proceedings of Hong Kong, parallel with the Cantonese-HK treebank. ([https://universaldependencies.org/treebanks/zh\\_hk/index.html](https://universaldependencies.org/treebanks/zh_hk/index.html))

<sup>70</sup> This will be explained in Section 3.1.1.



3. Gēnjù quánlì yāoqiú yī suǒ shù de fān qié hóngsù wēi wán fēn de gōngyì shēngchǎn fāngfǎ, qí tèzhēng zài yú:

3. The method of producing tomato lycopene microcapsules as described in claim 1, characterized by:

As shown in the examples provided above, we have identified five frequent structures in patent claims. Each distinct structure is indicated by a unique colour code. The identified structures, along with their corresponding frequencies within the corpus of the randomly selected sentences described above, are detailed in Table 2.12. These particular structures will be revisited and subject to syntactic analysis in Section 3.3.1 as part of the manual annotation procedure.

Original Chinese	English Translation	Number
其特征在於; 其特征是	characterized by	1,978
一种 ... 产品/方法	A ... product/method	788
包括 ...	comprising	1,265
包括 ... 步骤	including ... steps	139
由 ... 组分制成; 由 ... 组成	being composed of ...	60
根据权利要求1所述的 ... 产品/方法	product/method according to claim 1/as described in claim 1	871

**Table 2.12 - The claim-specific structures with translation and their total frequency in the sample.**

Another distinctive characteristic of claim sentences is the notably higher frequency of technical terminology usage in comparison to general corpora, such as news articles. Moreover, within a single patent, technical terms are recurrently employed, as exemplified below.

CN102019000230488

1. 一种集打孔倒角一体的钻头, 其特征在于: 包括依序连接的对接座(1)、平衡台(2)、钻头基体(3)和刀片(5); 所述对接座(1)、平衡台(2)与钻头基体(3)内部均设有冷却液通道(8); 所述平衡台(2)上设有出水口(6), 所述出水口(6)与冷却液通道(8)连通; 所述钻头基体(3)上设有冷却孔(7), 所述冷却孔(7)与冷却液通道(8)连通; 所述钻头基体(3)的侧面设有一个倒角头(4); 所述倒角头(4)凸出于钻头基体(3)表面, 所述倒角头(4)的截面为直角梯形, 直角梯形的锐角指向刀头(5)。

1. A drill bit with integrated hole punching and chamfering features, characterized by: comprising sequentially connected docking seat (1), balancing platform (2), drill bit body (3), and blade (5); internal cooling fluid channels (8) are provided within the docking seat (1), balancing platform (2), and drill bit body (3); a water outlet (6) is situated on the balancing platform (2), and the water outlet (6) is connected to the cooling fluid channel (8); cooling holes (7) are arranged on the drill bit body (3), and the cooling holes (7) are connected to the cooling fluid channel (8); a chamfering head (4) is located on the side of the drill bit body (3); the chamfering head (4) protrudes from the surface of the drill bit body (3), and the

cross-section of the chamfering head (4) is that of a right-angled trapezoid, with the acute angle of the right-angled trapezoid directed towards the blade (5).

2. 根据权利要求1所述的集打孔倒角一体的钻头, 其特征在于: 所述对接座(1)上设有外螺纹, 所述平衡台(2)上设有一个带有内螺纹的固定孔, 所述对接座(1)与平衡台(2)通过螺纹连接。

2. According to claim 1, an integrated hole punching and chamfering drill bit is disclosed, wherein an external thread is featured on the docking seat (1), a fixed hole with internal threads is situated on the balancing platform (2), and the threaded connection between the docking seat (1) and the balancing platform (2) is established.

3. 根据权利要求1所述的集打孔倒角一体的钻头, 其特征在于: 所述对接座(1)和平衡台(2)内均设有上下贯通的中空, 所述钻头基体(3)内设有筒状的中空; 所述中空连接在一起构成冷却液通道(8)。

3. In accordance with claim 1, the integrated hole punching and chamfering drill bit as described, characterized by the presence of vertically connected hollow structures within both the docking seat (1) and the balancing platform (2), and the incorporation of a cylindrical hollow structure within the drill bit base (3); these interconnected hollow structures collectively establish a cooling liquid passage (8).

4. 根据权利要求1所述的集打孔倒角一体的钻头, 其特征在于: 所述平衡台(2)内设有若干个由平横台(2)轴心向四周逐渐向下倾斜的冷却液支流通道(81), 所述冷却液支流通道(81)一端与冷却液通道(8)连通, 另一端通向平衡台表面的出水口(7)。

4. In accordance with claim 1, the drill bit with integrated hole punching and chamfering features, characterized by: within the balancing platform (2), a number of cooling fluid sub-channels (81) are established, gradually tilting downward from the central axis of the horizontal platform (2) towards the periphery; one end of the cooling fluid sub-channels (81) is connected to the cooling fluid channel (8), while the other end leads to an outlet (7) on the surface of the balancing platform.

5. 根据权利要求1所述的集打孔倒角一体的钻头, 其特征在于: 所述钻头基体(3)为圆柱状, 其底部设有一个与刀头(5)适配的缺口(31), 所述刀头(5)插入缺口(31)处, 通过螺丝(32)贯穿缺口(31)和刀头(5), 实现刀头(5)紧固。

5. In accordance with claim 1, the drill bit with integrated hole punching and chamfering features, characterized by: the drill bit base (3) is cylindrical and equipped with a notch (31) that matches the blade (5) at its bottom; the blade (5) is inserted into the notch (31), and fastened securely through the insertion of a screw (32) that passes through both the notch (31) and the blade (5).

6. 根据权利要求1所述的集打孔倒角一体的钻头, 其特征在于: 所述出水口(5)环绕钻头基体一圈; 所述钻头基体(3)内部设有若干个连接通道, 所述连接通道连通冷却液通道(8)与出水口(7)。

6. In accordance with claim 1, the drill bit with integrated hole punching and chamfering features, characterized by: the outlet (5) surrounds the drill bit base in a circle; the drill bit base (3) is internally equipped with several connecting passages, which establish communication between the cooling fluid channel (8) and the outlet (7).

Figure 2.13 - The Claims of Patent CN102019000230488

In a word, the characteristics of patent claim sentences encompass the use of straightforward sentence structures that are repetitively employed, the construction of lengthy sentences with intricate grammatical arrangements, the incorporation of a diverse and frequently updated repertoire of specialized vocabulary, the utilization of extended and fixed word groups with abundant modifying components, reflecting the robust combinatory nature intrinsic to technical attributes. Additionally, these sentences stipulate the need for precision in hierarchical relationships, both in terms of superordinate and subordinate concepts, while also acknowledging the recurring usage of specific terms within the same document. Consequently, this complex linguistic landscape has led to the emergence of a requirement for character-level dependency analysis.

## Chapter 3 - Syntactic Analysis of Chinese Patent Claims on Character-level

In this chapter, the emphasis is on analyzing the syntax of Chinese patent claim texts.

In the prevailing Chinese morphological theory, which is widely accepted and expounded by scholars such as Feng (1997), Zhang (2003), and Dong (2011), compounds in Chinese are categorized into two main groups: derivative words and compound words. The latter group comprises five subtypes, including modifier-head, coordinative, predicate-object, predicate-complement, and subject-predicate structures. Notably, He et al. (2012) and Chi et al. (2019) have highlighted a parallelism between the structure of compound words and syntactic structures in Chinese. This parallelism forms the basis for exploring the development of a new dependency schema that harmonizes character-level relationships with existing word-level relations.

To achieve this objective, it is imperative to incorporate these new inter-character relations into a dependency tree that aligns with established distributional criteria. Consequently, we have chosen to build upon the Surface-Syntactic Universal Dependencies (SUD), a variant of the Universal Dependencies (UD) framework presented by Gerdes et al. in 2018. SUD offers an alternative schema to UD, emphasizing surface-syntactic features while preserving a dependency structure rooted in word distribution that places importance on functional heads. Our approach involves applying syntactic tests designed to identify the head and internal structure of compound terms based on distributional patterns.

This chapter unfolds as follows: Section 3.1 introduces an innovative character-level annotation schema tailored specifically for Chinese patent claims. Section 3.2 delves into the creation and automated pre-annotation processing of Character-level Chinese Patent Treebanks. Section 3.3 offers insight into manual corrections conducted in alignment with the character-level annotation schema. In Section 3.4, we provide an overview of the annotated treebank, including pertinent statistics, and elucidate the methodology employed for enhancing the parser through bootstrapping and converting the character-level SUD treebank into a conventional UD word-level treebank.

### 3.1 A New Character-level Annotation Schema of Morphosyntactic Relations in Chinese

In this section, we propose a novel character-level annotation schema for the Chinese treebank. We begin by introducing both existing distributional and non-distributional tests used for parts of speech and each term's internal relation. Subsequently, we define new term internal labels specifically designed for the Surface Syntactic Universal Dependencies (SUD) framework. We then illustrate each type of technical term with concrete examples and adapt the tests to these real-world instances (detailed in Section 3.1.1).

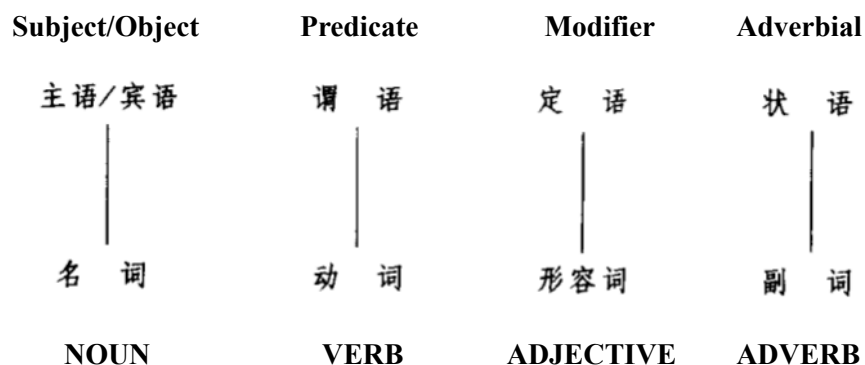
To provide a comprehensive view, we present all the newly created labels alongside the existing SUD syntactic labels within a hierarchical structure, elucidated in Section 3.1.2.

Finally, we offer a complete decision tree that serves as term-internal annotation guidelines, aiding in the consistent application of our proposed schema, as outlined in Section 3.1.3.

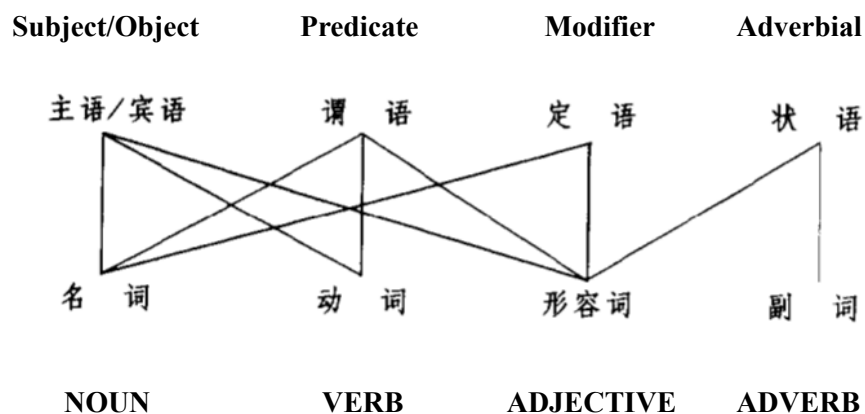
#### 3.1.1 Possible Distributional and Semantic Tests for Chinese Character Part-of-Speech and inter-characters Relations

##### 3.1.1.1 Tests for Part-of-Speech of the Characters

The issue surrounding the basis of Part-of-Speech (POS) tagging, namely whether it should rely on semantic meaning or syntactic distribution, has long been a prominent topic of inquiry in linguistic research (Xia, 2000b). This matter holds particular significance when dealing with the Chinese language, where the majority of characters exhibit multiple potential parts-of-speech tags (Figure 3.1 (a)) and lack the natural delimiters or inflectional markers commonly present in languages employing the Latin script (Figure 3.1 (b)) (Magistry et al., 2012). Consequently, the distinction between different part-of-speech tags in Chinese is primarily indicated by their distributional positions, with semantic considerations taking a secondary role. Thus, our approach to POS tagging, both at the word-level and character-level annotation, prioritizes distributional position over semantic factors.



**Figure 3.1 (b) - The Correspondence between Grammatical Functions and Lexical Classes in Indo-European languages (Lu Jianming, 2003)**



**Figure 3.1 (b) - The Correspondence between Grammatical Functions and Lexical Classes in Chinese (Lu Jianming, 2003)**

In their 2007 work on the modern Chinese language, Huang Borong and Liao Xudong provided a comprehensive overview of term formation classes that we have presented in Section 1.2.1. In their work, they also outline the three most frequently used criteria for determining a term's POS:

1. Form Change: This criterion examines whether a word undergoes changes in form, such as inflectional markers in European languages.
2. Meaning: This approach takes into account the semantic meaning of a word when classifying it. However, it's worth noting that this method can be ambiguous and challenging to quantify, as semantic meanings can evolve over time.
3. Grammatical Function in a sentence: This criterion focuses on a word's role and function within a sentence. According to Huang and Liao, this criterion is particularly applicable to Chinese.

They also pointed out that a term's grammatical function can be attributed to two principal factors:

1. Syntactic component: A term's ability to function as a specific syntactic component within a sentence.
2. Combination with other terms: A term's capacity to combine with other terms to form compounds.

A substantial portion of our tests is founded on these two factors.

In this work, the annotation of parts of speech is divided into two parts: UPOS (Universal Part-of-Speech) for single characters and ExtPos<sup>71</sup> (External Part-of-Speech') as features for multi-character units connected with a "@m" relation. The ExtPos is put on the head character of the character cluster connected by "@m" relations, and can be used as the POS when the syntactic analysis

<sup>71</sup> The ExtPos feature was introduced to facilitate the annotation of idioms, titles, and other multi-word units which behave like a certain part of speech, even though none of their constituents necessarily carry that part of speech. (from SUD guidelines: <https://surfacesyntacticud.github.io/guidelines/u/extpos/>)

is moved to the word level, as it is done for the UD version of our treebank, see Gerdes et al. 2019 for details.

Both Huang and Liao (2007) and Feng (1989) have presented a set of criteria for each lexical category, with particular emphasis on four primary open-class categories: nouns, verbs, adjectives, and adverbs. In our summary, the part-of-speech can be tested by the following rules:

### Noun

- a. Nouns are the bedrock of any language, representing entities, objects, or concepts. In Chinese, nouns can function as both subjects and objects in sentences;
- b. Additionally, noun stacking or repetition is generally avoided;
- c. They can be modified by quantity phrases but are typically not accompanied by adverbs like “不” (not);
- d. Nouns can be further modified by demonstratives such as “这种” (this kind), “这个” (this), or “一个” (one);
- e. It's important to note that Chinese nouns do not take the aspect markers “了”, “着”, or “过”.

### Verb

- a. Verbs serve as the action words in Chinese sentences and are commonly used as predicates;
- b. Many verbs can be followed by dynamic auxiliary markers such as “着”, “了”, and “过”;
- c. Some verbs, especially those describing continuous actions, can be reduplicated, but this is not a universal rule;
- d. They often take objects and can be modified by adverbs, including the negating adverb “不” (not);
- e. However, certain verbs, particularly those expressing psychological states or volition, may resist modification by adverbs denoting degree, like “很” (very);
- f. Chinese verbs do not take demonstratives like “这种” (this kind), “这个” (this), or “一个” (one) as modifiers;
- g. Most verbs can occur in V - “不 (bù, not)” - V (Verb or not Verb), such as “吃不吃 (chī bù chī, eat or not eat)”.

### Adjective

- a. Adjectives are words that describe qualities or attributes of nouns; They can function as predicates, noun modifiers, or even verb complements;

- b. Most Chinese adjectives directly modify nouns and often require reduplication or the addition of “地” (de) to form adverbial phrases, “轻轻地唱 (qīng qīng de chàng, ‘softly sing’);
- c. Adjectives do not take objects, although some adjectives can also function as verbs in specific contexts, and in those cases, they can be modified by degree adverbs like “很” (hěn, ‘very’);
- d. Some adjectives that denote inherent qualities can be reduplicated, but they should not be modified by “很” (hěn, ‘very’).

#### Adverb

- a. Adverbs are words that provide additional information about actions, adjectives, or other adverbs; In Chinese, adverbs primarily serve as modifiers for verbs or verb phrases and can convey details about manner, time, or frequency;
- b. Importantly, adverbs do not modify nouns in the same way adjectives do.

Throughout the annotation process, one commonly employed technique for determining the POS of a character involves testing its compatibility with a functional character specifically designated for a certain part-of-speech (further details are provided in Table 3.1). By examining the combinatorial possibilities, we can gain valuable insights into the appropriate POS categorization.

POS	functional characters tests	quantity phrases and demonstratives	modified by “不” (not)	modified by” “很” (not)	duplication
<b>NOUN</b>	After modifier particle 的 DE	Yes ✓	No ✗	No ✗	No ✗
<b>VERB</b>	Before aspect markers 了 (le) / 着 (zhe) / 过 (guo)	No ✗	Yes ✓	Some * (continuous actions)	Some * (psychological states or volition)
<b>ADJ</b>	Before modifier particle 的 DE After modifier particle 得 DE	No ✗	Yes ✓	Yes ✓	Some * (can't be modified by” “很”)

ADV	Before modifier particle 地 DE	No ✘	No ✘	No ✘	No ✘
-----	----------------------------------	------	------	------	------

Table 3.1 - Tests for Part-of-Speech<sup>72</sup>

These rules can be applied not only to multi-character units but also to single characters.

For character and word meanings and grammatical functions, given that certain original usages have ceased to be applicable in modern Chinese, and notwithstanding formal assessments, we turn to the following two sources:

- Han Dian (汉典): <https://www.zdic.net/>

"Han Dian (汉典, zdic.net)" is a free online Chinese character and word search website, founded in 2004. The mission of Han Dian is to introduce Chinese culture, history, and language. It aims to provide explanations and services for those interested in Chinese language learning and research. Additionally, Han Dian explores the norms and standards of Chinese language and writing usage.

Han Dian's database includes 93,898 Chinese characters, 361,998 words, phrases, and expressions, as well as explanations for 32,868 idiomatic expressions. The collection of classical texts in Han Dian comprises a total of 1,055 classical literary works spanning 38,529 chapters and 203 classical essays. Furthermore, the platform features 268,886 classical poems and collects 135,804 notable Chinese calligraphers' works in Chinese calligraphy.



<sup>72</sup> If none of these POS tests can be applied, the term will be annotated as X.



<p>驟 zhòu ㄓㄡˋ</p> <p>〈动〉</p> <p>(1) (形声。从马，聚声。本义:马奔驰)</p> <p>(2) 同本义 [horse trots]</p> <p>驟，马疾步也。——《说文》</p> <p>车驟徒趋。——《周礼·大司马》</p> <p>步及驟处兮。——《楚辞·招魂》。注:“走也。”</p> <p>车驱而驹(驟)。——《礼记·曲礼》</p> <p>驾彼四骆，载驟馵馵。——《诗·小雅·四牡》</p> <p>(3) 又如:驟馵(疾驰的骏马)</p> <p>(4) 使马奔驰 [whip a horse on]</p> <p>遇春驟马追到，便活擒于马上。——《英烈传》</p> <p>(5) 又如:驟马(策马奔驰)</p> <p>(6) 泛指奔驰 [gallop]</p> <p>麋鹿见之快驟。——《庄子》</p> <p>词性变化</p> <p>◎ 驟</p>	<p>驟 zhòu ㄓㄡˋ</p> <p>〈形〉</p> <p>(1) 迅疾，猛快 [fast;prompt]</p> <p>杞伯于是驟朝于晋。——《左传·成公十八年》</p> <p>飘风不终朝，驟雨不终日。——《老子》二十章</p> <p>驟雨初歇。——宋·柳永《雨霖铃》</p> <p>驟视之。——清·薛福成《观巴黎油画记》</p> <p>(2) 又如:驟膺(迅速接受);驟雨(暴雨);驟步(快走);驟进(快速前进);驟淹(迅速消失);驟兴(迅速兴起);驟断(迅速决断)</p> <p>◎ 驟</p> <p>驟 zhòu ㄓㄡˋ</p> <p>〈副〉</p> <p>(1) 突然 [suddenly;sudden]</p> <p>倘魏兵驟至，四面围定，将何策保之?——《失街亭》</p> <p>(2) 又如:驟见(突然相见);驟面(突然见面);天气驟变;狂风驟起;驟寒(天气突然变冷)</p> <p>(3) 屡次 [frequent;frequently]</p> <p>宣子驟谏。——《左传·宣公二年》</p> <p>(4) 又如:驟胜(屡胜);驟战(屡战);驟谏(屡次进谏)</p>
--	---

Figure 3.2 - The Definition of the Character “驟 (zhòu)” in Han Dian

- Online Xin Hua Dictionary (在线新华字典): <https://zd.hwxnet.com/>

As part of hwxnet (汉文学网) (<https://www.hwxnet.com/>), it does not provide detailed information about its founders and contributors.

Online Xin Hua Dictionary (在线新华字典) has an extensive collection, encompassing over 20,000 simplified and traditional Chinese characters. It serves as a valuable resource for Chinese language learners, offering information such as pinyin pronunciation, radical categorization, stroke count, character meanings, synonyms, antonyms, homophones, English translations, word definitions, part-of-speech variations, Kangxi Dictionary explanations, Shuowen Jiezi interpretations, character etymology, and commonly used word combinations. This resource supports character lookup through pinyin input and handwriting recognition.

在线新华字典 > 骤字的解释及意思



繁体: 驟 分类: 通用字、常用字  
 拼音: zhòu  
 部首: 马 部外笔画: 十四画 总笔画: 十七画  
 繁体: 馬 部外笔画: 十四画 总笔画: 十七画  
 笔顺: ㇇㇇一丨丨一一フ、ノ丨ノノノ、  
 仓颉: NMSEO 四角号码: 77132 U+9AA4 五笔86/98: CBCICGBI

分享到: [新浪微博](#) [微信](#) [QQ空间](#) [豆瓣](#) [百度贴吧](#) [复制网址](#) [更多](#)  
 阅读(7453次)

#### 详细字义解释

◎ 骤 骤 zhòu

〈动词〉

(1) (形声。从马,聚声。本义:马奔驰)

(2) 同本义

骤,马疾步也。——《说文》

车骤徒趋。——《周礼·大司马》

步及骤处兮。——《楚辞·招魂》。注:“走也。”

车驱而驰(骤)。——《礼记·曲礼》

驾彼四骆,载骤馵馵。——《诗·小雅·四牡》

(3) 又如:骤骥(疾驰的骏马)

(4) 使马奔驰

遇春骤马追到,便活擒于马上。——《英烈传》

(5) 又如:骤马(策马奔驰)

(6) 泛指奔驰

麋鹿见之快骤。——《庄子》

#### 词性变化

◎ 骤 骤 zhòu

〈形容词〉

(1) 迅疾,猛快

杞伯于是骤朝于晋。——《左传·成公十八年》

飘风不终朝,骤雨不终日。——《老子》二十章

骤雨初歇。——宋·柳永《雨霖铃》

骤视之。——清·薛福成《观巴黎油画记》

(2) 又如:骤膺(迅速接受);骤雨(暴雨);骤步(快走);骤进(快速前进);骤淹(迅速消失);骤兴(迅速兴起);骤断(迅速决断)

Figure 3.3 - The Definition of the Character “骤 (zhòu)” in Online Xin Hua Dictionary

In Figures 3.2 and 3.3, we can see an example featuring the character “驟 (zhòu ‘(horse) run’)” in Han Dian (zdic.net) and Online Xin Hua Dictionary. Originally, this character meant “horse trots” or “gallop.” However, in modern Chinese, it has retained its meaning of “fast” as an adjective or “suddenly” as an adverb.

In the term “步驟 (bù zhòu, 'steps\_N)”, which is a highly frequent term in our patent corpus, “驟 (zhòu)” is in coordination with “步 (bù, ‘step’)”, which was also a verbal character<sup>73</sup> in classical Chinese. In this context, the relationship between them should be “conj@m.”

However, when we only consider the distributional tests, the term appears to be a combination of a NOUN and an ADV. Giving consideration to the fact that nouns are normally not accompanied by adverbs, the relationship changes to “flat@m.” This discrepancy highlights the complexity of character usage evolution in modern Chinese and the challenges in categorizing and annotating such terms based solely on distributional tests.

The preference in our approach leans more towards distributional tests than etymological semantics. We only resort to etymological meaning when distributional tests cannot provide a clear classification.

Open class words	Closed class words	Other
<a href="#">ADJ</a>	<a href="#">ADP</a>	<a href="#">PUNCT</a>
<a href="#">ADV</a>	<a href="#">AUX</a>	<a href="#">SYM</a>
<a href="#">INTJ</a>	<a href="#">CCONJ</a>	<a href="#">X</a>
<a href="#">NOUN</a>	<a href="#">DET</a>	
<a href="#">PROPN</a>	<a href="#">NUM</a>	
<a href="#">VERB</a>	<a href="#">PART</a>	
	<a href="#">PRON</a>	
	<a href="#">SCONJ</a>	

**Table 3.2 - List of UD POS tags.**

Within our annotation schema, we encompass all 17 parts-of-speech (UPOS) tags<sup>74</sup> derived from the Universal Dependencies (UD) framework (Nivre et al., 2016), as depicted in Table 3.2. However, owing to the distinct writing style prevalent in patent texts, it is worth noting that the INTJ (interjection) tag does not appear in the final treebank.

<sup>73</sup> The verbal characters are characters that function on their own as a verb.

<sup>74</sup> For the POS other than NOUN, ADJ, VERB and ADV, we follow the definition of the Universal Dependencies and examples from other existing treebanks.

### 3.1.1.2 Tests for Inter-characters Relations

In this section, we explore the relationship between the term internal dependency relations and syntactic relations in the modern Chinese lexicon. Building upon the framework presented in Section 1.2.1, which identifies six classes<sup>75</sup> of character-level dependency relations, we establish connections between word-level structures and sentence-level structures annotated as dependency relations in treebanks like UD and SUD projects.

For each class, we provide a definition and establish its correspondence with syntactic relations in SUD by giving examples from our patent claim corpus.

As mentioned in Section 3.1.1 above, the assignment of classes takes into consideration three factors:

- the boundary between character-level relations and conventional syntactic relations for the assignment of “@m”;
- the internal structure of the term and its corresponding SUD label (dependency relation type and head position);
- and the lexical classes of both the entire term (ExtPos) and its constituent component characters (UPOS).

Additionally, we propose criteria and tests to determine whether a term belongs to a particular class, which are further detailed in the decision tree outlined in Section 3.3.

Our methodology involves the utilization of an automatic POS tagging system, as outlined below in Section 3.2, which then subjects to manual correction using the POS tests in Section 3.1.1.1 and the Part-of-Speech Tagging Guidelines for the Penn Chinese Treebank (3.0) (Xia, 2000b) for word-level annotation and the Xinhua Dictionary for character-level annotation. It is worth highlighting that the process of selecting POS tags and word-internal relation labels for characters occurs concurrently, with the relationship type exerting a substantial influence on the POS assignment. This intricate interplay between relation types and POS choices is extensively discussed in the subsequent section.

As mentioned in Section 1.2.1.2, based on statistics provided by Zhou Jian in the 1983 edition and Bian Chenglin in the 1990 edition of the “Modern Chinese Dictionary” (《现代汉语词典》), disyllabic terms in Chinese can be categorized into five basic types. Notably, over 50% of these terms fall into the attributive type, followed by the coordinative type, which accounts for over 20%. In the third position is the predicate-object type, with percentages of 15.6% and 19.31% in the respective editions. Conversely, the subject-predicate and predicate-complement types together comprise less than 5% of disyllabic terms.

In terms of quantity statistics, it's worth noting that the attributive type and coordinative type types are the most prevalent, and our annotation process will prioritize these two types.

---

<sup>75</sup> Coordinative type (联合型, lián hé xíng; 并列式, bìng liè shì), Attributive type (偏正型, piān zhèng xíng), Predicate-complement type (补充型, bǔ chōng xíng), Verb-object type (动宾型, dòng bīn xíng; 支配式, zhī pèi shì) and Subject-predicate type (主谓型, zhǔ wèi xíng).

## 3.1.1.2.1 Coordination compounds

Coordination compound terms are created from two or more morphemes, which are typically synonyms, antonyms, or semantically related. The meaning of the compound term can either be a combination of its morphemes, leaning towards one of its characters, or entirely independent of the meanings of its components.

Our tests rely on analyzing the part-of-speech composition of the component characters within coordination compound terms. In terms of lexical class, a coordination compound term can comprise two nominal characters, two verbal characters, or two adjective characters. Examples of each subclass are provided below:

## 1) Two nominal characters: N1 + N2

These lexicalized terms differ from phrases in that they allow for the rearrangement of their component characters and do not necessitate the conjunction “和 (hé, ‘and’)” between them. In the following examples from Feng (2004b) and our patent corpus, we contrast the coordinative phrase (1)-(2) and compound (3)-(5):

(1)	字	母	数	字
	zì	mǔ	shù	zì
	‘character_N’	‘mother_N’	‘digit_N’	‘character_N’
	‘ <i>alphabet_N</i> ’		‘ <i>digit_N</i> ’	
	‘ <i>alphabet and digit</i> ’			
(2)	输	入	输	出
	shū	rù	shū	chū
	‘transport_V’	‘enter_V’	‘transport_V’	‘get out_V’
	‘ <i>input_N/V</i> ’		‘ <i>output_N/V</i> ’	
	‘ <i>input and output</i> ’			
(3)	信	息		
	xìn	xī		
	‘letter_N’	‘news_N’		
	‘ <i>Information_N</i> ’			
(4)	权	利		
	quán	lì		
	‘power_N’	‘profit_N’		
	‘ <i>Right_N</i> ’			
(5)	时	刻		
	shí	kè		
	‘hour_N’	‘quarter (hour)_N’		
	‘ <i>Moment_N; Always_ADV</i> ’			

Among the examples above, “信息 (xìn xī, ‘information’)” and “权利 (quán lì, ‘right’)” are two examples of this particular subclass, also appearing in the patent treebank. While in “信息 (xìn xī)”, the two characters are synonymous, and the compound's meaning is a synthesis of these two characters, in the case of “权利 (quán lì)”, the meaning is more inclined towards “权 (quán)”.

It's important to note that the external part-of-speech (ExtPos) is not necessarily a noun. In terms like “时刻 (shí kè)”, its ExtPos may be both a noun and an adverb.

In the patent claims, although there are a great number of terms that have this internal structure, as in example (3)-(5), it is very difficult to detect this subclass by any distributional tests.

The identification of this type of coordination compound term relies heavily on semantics, where the two component characters are often either synonyms or antonyms. However, when considering the combination of the component characters' part-of-speech, there can be confusion with the NOUN-NOUN type attributive compound terms, which will be discussed in the next subsection.

According to Dong (2011), the process of lexicalization involves the fusion of two parallel abstract semantic elements or two concrete nouns, transforming them into words through metaphor or generalization. In this process, each component loses its individual meaning.

One proposed test involves expanding the term using the conjunction “和 (hé, ‘and’)”. We first identify disyllabic synonyms for the two-component characters and create a coordinative structure with “和 (hé, ‘and’)”, as demonstrated in the following example:

权 利	=	权 力	和	利 益
quán lì		quán lì	hé	lì yì
‘right’		‘power’	‘and’	‘profit’

“权 (quán)” in “权利 (quán lì)” can be regarded as the abbreviation of the term “权力 (quán lì)”, and as the same “利 (lì)” in “权利 (quán lì)” can be regarded as abbreviation of the term “利益 (lì yì)”.

## 2) Two verbal characters: V1 + V2

Similar to the first category, there is a wealth of examples of this type within the technical domain.

(6)	接	收
	jiē	shōu
	‘receive’	‘receive’
	‘Receive_V’	

- (7) 计            算  
 jì            suàn  
 ‘count’      ‘calculate’  
 ‘Calculate\_V’
- (8) 操            作  
 cāo           zuò  
 ‘manipulate’ ‘do’  
 ‘Operation\_N’

Within this subclass, there are instances such as 接收 (jiē shōu) and 计算 (jì suàn), which are composed of a sequence of two verbs and function as verbs themselves. In contrast, 操作 (cāo zuò), which combines a pair of verbs, is typically employed as a noun.

Dong (2011) suggests that the combination of two transitive verbs is more prone to forming compound words.

Furthermore, there are specific examples, such as 步骤 (bù zhòu), as discussed earlier, where the parts-of-speech of the component characters are no longer evident. In such cases, we had to consult dictionaries to uncover their original meanings.

- (8) 步            骤  
 bù            zhòu  
 ‘walk\_V’      ‘(horse) run\_V’  
 ‘steps\_N’

There are some distribution tests for this subclass. Having a disyllabic term that is composed of two characters: c1 and c2, the term can be transformed into the following structures:

- Test 1: “c1 而 (ér) c2 之 (zhī)<sup>76</sup>”
- Test 2: “先 (xiān, ‘first’) c1 后 (hòu, ‘then’) c2”

We include the successive type (递续式复合词, dì xù shì fùhécí)<sup>77</sup> (Zhou, 2016) in this subclass (Test 2).

In contrast, the verbal coordinative phrases are mostly composed of more than two characters and can be transformed into the following structures:

- “c1 并/并且 (bìng / bìng qiě, ‘and’) c2”
- “一边 (yī biān, ‘simultaneously’) c1 一边 (yī biān, ‘simultaneously’) c2”

<sup>76</sup> Can be translated as “While c1, (then) c2”.

<sup>77</sup> It is different from words with predicate-object structure, parallel structure, and words with complementary structure. The two actions represented by their two morphemes, whether issued by the same subject or by different subjects, all occur one after another.

3) Two adjective characters: A1 + A2

There are few examples of this kind in technical domains.

(9) 弯 曲  
wān qū  
'bent' 'curved'  
'Bent, curved\_ADJ'

(10) 安 全  
ān quán  
'safe' 'complete'  
'Safe\_ADJ'

(11) 疏 密  
shū mì  
'sparse' 'dense'  
'Density\_N'

Similar to subclasses (1) and (2), compared to that of their characters, the external POS of the compound can be the same as (e.g. 弯曲 (wān qū) is an adjective) or different to (e.g. 安全 (ān quán) and 疏密 (shū mì) is a noun) the POS of its component characters.

Some possible distributional tests would be the transformation into the following structures:

- “c1 c1 c2 c2”
- “c1 而不 (ér bù, ‘instead of’) c2”
- “又 (yòu, ‘also’) c1 又 (yòu, ‘also’) c2”

The first and second examples passed the first test: 弯弯曲曲 (wān wān qū qū, ‘twisty’), 安安全全 (ān ān quán quán, ‘safe’). The third example 疏密 (shū mì, ‘sparse dense’, ‘density’) passed the third test as it can be transformed into “疏而不密 (shū ér bù mì, ‘sparse but not dense’)”.

Another distributional test for all three subclasses of coordinative compound terms would be the substitution test like below:

- A B → C A and C B  
更 动 → 变 更 变 动  
gēng dòng biàn gēng biàn dòng  
'replace' 'move' 'change' 'replace' 'change' 'move'
- A B → A C and A B  
位 置 → 位 于 置 于  
wèi zhì wèi yú zhì yú



‘position’ ‘place’

‘position’ ‘at’

‘place’ ‘at’

All coordination structures are considered as “conj@m” relations in SUD with edge direction left-to-right. While the constituent characters must share the same UPOS, the ExtPos of the whole can vary in spite of the lexical class of its constituent characters.

### 3.1.1.2.2 Attributive compounds

Commonly, attributive compound terms or modifier-head compound terms may consist of two or three characters.

In the first case the term “AB”, where “A” (or the modifier character) modifies “B” (the head character, which can be a noun, an adjective or a verb).

There are also three subclasses depending on the part-of-speech of the centre character of the terms:

(1) The centre character is a NOUN.

“电源 (diàn yuán)”, “单元 (dān yuán)” and “走线 (zǒu xiàn)” have a noun as the centre character. However, the modifier can be also a noun (as in the first term), an adjective (as in the second term) or a verb (as in the third term).

- |      |                  |               |
|------|------------------|---------------|
| (12) | 电                | 源             |
|      | diàn             | yuán          |
|      | ‘electricity_N’  | ‘origin_N’    |
|      | ‘Power source_N’ |               |
| (13) | 单                | 元             |
|      | dān              | yuán          |
|      | ‘single_ADJ’     | ‘component_N’ |
|      | ‘Unit_N’         |               |
| (14) | 走                | 线             |
|      | zǒu              | xiàn          |
|      | ‘walk_V’         | ‘wire_N’      |
|      | ‘Routing_N/V’    |               |

(2) The centre character is an ADJ.

As for modifier compounds with adjective head characters, the only example that we annotated in our patent treebank is 最终(zuì zhōng).

- |      |                        |            |
|------|------------------------|------------|
| (15) | 最                      | 终          |
|      | zuì                    | zhōng      |
|      | ‘most_ADV’             | ‘last_ADJ’ |
|      | ‘Eventual(ly)_ADJ/ADV’ |            |

(3) The centre character is a VERB.

While 触控 (chù kòng) is an example with a verbal head character with a verbal as a modifier, 预设 (yù shè) is an example with a verbal head character with an adverbial character as a modifier, noun characters can also serve as modifiers of verbal characters as shortened forms of oblique structures such as “V as N”, “V with N”, “V towards N”, etc., like in 压感 (yā gǎn).

- |      |                           |             |
|------|---------------------------|-------------|
| (16) | 触                         | 控           |
|      | chù                       | kòng        |
|      | ‘touch_V’                 | ‘control_V’ |
|      | ‘Controlled by touch_ADJ’ |             |
|      |                           |             |
| (17) | 预                         | 设           |
|      | yù                        | shè         |
|      | ‘in advance_ADV’          | ‘set up_V’  |
|      | ‘Presuppose_V’            |             |
|      |                           |             |
| (18) | 压                         | 感           |
|      | yā                        | gǎn         |
|      | ‘pressure_N’              | ‘sensing_V’ |
|      | ‘Pressure sensing_ADJ’    |             |

And in the second case the term “ABC”, where “AB” together modify “C” (head character). In contrast to the diverse ExtPos patterns observed in bisyllabic modifier compound terms, the majority of trisyllabic modifier compound terms are classified as nouns. In many studies, the head character in these terms is often regarded as a suffix due to its productivity. We label them as single terms because the last head character cannot be used independently, signifying their status as independent characters.

- |      |                 |           |          |
|------|-----------------|-----------|----------|
| (19) | <u>信</u>        | <u>号</u>  | 线        |
|      | xìn             | hào       | xiàn     |
|      | ‘message_N’     | ‘sign_N’  |          |
|      | ‘signal_N’      |           | ‘wire_N’ |
|      | ‘Signal wire_N’ |           |          |
|      |                 |           |          |
| (20) | <u>输</u>        | <u>入</u>  | 端        |
|      | shū             | rù        | duān     |
|      | ‘transport_V’   | ‘enter_V’ |          |
|      | ‘input_N’       |           | ‘side_N’ |
|      | ‘Input side_N’  |           |          |

Compound terms in this group are annotated with “mod@m” label with edge direction right-to-left. The syntactic head (head character) of a modifier compound is always the last character, and the ExtPos of the whole term is always the same as the UPOS of its head character.

For attributive compound terms, the test set includes the check on if there exists:

1. Possible expansion with “的/地” (de)<sup>78</sup>, e.g. “电源 (diàn yuán, ‘power source’)” can be extended into “电气的源头 (diàn qì de yuán tóu, ‘source of the electricity’)”, where “电 (diàn, ‘electricity’)” stands for “电气 (diàn qì, ‘electricity’, which is itself a modifier-head compound with 电 (diàn, ‘electricity’) as its head)”, “源 (yuán, ‘source’)” stands for “源头 (yuán tóu, ‘source’, which is itself a modifier-head compound with 源 (yuán) as its head)”. And when the insertion does not need expansion with “的/地” (de), we consider that it should be a syntactic-level relation.

(21) 电 气 的 源 头  
 diàn qì de yuán tóu  
 ‘electricity\_N’ ‘of’ ‘source\_N’  
 ‘The source of the electricity’

2. The paradigm of the head character, such as the productive character “感 (gǎn)” that can combine with “光 (guāng, ‘light’)” and “声 (shēng, ‘sound’)”.

(22) 光 感  
 guāng gǎn  
 ‘light\_N’ ‘sensing\_V’  
 ‘Light sensitive\_ADJ’

(23) 声 感  
 shēng gǎn  
 ‘sound\_N’ ‘sensing\_V’  
 ‘Sound sensitive\_ADJ’

3. Possible expansion into a corresponding phrase for those with a verbal head character, e.g. “压感 (yā gǎn)” is expanded into “用压力的方式感 (yòng yā lì de fāng shì gǎn)”.

(24) 用 压力 的 方式 感  
 yòng yā lì de fāng shì gǎn  
 ‘use’ ‘pressure’ ‘of’ ‘method’ ‘sense’  
 ‘Sense with pressure’

One specific type of attributive compound term consists of a noun head character in its first position and a classifier (e.g. “文件 (wén jiàn)”) or a second noun character indicating the category or form of the first noun (e.g. “模块 (mó kuài)”). The external POS of compounds of this type is always NOUN. This type can be easily identified by the presence of a classifier as the second character.

(25) 文 件  
 wén jiàn  
 ‘article\_N’ ‘item\_N’; classifier  
 ‘Document\_N’

<sup>78</sup> 的/地/得 DE are noun modifier particle, adjective modifier particle and verb modifier particle in Chinese.

- (26) 模                      块  
 mó                         kuài  
 ‘model\_N’                ‘block\_N’; classifier  
 ‘Module\_N’

### 3.1.1.2.3 Subject-predicate compounds

In subject-predicate compounds, similar to modifier compounds, the head is also the last character, which is a verb (e.g. “水淹 (shuǐ yān)”, “针对 (zhēn duì)”) or an adjective<sup>79</sup>, and the first character is a noun and serves as the subject of the head. Different from modifier compounds, the external POS of the term does not always correspond to the POS of the head character.

- (27) 水                      淹  
 shuǐ                        yān  
 ‘water\_N’                ‘submerge\_V’  
 ‘Submerged by water\_V’

- (28) 针                      对  
 zhēn                        duì  
 ‘needle\_N’                ‘point\_V’  
 ‘Be directed against\_V’

Subject-predicate structures are annotated as “subj@m” with edge direction right-to-left.

Together with predicate-object compounds and predicate-complement compounds, the test for these three last classes is that at least one character of the compound can have one of the aspect markers “了 (le) /着 (zhe) /过 (guo)” without a change of meaning, which means that this character is a verbal character. And subject-predicate compounds distinguish themselves from the other two by the fact that they have only one verbal character that is in the second position and its first character can be modified by the noun modifier particle “ADJ 的 (de)” without a change of meaning, which means that it is a noun character. In the example of “水淹 (shuǐ yān)”, it is possible to say “淹着 (yān zhe, ‘submerged’)” and “ADJ 的水 (ADJ de shuǐ, ‘ADJ water’)”.

### 3.1.1.2.4 Predicate-object compounds

Contrary to the subject-predicate structure, the predicate-object compounds have their first character as head and its second character as the direct object of the verbal head (e.g. “结果 (jié guǒ)”, “通信 (tōng xìn)”, “传感 (chuán gǎn)”), which is usually a noun character.

Predicate-object structures are considered equal to “comp:obj@m” relation with a left-to-right edge.

- (29) 结                      果  
 jié                         guǒ  
 ‘bear\_V’                    ‘fruits\_N’

<sup>79</sup> No example from the patent treebank.

‘Results\_N’

(30) 通            信  
 tōng            xìn  
 ‘go through\_V’ ‘letter\_N’  
 ‘Communicate\_V’

(31) 传            感  
 chuán            gǎn  
 ‘transfer\_V’    ‘sense\_N’  
 ‘Sensing\_V’

In contrast to subject-predicate compounds, predicate-object and predicate-complement compounds have a verbal head character in the first position. Though both of them are annotated as “comp”, the predicate-object compounds have a noun character on the second position, while the second character of predicate-complement compounds is never a noun.

### 3.1.1.2.5 Predicate-complement compounds

A predicate-complement compound of this type is always verbal and has a comp-like relation marked as different sub-relations in SUD, such as “comp:obl” (for oblique arguments of verbs, adjectives, adverbs, nouns or pronouns, e.g. “来自 (lái zì, ‘come from’)”), “comp:dir” (for directional arguments of verbs, e.g. “接入 (jiē rù, ‘gain access to’)”), “comp:res” (for resultative arguments of verbs, e.g. “输出 (shū chū, ‘output’)”) and “comp:aux” (for the argument of auxiliaries, e.g. “可变 (kě biàn, ‘variable’)”), and corresponds to the “aux” relationship as defined by UD).

(32) 来            自  
 lái              zì  
 ‘come\_V’      ‘from\_PREP’  
 ‘Come from\_V’

(33) 接            入  
 jiē              rù  
 ‘connect\_V’    ‘enter\_V’  
 ‘Gain access to\_V’

(34) 输            出  
 shū              chū  
 ‘transport\_V’    ‘get out\_V’  
 ‘Output\_N/V’

(35) 可            变  
 kě                biàn  
 ‘can\_AUX’      ‘change\_V’  
 ‘Variable\_ADJ’

In predicate-complement compound terms, the first character is a verbal head character.

In this first version of annotation, all subtypes of predicate-complement relations are simply annotated as “comp@m”, except the “comp:obl@m” relation in which the second character is an adposition.

### 3.1.1.2.6 Non-compound Terms with Unclear Internal Structures

Besides the compound terms in modern Chinese, there are other types of words that contain more than one word but whose internal structures have no direct correspondence to modern Chinese syntactic relations, such as polysyllabic simple words, transliterated words and onomatopoeia. We borrowed the label “flat”<sup>80</sup> from SUD/UD schema and created the corresponding character-level relation “flat@m” for them.

- The first usage of “flat@m” is for unclear internal structures often involving the lexicalization of cross-layer structures<sup>81</sup>. And these terms are mostly function terms. The internal structure of compounds generated through this process is quite concealed and challenging to analyze in a synchronic context.

(36) 之            间  
 zhī            jiān  
 ‘of\_PREP’    ‘gap\_N’  
 ‘Between\_ADV’

(37) 之            前  
 zhī            qián  
 ‘of\_PREP’    ‘before\_N’  
 ‘Before\_ADV’

- The second application pertains to loanwords, encompassing both transliterated terms and borrowed terms from Japanese that also incorporate Chinese characters.

(38) 以            太  
 yǐ            tài  
 ‘Ether\_N’

(39) 微    积    分  
 wēi    jī    fēn  
 ‘Microaccumulation\_N’

<sup>80</sup> The flat relation in UD is used to combine the elements of an expression where none of the immediate components can be identified as the sole head using standard substitution tests.

<sup>81</sup> The so-called “cross-layer structure” refers to a structure composed of two elements that do not constitute a direct pair but belong to different syntactic units and are adjacent in linear order (Haò Jìngcún, Liáng Bóshū, 1992)

Note that the subclass “@m” is specially designed for relations between Chinese characters. As a result, transliterated words using Chinese characters are labelled as “flat@m”, but foreign words are always labelled as “flat”.

It is worth noting that the flat@m category holds a distinctive position within the character-level relation framework. This category not only encompasses multi-character simple words like transliterated words, onomatopoeia, and reduplicated words but also includes words that do not fit into other predefined categories. This particularly applies to words containing structures that may come from the ancient Chinese, which can make it hard to notice the original structure from a modern point of view.

Another remark is that our annotation schema does not contain the confusing label “compound” anymore. In the original UD schema, the “compound” relation contains noun-noun compounds and verb and verb-object compounds, whose boundaries with “nmod”, “scomp”, “xcomp” and word segmentation are not very clear.

### 3.1.1.2.7 Terms Composed of More Than Two Characters

In contemporary Chinese, terms can have varying numbers of characters. Chinese terms can be monosyllabic, composed of a single character (unigrams), or polysyllabic, consisting of multiple characters (polygrams).

In a comprehensive study conducted by Su Xinchun (2001), an extensive investigation was undertaken to analyze the frequency distribution of modern Chinese words based on their respective number of characters. The research aimed to provide valuable insights into the usage patterns and prevalence of different word lengths in the Chinese language. The findings of this study, including the frequency data, are presented in Table 3.3.

It becomes apparent that it is rare to find a term composed of four or more characters in the patent corpus. Particularly in the specific technical domain of patents, the noteworthy technical terms predominantly appear in the form of bigrams or trigrams.

	<b>total entries</b>	<b>monosyllabic</b>	<b>bisyllabic</b>	<b>trisyllabic</b>	<b>quadrisyllabic and so on</b>
《现代汉语常用词词频词典》 Modern Chinese Frequency Dictionary of Commonly Used Words	77,482	7,611	46,729	11,213	11,929
		10%	60%	15%	15%
《现代汉语词典》 Modern Chinese Dictionary	56,147	10,540	35,056	5,703	4,766
		19%	62%	10%	9%

**Table 3.3 - Statistics of Character Count of Chinese Words (Su Xinchun, 2001)**

Up until this point, all the tests described have focused on disyllabic terms in Chinese. However, in this section, we will delve into the explanation of how to handle trisyllabic terms. These longer terms require a distinct approach to ensure accurate annotation and understanding within the context of the patent corpus.

According to Zhou (2016), three-character combinations may have three structural patterns that are explained with examples below: 1+1+1, 2+1 or 1+2. According to the 1996 edition of the Modern Chinese Dictionary (《现代汉语词典》), there are only 7 three-character combinations with a 1+1+1 pattern, accounting for approximately 0.14% of the total number of three-character combinations. Thus, we do not give consideration to this case. Three-character combinations with a 2+1 pattern amount to 2,997, constituting approximately 62.08% of the total, while those with a 1+2 pattern reach 1,824, making up roughly 37.78% of the total. Both the 2+1 and 1+2 patterns of three-character combinations exhibit some similar structural relationships. In both cases, the predominant structure involves a modifier-head relationship.

One typical structure of trisyllabic terms in Chinese patents is the combination of a disyllabic term and a highly productive character at the beginning or the end, corresponding to the 2+1 and 1+2 terms. Some researchers also categorize these two types as derived terms.

- 2+1 pattern

“传输机 (chuán shū jī)” is a typical example of the 2+1 pattern, with the hyperproductive character “机 (jī, ‘machine’)” in the end. With the same character “机 (jī)”, we have many examples from the corpus, such as “发电机 (fā diàn jī)”, “宿主机 (sù zhǔ jī)”, etc. This last character is mostly a nominal character.

(40)	传                输	机
	chuán                shū	jī
	‘transfer_V’        ‘transport_V’	
	‘transmission_N’	‘machine_N’
	‘Transmitter_N’	

(41)	发                电	机
	fā                        diàn	jī
	‘generate_V’        ‘electricity_N’	
	‘generate electricity_V’	‘machine_N’
	‘Electricity generator_N’	

(42)	宿                主	机
	sù                        zhǔ	jī
	‘lodge_V’            ‘master_N’	
	‘host_N’	‘machine_N’
	‘Host computer_N’	



Other frequently used characters are “器 (qì, ‘machine\_N’), “区 (qū, ‘block; area\_N’), “端 (duān, side\_N)”, etc.

One exception is “可视化 (kě shì huà)”. While it corresponds to the 2+1 pattern of trisyllabic terms, it is not a modifier-head relationship, but a predicate-complement structure with “化 (huà)” as head. In this specific case, the character “化 (huà)” is a special suffix mark that transforms an adjective into a verb.

(43)	<u>可</u>	<u>视</u>	化
	kě	shì	huà
	‘capable_AUX’	‘see_V’	
	‘visual_ADJ’		suffix
	‘Visualize_V’		

- 1+2 pattern

For the 1+2 pattern, there are more variations, such as ADV+V (e.g. “未保存 (wèi bǎo cún)”), ADJ+N (e.g. “互信息 (hù xìn xī)”), AUX+V (e.g. “可插接 (kě chā jiē)”) and V+V (e.g. “待验证 (dài yàn zhèng)”).

(44)	未	<u>保</u>	<u>存</u>
	wèi	bǎo	cún
	‘not_ADV’	‘keep_V’	‘safe_V’
	‘Unsafe_ADJ’	‘safe_V’	
(45)	互	<u>信</u>	<u>息</u>
	hù	xìn	xī
	‘mutual_ADJ’	‘letter_N’	‘message_N’
	‘Mutual information_N’	‘information_N’	
(46)	可	<u>插</u>	<u>接</u>
	kě	chā	jiē
	‘capable_AUX’	‘plug_V’	‘attach_V’
	‘Plugable_ADJ’	‘plug_V’	
(47)	待	<u>验</u>	<u>证</u>
	dài	yàn	zhèng
	‘wait_V’	‘exam_V’	‘confirm_V’
	‘Awaiting verification_ADJ’	‘validate_V’	

One problematic example is “自适应 (zì shì yīng)”, in which “自 (zì)” is a pronoun. We annotate it as “subj@m”.

(48)	自	<u>适</u>	<u>应</u>
	zì	shì	yīng
	'self_PRON'	'adapt_V'	
	'Adaptive_ADJ'		

The annotation process for trisyllabic terms consists of three main steps.

1. Firstly, the productive “affix” character within the term is identified;
2. Next, the inter-character relation of the remaining bigram is annotated using tests specific to inter-character relations above;
3. Finally, the relation between the “affix” character and the head character of the remaining bigram is annotated. In the patent corpus, this relation is typically mod@m, with occasional exceptions where it may be comp@m. The determination of the head character depends on the position of the “affix” character.

Furthermore, the treebank includes several terms consisting of more than three characters, necessitating distinct annotation procedures.

(50)	主	<u>交 换</u>	<u>芯 片</u>
	zhǔ	jiāo huàn	xīn piàn
	'mian_ADJ'	'exchange_V'	'chip_N'
	'Main switching chip_N'		

(49)	<u>非</u>	<u>显 示</u>	区
	fēi	xiǎn shì	qū
	'no_ADV'	'demonstrate; display_V'	'block; area_N'
	'Non-display area_N'		

Much like trisyllabic terms, these terms are regarded as indivisible due to the fact that their initial character in the first example or their concluding character in the second example can not be used independently.

The challenge of automated segmentation is primarily attributed to these multi-syllabic terms, which will be addressed in Section 3.2.2.

### 3.1.2 Hierarchizing the Morphological and Syntactic Relation Labels for Character-level Treebanks

The annotation of syntactic relations of the Chinese Patent Treebank is based on the surface-syntactic universal dependencies (SUD) schema proposed by (Gerdes et al., 2018). And on top of it, we add our own character-level annotation tags by analogy with SUD’s surface-syntactic relations.

We categorize these tags into three granularity levels based on the transparency of their internal structure, ranging from high to low transparency: 1. The regular syntactic level (including all SUD syntactic relations tags); 2. The computational lexical level (including all tags with “@m” except “flat@m”); and 3. Finally, the syntactic level of relations between Chinese characters that do not exist internally or are not analogous to syntactic structures (using flat@m).

#### → LEVEL 1: syntactic relations

Tags	Syntactic Relations
appos <sup>82</sup>	An appositional modifier of a noun is a nominal immediately following the first noun that serves to define, modify, name, or describe that noun.
conj <sup>83</sup>	A conjunct is a relation between two elements connected by a coordinating conjunction, such as and, or, etc.
comp <sup>84</sup>	The <i>comp</i> relation is used for arguments of verbs, nouns, adjectives, adverbs, auxiliaries, adpositions and conjunctions.
comp:obj	The <i>comp:obj</i> relation is used for direct object complements, including direct complements of an adposition or a subordinating conjunction.
comp:obl	The <i>comp:obl</i> relation is used for oblique arguments of verbs, adjectives, adverbs, nouns or pronouns, regardless of their form.
comp:cleft	The <i>comp:cleft</i> relation is used in cleft sentences for the dependency from the head of the sentence to the head of the complement clause.
comp:pred	The <i>comp:pred</i> relation is used for predicative arguments of verbs.
	The <i>comp:aux</i> relation is used for the argument of auxiliaries, and

<sup>82</sup> <https://universaldependencies.org/u/dep/appos.html>

<sup>83</sup> <https://universaldependencies.org/u/dep/conj.html>

<sup>84</sup> <https://surfacesyntacticud.github.io/guidelines/u/relations/comp/>

comp:aux	corresponds to the aux relationship as defined by UD.
comp:svc	The <i>comp:svc</i> relation is used for the serial verbs construction.
comp:dir	The <i>comp:dir</i> relation is used for the directional arguments of verbs.
comp:res	The <i>comp:res</i> relation is used for the resultative arguments of verbs.
subj <sup>85</sup>	The <i>subj</i> relation is used for all subjects, regardless of their form (nominal or verbal). This relationship encompasses both the <i>nsubj</i> and <i>csubj</i> relationships as defined by UD, as the following examples show.
mod <sup>86</sup>	The <i>mod</i> relation is used for modifiers of verbs, nouns, adjectives, adverbs, auxiliaries, adpositions and conjunctions.
flat <sup>87</sup>	The <i>flat</i> relation is used for non-Chinese expressions, such as “DIN ISO 4590-86” “Bloking/Reacting Buffer”
parataxis <sup>88</sup>	The <i>parataxis</i> relation (from Greek for “place side by side”) is a relation between a word (often the main predicate of a sentence) and other elements, such as a sentential parenthetical or a clause after a “:” or a “;”, placed side by side without any explicit coordination, subordination, or argument relation with the head word.
punct <sup>89</sup>	The <i>punct</i> relation is used for any piece of punctuation in a clause, if punctuation is being retained in the typed dependencies.

→ LEVEL 2: morphological relations

<sup>85</sup> <https://surfacesyntacticud.github.io/guidelines/u/relations/subj/>

<sup>86</sup> <https://surfacesyntacticud.github.io/guidelines/u/relations/mod/>

<sup>87</sup> <https://universaldependencies.org/u/dep/flat.html>

<sup>88</sup> This approach can be used to analyse two elements that are placed side by side with no explicit marker of coordination, subordination, or argument relation with the head word (<https://universaldependencies.org/u/dep/parataxis.html>)

<sup>89</sup> <https://universaldependencies.org/u/dep/punct.html>

Tags	Morphological relations
conj@m	Coordinative type (联合型, lián hé xíng; 并列式, bìng liè shì): Composed of two morphemes with similar, related, or opposite meanings.
comp:obj@m	Verb-object type (动宾型, dòng bīn xíng; 支配式, zhī pèi shì): The first morpheme represents an action or behaviour, while the second morpheme represents the entity or object associated with that action or behaviour.
comp:aux@m	Auxiliary-verb type (助宾型, zhù bīn xíng): The first morpheme is an auxiliary and the second morpheme is a verb.
comp@m	Predicate-complement type (补充型, bǔ chōng xíng): The subsequent morpheme provides supplementary information to the preceding morpheme.
*comp:obl@m	A subtype of predicate-complement type: The subsequent morpheme is an adposition.
*comp:res@m	A subtype of predicate-complement type: The subsequent morpheme is the resultative complement of the preceding morpheme.
subj@m	Subject-predicate type (主谓型, zhǔ wèi xíng): The latter morpheme predicates the thing or object described by the former morpheme.
mod@m	Attributive type (偏正型, piān zhèng xíng): The preceding morpheme restrictively modifies the following morpheme.

## → LEVEL 3: flat@m

flat@m	<p>When the internal relationship of the term is ambiguous and cannot be classified into any of the previously mentioned types. It includes examples like transliterated terms, onomatopoeia, reduplicated terms, and functional terms that have been lexicalized for a long time without any discernible syntactic structures.</p> <p>This type stands as the only category that cannot be divided and unequivocally qualifies as “words” in this study.</p>
--------	---

### 3.1.3 The Final Decision Tree

In Section 3.1.1, we have already discussed certain tests along with relevant examples. The objective of this section is to consolidate the complete set of tests into an annotation decision tree by examining actual examples from our patent treebank corpus.

In the context of a term comprising two characters “c1c2”, the task at hand involves the determination of the character-level relation between “c1” and “c2”. This section aims to elucidate the decision-making process by employing a range of distributional and semantic tests, thus enabling a comprehensive analysis of the syntactic and semantic properties associated with these characters.

Note that it is crucial to consider the impact of the bi-syllabication phenomenon in Chinese, as also demonstrated in the tests below. When confining the test to individual characters alone, its applicability becomes limited, encompassing only a small subset of terms exhibiting “@m” relations. Acknowledging this constraint, we recognize the necessity to broaden the scope of the tests to include bigrams that contain the character in question and share the same semantic meaning.

#### “mod@m”

The first set of tests (A1-A3) is dedicated to modifier relations within a term in both semantic (A1) and distributional (A2 and A3) aspects. We expect the *mod* relations to go from the second to the first character, and we can thus formulate the test as follow:

A1: Is the whole term “c1c2” a kind of or a method of the second character “c2”?

This test is especially dedicated to terms with a nominal head character. The four specific cases with a verbal head character are listed below:

A1 (a)

Expressing Comparison Relation:

“c2 V” like “c1 N”<sup>90</sup>

像 c1\_N 一样地 c2\_V

xiàng c1\_N yīyàng de c2\_V

- 水平 (shuǐ píng, water flat): Horizontal - ‘flat like the water’<sup>91</sup>

---

<sup>90</sup> This type of word-forming structure is more literal in nature and less observed in technical and scientific corpus, like patent.

<sup>91</sup> This and most of the following examples have been encountered in our patent treebank.

A1 (b)

Expressing Relation with Dependency:

“c2\_V” with the use of “c1\_N”

以/用 N 来 V  
yǐ/yòng N lái V

- 笔谈 (bǐ tán, pen talk): Written discussion - ‘*discuss with the use of pen*’<sup>92</sup>

A1 (c)

Expressing Spatial Relation:

“c2\_V” from/to “c1\_N”

在/向/从 N 处 V  
zài/xiàng/cóng N chù V

- 上传 (shàng chuán, up transport): Upload - ‘*transfer to an upper position*’
- 下载 (xià zài, down load): Download - ‘*transfer to a lower position*’

A1 (d)

Expressing Temporal Relation:

“c2\_V” during “c1\_N”

在 N 时 V  
zài N shí V

- 午睡 (wǔ shuì, noon sleep): Siesta - ‘*sleep at noon*’<sup>93</sup>

A2: Possible expansion of the term with “的/地/得 (de)” (separately, noun modifier particle, adjective modifier particle and verb modifier particle in Chinese<sup>94</sup>) with bi-syllabication<sup>95</sup>?

---

<sup>92</sup> No example from the patent claim treebank.

<sup>93</sup> No example from the patent claim treebank.

<sup>94</sup> See the tests for parts-of-speech in Section 3.1.1.1.

<sup>95</sup> Compared to the “expansion method” or “insertion method” proposed by Huang et Liao (2007) mentioned in Section 1.2.1.3, the expansion here requires the bi-syllabication of single characters, namely the transformation of monosyllabic terms into correspondent disyllabic terms of the same meaning. This difference also distinguishes inseparable terms from phrases.

A3: Existence of a paradigm of the head character.

Example: The character 面 (miàn, *face*) appears in a wide range of terms: “界面 (jiè miàn, ‘interface’), “页面 (yè miàn, ‘page’), “桌面 (zhuō miàn, ‘desktop’)” all having the character “面 (miàn, *face*)” as head preceded by a modifier.

The next test in this set is devoted to a special case of “mod@m” relation - the structure of a NOUN character and a classifier<sup>96</sup> character. These mod@m relations go from left to right, which is untypical for Chinese modifications.

A4: Whether the term “c1c2” can be reformed into the structure “一 (yī, ‘one’) c2 c1” (while changing the number of the term from plural to singular)?

Examples:

车辆 (chē liàng, “vehicles”) = 一辆车 (yī liàng chē, “one car”)

花朵 (huā duǒ, “flowers”) = 一朵花 (yī duǒ huā, “one flower”)

While the original term is a collective noun, the reformed expression signifies a single object.

### “conj@m”

The initial set of three tests (B1-B3) is specifically designed to discern conj@m relations.

B1 and B2 constitute distributional tests, with B1 focusing on adjectives and B2 on verbs. These tests aim to determine whether a character-level structure can be considered a conjunction by placing the two characters within a syntactic-level conjunction structure.

In contrast, B3 entails more semantic criteria that rely heavily on the meaning of a character. This test explores the existence of replication of the meaning and part-of-speech of characters<sup>97</sup> and the possibility of the replacement of the whole term by one of the characters to ascertain their potential for conj@m relations.

B1: Can the characters “c1” and “c2” in a term be used in the following structures?

- “c1 而不 (ér bù, ‘instead of’) c2”
- “又 (yòu, ‘also’) c1 又 (yòu, ‘also’) c2”
- Duplication: “c1 c1 c2 c2”

---

<sup>96</sup> The Chinese classifier, also called “measure word” (量词, liàngcí), is used before a noun when the noun is qualified by a numeral or demonstrative, e.g. “个 (gè)” in “一个人 (yī gè rén, ‘one person’)”.

<sup>97</sup> The parts-of-speech of characters in a term with a conj@m is always identical.



B2: Can the characters “c1” and “c2” in a term be used in the following structures expressing time sequence?

- “先 (xiān, ‘first’) c1 后 (hòu, ‘then’) c2”
- “c1 而 (ér, ‘and’) c2 之 (zhī, ‘that’)”

B3: Can the characters “c1” and “c2” in a term be used in the conjunct construction (with bi-syllabication)?

- “c1 和 (hé, ‘and’) c2”

Among all tests, Test B3 is the one that poses the most problems, due to the unclear criterion of bi-syllabication relating to the independence of the single characters. In Section 3.3.2.1, there are some examples of analysing the problematic terms of this issue.

B4: Can the characters “c1” and “c2” be combined with one same third character “c3”?

Example: “获取” (huò qǔ, ‘obtain’ ‘pick’): gain

= “获得 (huò dé, ‘obtain’ ‘get’)” + “取得 (qǔ dé, ‘pick’ ‘get’)”

### “comp@m” and “subj@m”

The final and most intricate set of tests focuses on the “comp@m” and “subj@m” relations. These two types share a common characteristic, which is the presence of a verbal head character and a dependent character serving as the subject or complement<sup>98</sup> of the verb. However, the distinction lies in the order of the characters. In “comp@m”, the second character is always the verbal head, while in “subj@m”, it is always the first character. Another common characteristic of these two types is that the ExtPos of the whole term is always VERB.

An exception to this pattern is the “comp:aux@m” relation, where an auxiliary character serves as the head instead of a verbal character. This specific case is addressed in Test C1.

Another challenge arises when differentiating “comp@m” from the “conj@m” relation within a term composed of two verbal characters, as both their first and second characters are verbal. Test C5 is specifically designed to tackle this issue as a complement to the test set B1-B3, with the aim of helping to identify a “conj@m” structure.

---

<sup>98</sup> Can be a nominal character, an adjective or adverb character or also a verbal character according to different subtypes of the relation comp@m.

C1: Does the term contain one in the closed list of auxiliary characters in the Chinese language?

- “可(以) (kě yǐ)”: Can, may

Example: “可变” (kě biàn): variable - ‘*can verify*’

- “能(够) (néng gòu)”: Can, able to

Example: “能动” (néng dòng): dynamic - ‘*can move*’

- “应(该) (yīng gāi)”: Should

Example: “应得” (yīng dé): deserving - ‘*should gain*’

If the answer is yes, the term should be annotated with a “comp:aux@m” relation.

The subsequent tests C2-C5 form a sequential set that aims to distinguish relations containing at least one verbal character: “subj@m”, “comp:obj@m”, “comp:obl@m”, “comp:res@m”, and certain ambiguous cases of “conj@m”.

C2: To check which character is verbal in a term, we combine each of its characters in the term “c1c2” with the following aspect markers:

- “了 (le)”: Aspect marker indicating a completed or changed action/state.
- “着 (zhe)”: Aspect marker indicating an ongoing or continuous action/state.
- “过 (guò)”: Aspect marker indicating a past experience or action.

While in a subj@m structure, it should only be able to be combined with the second character, and for a verbal conj@m structure it should be able to be combined with both of the characters, for a comp:obj@m structure only the first character theoretically, for all other subtypes of comp@m the result of the second character is uncertain.

An additional test completing C2 is C3, checking if the other character is nominal.

C3: Does the other character have a “的 (de, possessive marker) + char” structure?

If the answer is true, then the other character is a NOUN, and the choice of the label can be limited to “subj@m” and “comp:obj@m”, depending on the position of the character.

At this point, our focus shifts to the remaining relations: “comp:obj@m”, “comp:obl@m”, “comp:res@m”, “comp:dir@m”, and verbal “conj@m”. Among these, “comp:obl@m” and

“comp:res@m” share a common characteristic. In both cases, the first character serves as the verbal head, as examined in Test C2. However, the second character in each relation falls into a specific category with a closed list in Chinese. For “comp:obl@m”, it corresponds to a preposition character, while for “comp:res@m”, it represents a directional complement character.

C4: Does the second character “c2” belong to one of the following lists?

Adposition Character	Resultitive Complement Character
自 (zì) - from	上 (shàng) - up, above, on, go up
于 (yú) - in, at	下 (xià) - down, below, go down
在 (zài) - in, at	进 (jìn) - enter, go in, advance
当 (dāng) - during	出 (chū) - go out, exit, leave
向 (xiàng) - towards, to	回 (huí) - return, go back
从 (cóng) - from	过 (guò) - pass, cross, go over
以 (yǐ) - with	起 (qǐ) - rise, get up, start
...	...

**Table 3.4 - The List of Adposition Characters and Resultitive Complement Characters in Modern Chinese**

To address the rest of our relation types, we apply another test C5.

C5: By asking the question “(SUBJ) c1c2 (OBJ) 了吗?” (Is SUBJ c1c2 (OBJ)?), the answer can be expanded in which form?

- Answer 1: “(SUBJ) c1 得/不 c2” (SUBJ c1 (not) c2) - suggests the presence of a “comp:res@m” relation between “c1” and “c2”.

Example: “通过” (tōng guò): pass through - ‘pass’

Example: “渗透” (shèn tòu): penetrate through - ‘penetrate’

- Answer 2: “(SUBJ) c1 了/着/过 c2” (SUBJ c1 aspect marker c2) - indicates a “comp:obj@m” relation between “c1” and “c2”.





## 3.2 Construction and Automatic Pre-annotation Processing of the Character-level Chinese Patent Treebank

In this section, we will outline our approach to gathering and automatically preprocessing patent claim sentences, encompassing tasks such as automated tokenization, tagging, and dependency parsing. Additionally, within this section, we will highlight certain challenging outcomes that we aim to address through the implementation of character-level annotation.

We built the Chinese Patent Treebank by randomly selecting sentences from patent claims<sup>99</sup> submitted to the Chinese patent office between November 2017 and September 2018. The first 200 sentences are all from Section G - Physics of the International Patent Classification. Section G is the most accessible to us as it also contains computer science and less formula than for example the sections on chemistry, pharmacy, or materials.

Each of the sentences has been sentence-segmented narrowly in order to obtain short syntactic units<sup>100</sup> by splitting on “。”, “;” and “:” in addition to the newline character.

Splitting also at columns and new lines can result in syntactically incomplete sentences such as the sentence 3.5 (a) “characterized by” below. Yet, most of the sentence units of our treebank are syntactically complete where the main verb contains its dependents.

Then, the shortened sentences are segmented into single characters as in Figure 3.5 (a-c). After the sentence segmentation, the average length of the patent claim is 42.54 characters per sentence.

其 特 征 在 于 :  
qí tè zhēng zài yú  
*‘It is characterized by :’*

**Figure 3.5 (a)**

所 述 壳 体 内 容 纳 有 制 冷 剂 ;  
suǒ shù ké tǐ nèi róng nà yǒu zhì lěng jì  
*‘The housing contains refrigerant;’*

**Figure 3.5 (b)**

<sup>99</sup> All patents are collected from the official site of CNIPA (China National Intellectual Property Administration, former SIPO): <http://patdata1.cnipa.gov.cn/>

<sup>100</sup> A Chinese patent claim sentence contains between 50 and 70 characters on average, which is extremely long compared to general texts (while the average number of characters of the Chinese Grammar Wiki treebanks ([https://arboratorgrew.elizia.net/?#/projects/~:text=chinese\\_grammar\\_wiki\\_morphSUD](https://arboratorgrew.elizia.net/?#/projects/~:text=chinese_grammar_wiki_morphSUD)) is between 10 and 18 per sentence), and even harder to parse.

所 述 制 冷 剂 为 硅 油 。

suǒ shù zhì lěng jì wéi guī yóu 。

*‘The refrigerant is silicone oil.’*

**Figure 3.5 (c)**

In the context of this study, the initial set of 100 claim sentences underwent meticulous manual annotation, while a subsequent set of 100 sentences underwent a preliminary automated pre-annotation process, subsequently followed by a manual correction procedure.

These second 100 sentences in the character-level treebank underwent a multi-step automated annotation process. Primarily, it subjects to automatic annotation encompassing (1) word segmentation, (2) Part-of-Speech (POS) tagging, and (3) dependency parsing. These annotations were conducted based on a collaborative consensus established by three state-of-the-art language processing pipelines: Spacy<sup>101</sup>, Stanza<sup>102</sup>, and Trankit<sup>103</sup>. In the case of word segmentation, the outcomes are indicated using the “@m” designation within the relation labels.

In parallel to the methodology employed for character-level annotation, the automatic POS tagging at the word level was also determined through consensus among the aforementioned language processing pipelines. Unlike the character-level approach, where labels are assigned individually, the word-level annotation retains a singular label corresponding to the part-of-speech assigned to each individual character. This part-of-speech information is archived as external POS (“ExtPos”) linked to character combinations.

The various tests and results with Chinese dependency parsing on words and characters will be presented in Section 4.2.

---

<sup>101</sup> <https://spacy.io/>

<sup>102</sup> <https://stanfordnlp.github.io/stanza/>

<sup>103</sup> <https://trankit.readthedocs.io/en/latest/>

### 3.3 Manual Annotation and Correction of the Pre-Annotated Chinese Treebanks

In this section, based on the annotation schema proposed in previous sections, we start with the manual annotation of the Chinese patent treebank. In Section 3.3.1, we analyse the general sentence structures in Chinese patent claims and present the annotation of certain frequent structures. Section 3.3.2 selects a group of problematic cases during the annotation, in which the direct application of our decision tree is difficult. These cases are subclassified into four subgroups: content terms with unclear internal structure, function words with unclear internal structure, expressions with unclear word boundaries, and unclear syntactic level structure.

All figures of dependency trees (with an index) can be found online in the Arborator annotation project: [https://arboratorgrew.elizia.net/?#/projects/CNPatent/zh\\_patentchar-sud-test](https://arboratorgrew.elizia.net/?#/projects/CNPatent/zh_patentchar-sud-test).

#### 3.3.1 Frequent Syntax Structures Specific to Chinese Patent Claims

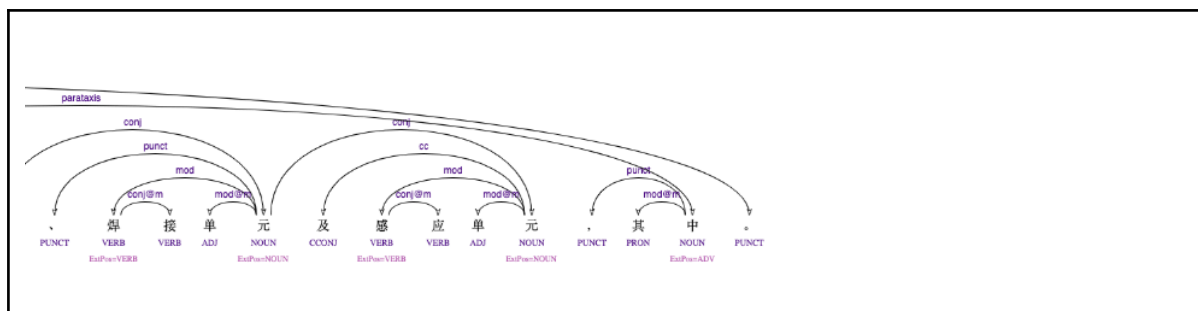
In the patent claims section, sentences are constructed not so much in accordance with modern Chinese grammar rules as they are a combination of specific sentence patterns. Please refer to Table 2.12 in Section 2.5 for further details.

It's important to note that this study does not extensively delve into Chinese syntax analysis but rather focuses on analyzing the linguistic characteristics and structures within patent documents.

	Original Chinese	English Translation	Number
1	其特征在于/其特征是 qí tèzhēng zài yú/qí tèzhēng shì	characterized by	42
2	步骤一/步骤1/步骤S1	Step one/Step 1/Step S1	12
3	(,) 其中(,)	(,) among them(,)	14
4	根据/如权利要求x所述的xxx	According to/As described in claim x, xxx	22
5	一种xxx, 用于xxx	A type of xxx, used for xxx	2
6	一种xxx, 所述xxx	A type of xxx, the described xxx	6
7	一种xxx, 包括xxx	A type of xxx, including xxx	13
8	由xxx组成	composed of xxx	0



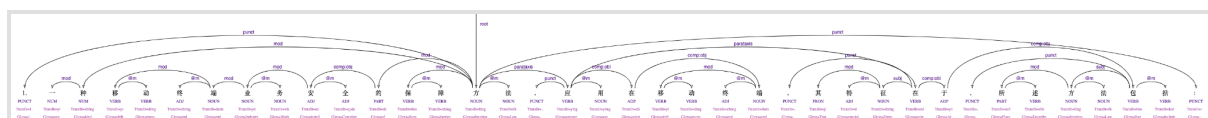




**Figure 3.6 - Example of a Complicated Combination of Frequent Syntactic Patterns in Chinese Patents**

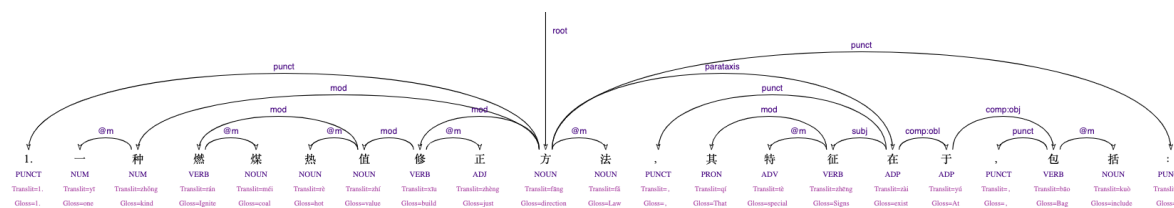
Between phrases starting with “一种 (yī zhǒng, ‘a’), “用于 (yòng yú, ‘for’), and “包括 (bāo kuò, ‘comprising’), we simply annotate them as a “parataxis” relation.

In “其特征在于 (qí tè zhēng zài yú, ‘characterized by’), between “在 (zài)” and “于 (yú), we annotate “comp:obl”; and between “其特征在于” itself and the following syntactic unit, especially the very frequent combination of “其特征在于/其特征是 (qí tèzhēng zàiyú/qí tèzhēng shì, ‘It is characterized by’), (xxx) 包括 (bāo kuò, ‘including’) xxx”, we annotate “comp:obj”.



zh\_patentchar-sud-test\_89

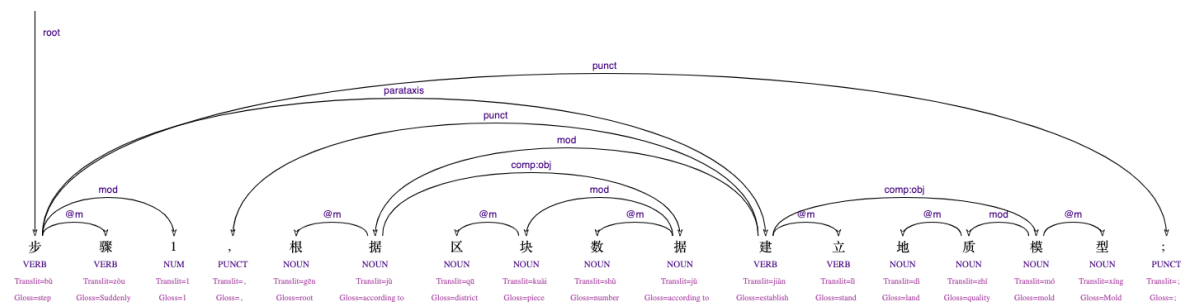
**Figure 3.7 (a)**



zh\_patentchar-sud-test\_118

**Figure 3.7 (b)**

Additionally, the “步骤一/步骤1/步骤S1” is annotated as “appos”.



zh\_patentchar-sud-test\_77

Figure 3.8 (a)



zh\_patentchar-sud-test\_196

Figure 3.8 (b)

### 3.3.2 Practical Choices for Challenging Examples

In this section, we delve into a comprehensive examination of the practical choices made during the annotation process of our patent treebank corpus, specifically focusing on the assignment of labels to challenging examples. These examples encompass a diverse range of linguistic elements, including content and function words, characterized by inherent ambiguity in their internal character relations. Additionally, we encounter words that present considerable difficulty in determining whether they should be assigned a word-internal relation or a conventional syntactic relation. Furthermore, we encounter intricate syntactic structures that pose significant challenges in the annotation process. By addressing these concrete decisions, we aim to shed light on the intricacies involved in annotating such complex linguistic phenomena within the domain of patent texts.

There are three sources for the etymology of the individual terms:

- Federico Masini. *The formation of the modern Chinese lexicon and its evolution toward a national language: the period from 1840 to 1898*. (1993).<sup>105</sup>
- Dong Xiufang. *Lexicalization: the origin and evolution of Chinese disyllabic words*. (2011).<sup>106</sup>
- Zhou Jian. *On Lexicology*. 1st edition (2016).<sup>107</sup>

<sup>105</sup> 马西尼.《现代汉语词汇的形成:十九世纪汉语外来词研究》.(1993).

<sup>106</sup> 董秀芳.《词汇化:汉语双音词的衍生和发展》.(2011).

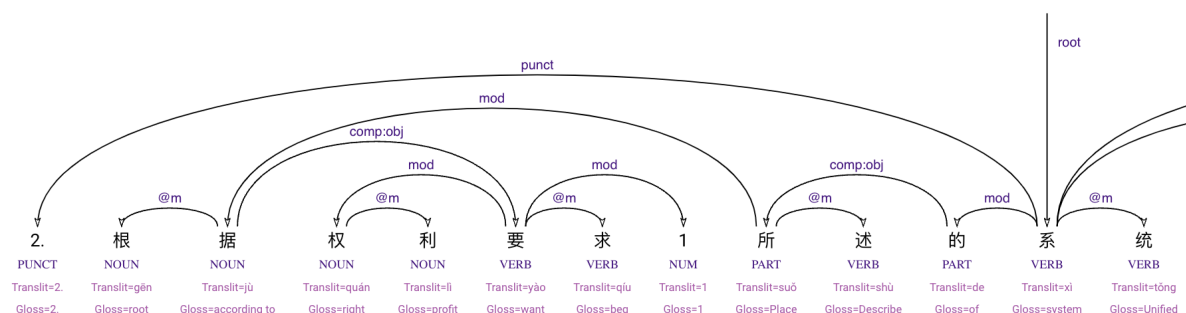
<sup>107</sup> 周荐.《词汇论》.商务印书馆,第一版(2016).

As discussed in Section 3.1.1.1, we strike a balance between etymology and distributional criteria. For many terms with unclear internal structures that were lexicalized a long time ago or are loanwords from other languages, even though their structure is accessible, due to their high degree of lexicalization, we have chosen to annotate them as “flat@m”.

### 3.3.2.1 Content terms with @m

In this section we will present 9 cases where our decision tree does not easily provide a syntactic structure. In these cases, we look into the etymology of the term and try to find paradigms that help us to establish the structure.

#### A. 权利 (quán lì, ‘right’)



zh\_patentchar-sud-test\_73

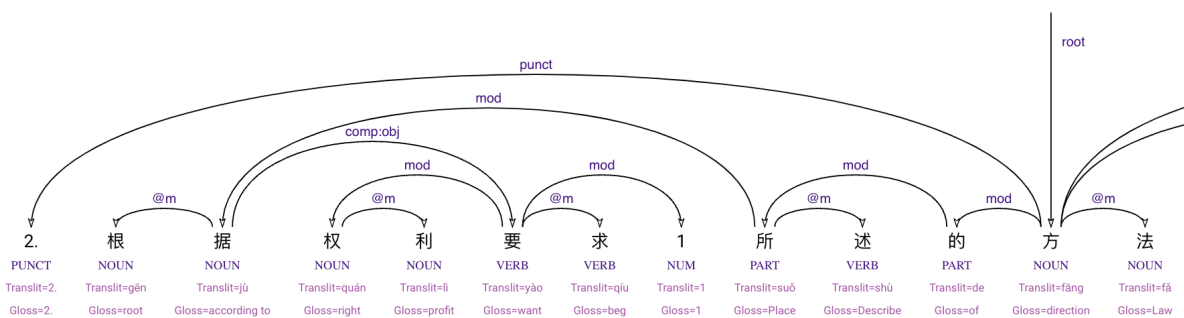
Figure 3.9 (a)

According to Masini (1993), the structure of the term “权利 (quán lì)” is coordinative. In our annotation schema, the word “权利 (quán lì, ‘right’)” is frequently encountered in the context of “权利要求 (quán lì yāo qiú, ‘claim’)”, and we considered it to be an inter-character conjunction relation (conj@m) although it does not fulfil any of the established tests for conj@m according to our decision tree. This is particularly true as there are no evident syntactic criteria applicable to this specific case. Nonetheless, the closest criterion that aligns with this scenario is Test A3, which centres around the semantic criterion. It is noteworthy that “权利 (quán lì, ‘right’)” can be expanded into “权力 (quán lì, ‘power’)” and “利益 (lì yì, ‘interest’)”. This potential expansion aligns with the semantic composition criterion outlined in “Wordhood in Chinese” by San Duanmu (1998).

Overall, our analysis overlaps with the examples that Masini provides for coordinative terms, although he remains rather informal and it is not clear how to systematically extent his analysis beyond the given examples.

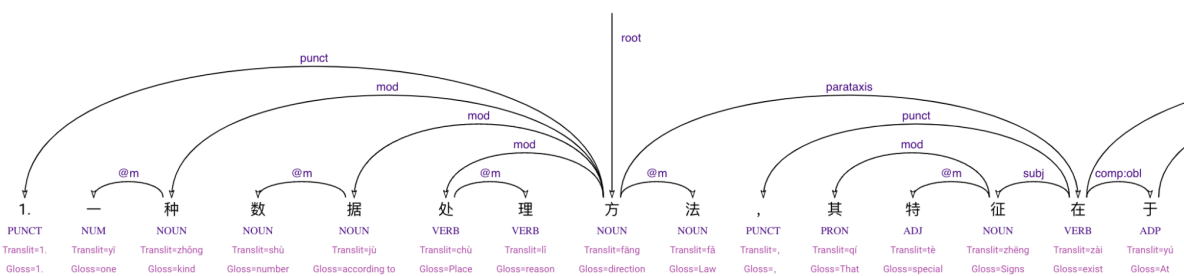
#### B. 根据 (gēn jù, ‘according to’) and 数据 (shù jù, ‘data’)<sup>108</sup>

<sup>108</sup> The term 根据 (gēn jù, ‘according to’) is classified as a function term, and its inclusion in this discussion is crucial due to the insightful comparison with the content term 数据 (shù jù, ‘data’),



zh\_patentchar-sud-test\_139

Figure 3.9 (c)



zh\_patentchar-sud-test\_9

Figure 3.9 (d)

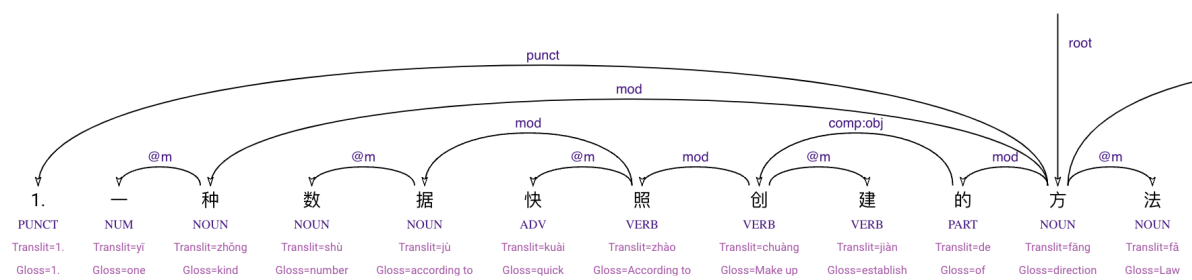
There is a paradigm for the character “据 (jù, ‘evidence\_NOUN; occupy\_VERB’):”:

- 票据 (piào jù, ‘bill’)
- 单据 (dān jù, ‘documents’)
- 证据 (zhèng jù, ‘evidence’)

When considering the inclusion of the term “数据 (shù jù, ‘data’)” in the list, the decision-making process is relatively straightforward. However, the case of “根据 (gēn jù, ‘according to’)” presents a more intricate situation, as it can be interpreted in two distinct ways: “the root evidence” and “the root occupies”. In light of this ambiguity, we have made the deliberate choice to opt for the former interpretation in order to ensure a more generalized and inclusive list, thereby minimizing the proliferation of independent cases. Finally, we decided to assign “mod@m” label to this term.

C. 方法 (fāng fǎ, ‘method’)

which shares the character 据 (jù). By juxtaposing these terms, we can effectively address the challenges associated with character-level labeling in a more comprehensive manner.

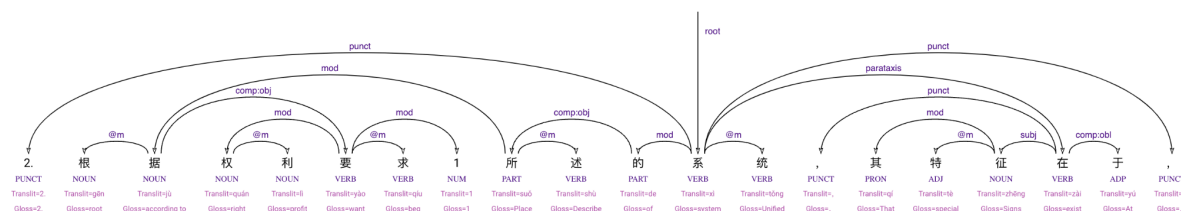


zh\_patentchar-sud-test\_174

Figure 3.9 (e)

Similar to the aforementioned case of “权利 (quán lì, ‘right’), another prominent term in Chinese patents, namely 方法 (fāng fǎ, “method”), can also be decomposed into “方式 (fāng shì, ‘way’)” and “办法 (bàn fǎ, ‘means’)”. Consequently, this specific word is assigned the label conj@m in our annotation schema. Also, according to Masini (1993), this term is the return of an originally Chinese term from Japanese. And during the Tang Dynasty, this term already had its current meaning. This designation aligns with the observed semantic decomposition criterion, wherein the compound term can be disassembled into constituent parts that signify distinct concepts.

D. 特征 (tè zhēng, ‘feature, characteristic’)



zh\_patentchar-sud-test\_73

Figure 3.9 (f)

The term “特征 (tè zhēng, ‘feature, characteristic’)” often appears in the common expression “其特征在于 (qí tèzhēng zàiyú, ‘It is characterized by’)” within Chinese patents. However, it poses an interesting challenge due to its potential for multiple interpretations. Specifically, it can be understood as a composition of the constituents “特点 (tè diǎn, ‘trait’)” and “征迹 (zhēng jī, ‘signs’), or as the adjective “特别的 (tèbié de, ‘special’)” modifying “征迹 (zhēng jī, ‘signs’)”. Given the presence of a paradigmatic list centred around the character “特 (tè, ‘special’), we have taken into account its constituent elements:

- “特点 (tè diǎn, ‘feature, characteristic’), where “点 (diǎn)” carries the meaning of “point” in a literal sense.
- “特性 (tè xìng, ‘feature, characteristic’), where “性 (xìng)” signifies “nature” in its literal sense.

Drawing from this common structural pattern, we have made the discerning choice to assign the relation mod@m to the term “特征 (tè zhēng, ‘feature, characteristic’)”.

E. 位置 (wèi zhì, ‘location’)

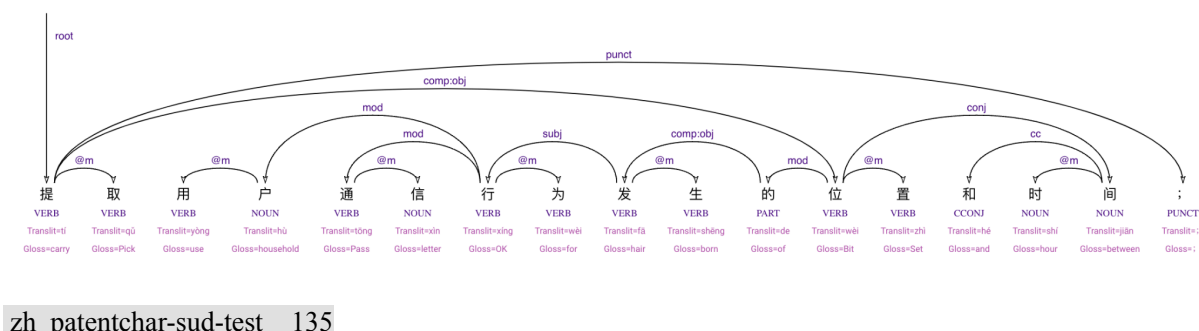


Figure 3.9 (g)

The term “位置 (wèi zhì, ‘location’)” presents a challenge when it comes to assigning a label due to the dual nature of its constituent characters. The first character “位 (wèi, ‘position’)”, when used independently, typically functions as a noun in modern Chinese (e.g. 订了个位, dīng le gè wèi, ‘booked a seat’). However, in classical Chinese and the commonly used expression “位于 (wèi yú, ‘located’)”, it operates as a verb. On the other hand, the character “置 (zhì, ‘put, place’)” is not commonly used independently in modern Chinese but does have the structure “置于 (zhì yú, ‘placed’)”. Considering these complexities, we have made the deliberate decision to assign the label conj@m to the term “位置 (wèi zhì, ‘location’)”.

F. 信息 (xìn xī, ‘information’)

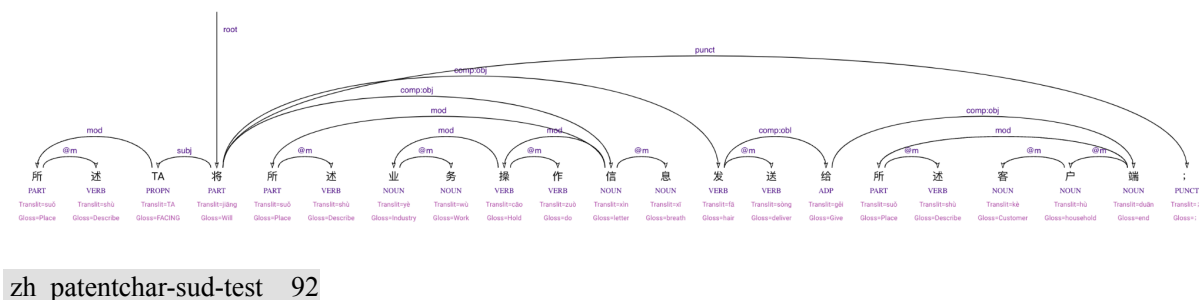


Figure 3.9 (h)

The term “信息 (xìn xī, ‘information’)” poses a challenge in label assignment due to the composition of its constituent characters. It is formed by combining “信 (xìn, ‘letter, message; trust’)” and “息 (xī, ‘breath; interest; message’)”. The difficulty lies in the fact that the character “息 (xī)” when used independently consistently signifies “interest” (e.g., “银行加息了 (yín háng jiā xī le, ‘The bank raised interest rates’)”). It requires a combination of other characters to express other concrete meanings. For instance, “叹息 (tàn xī, ‘sigh’)” conveys the meaning of ‘breath’, while “信息 (xìn xī, ‘information’)” conveys the sense of ‘message’. This example illustrates the challenge of applying syntactic tests to determine inter-character relations in Chinese words. “息 (xī)” exemplifies why it can be difficult to rely solely on syntactic criteria. Conversely, the character “信 (xìn)” is relatively independent and can also

be bi-syllabized to express the same meaning. For instance, “信件 (xìn jiàn)” refers to a ‘letter’, “信息 (xìn xī)” denotes ‘message’, and “信任 (xìn rèn)” signifies ‘trust’.

In line with Test A3, we can explore potential replacements for the term “信息 (xìn xī, ‘information’)” by considering “音信 (yīn xìn, ‘news’)” and “消息 (xiāo xī, ‘information, news, message’)”. Upon examination, it becomes apparent that while the latter, 消息 (xiāo xī), can substitute the original term “信息 (xìn xī)” in certain contexts with a slight shift in meaning, the former, “音信 (yīn xìn)”, is unable to serve as a suitable replacement in this specific case.

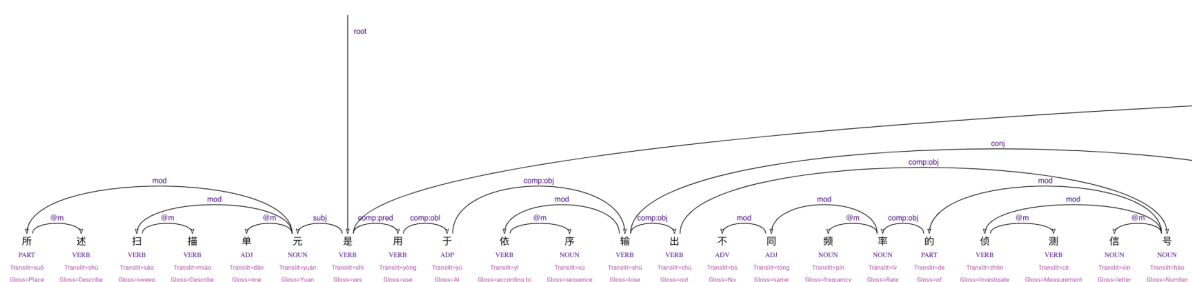
他发了一条信息。  
tā fā le yī tiáo xìn xī  
‘He sent a piece of information.’

他发了一条消息。  
tā fā le yī tiáo xiāo xī  
‘He sent a message.’

\* 他发了一条音信。  
tā fā le yī tiáo yīn xìn  
‘He sent news.’

In the end, we decided to assign “conj@m” to the term “信息 (xìn xī, ‘information’)”.

G. 信号 (xìn hào, ‘signal’)



zh\_patentchar-sud-test\_87

Figure 3.9 (i)

According to Masini (1993), this term is a loanword borrowed from Japanese.

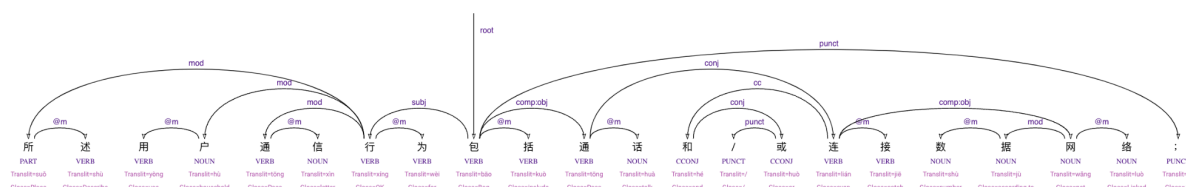
The term “信号 (xìn hào, ‘signal’)” can be annotated in two possible ways: “信 <-mod@m-号” and “信 -conj@m-> 号”. In the first annotation, “信 (xìn, ‘message’)” is treated as a modifier of “号 (hào, ‘signal’)”. Conversely, the second annotation considers both characters



to hold an equal position, functioning in conjunction with each other. After careful consideration, the final choice favours the second annotation, wherein the two characters are treated as conj@m.

This decision is influenced by the presence of a paradigm of “号 (hào, ‘signal’)” within the language. Examples such as “暗号 (àn hào, ‘cypher’)” and “记号 (jì hào, ‘note’)” demonstrate the existence of a common structure involving “号 (hào, ‘signal’)”.

### H. 网络 (wǎng luò, ‘network’)



zh\_patentchar-sud-test\_\_140

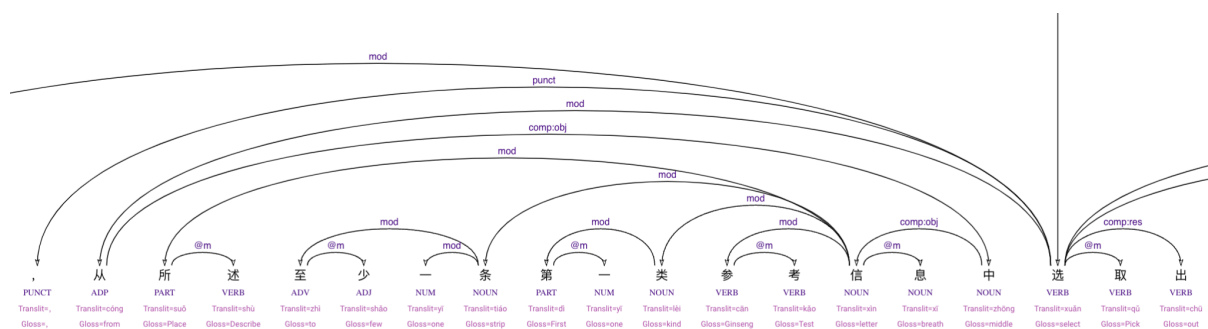
Figure 3.9 (j)

The term “网络 (wǎng luò, ‘network’)” holds significant usage in everyday language, representing a relatively modern concept. The character “网 (wǎng)” independently carries the meaning of ‘net’ and is commonly understood as such. However, the character “络 (luò, ‘network-like thing’)” is less frequently used in isolation. Notably, “络 (luò)” appears in two other compound words: “经络 (jīng luò, referring to ‘meridians’ in the context of Chinese medicine)” and “脉络 (mài luò, denoting ‘arteries and veins; thread of thought’)”. Additionally, the term “络子 (lào zi)” exemplifies a common structure of bi-syllabization, where the suffix “子 (zi)” is appended to the non-lexical character (Huang et Liao, 2012) “络 (lào)<sup>109</sup>”. This construction results in a specific term that signifies a particular type of Chinese knotting, deviating from its literal interpretation as a ‘network-like thing’.

Indeed, when considering the word-internal relation within “网络 (wǎng luò, ‘network’)”, categorizing it as conj@m is supported by these linguistic observations, particularly from a semantic perspective. However, it is crucial to recognize and address the inherent ambiguity that arises when applying semantic criteria. The meanings of terms are dynamic and subject to change over time, and a single term can encompass various senses and even adopt different parts-of-speech, particularly in the case of single characters. In the context of “网 (wǎng, ‘net’), it predominantly maintains a stable meaning and serves as a noun in modern Chinese usage. Conversely, the second character “络 (luò)” carries the meaning of “network-like thing” as a noun, despite its limited occurrence in modern Chinese usage. Interestingly, “络 (luò)” can also function independently as a verb, signifying actions such as “wind, bind,” or “hold something in place with a net” in classical Chinese.

### I. 参考 (cān kǎo, ‘refer to; reference’)

<sup>109</sup> And with the changement of the pronunciation.



zh\_patentchar-sud-test\_17

Figure 3.9 (k)

The term “参考 (cān kǎo, ‘refer to; reference’)” is composed of “参 (cān, ‘join; consider; examine, inspect’)” and “考 (kǎo, ‘examine’)”. It possesses a similar problem with the previous examples “网络 (wǎng luò, ‘network’)” and “信息 (xìn xī, ‘information’)”, which means that it has a duplication of sense of its characters, but is hard to apply to the syntactic tests.

This is also seen as a “conj@m” relation in our annotation schema.

### 3.3.2.2 Function words constructed with @m

Following the analysis of the internal relations within content terms in our patent corpus, we now shift our focus to the function terms that present challenges in delivering character-level labels.

Unlike content terms, where the constituent characters typically possess concrete meanings and ambiguity arises from the uncertainty surrounding the exact meaning of each character when combined, annotating function terms poses distinct difficulties. The complexity of annotating function terms is further compounded by the fact that the difficulty lies in their functional nature rather than the specific meanings of individual characters. Function terms, as their name suggests, serve a grammatical or functional role in the language and often lack concrete semantic interpretations.

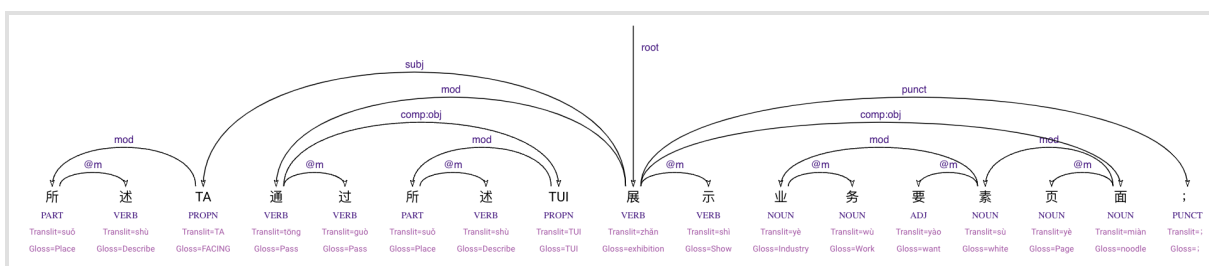
The challenge in annotating function characters stems from the intricate interplay between their syntactic and grammatical functions, making it challenging to assign character-level labels. These characters may exhibit variations in usage, grammatical roles, and syntactic patterns that do not neatly align with the criteria applied to content characters.

Moreover, function characters are often context-dependent and rely heavily on the surrounding linguistic context for their interpretation and understanding. According to Dong (2011), the function terms are often the results of the lexicalization of syntactic structures or cross-layer structures. Their meaning and usage can vary across different syntactic constructions and discourse contexts, further complicating the annotation process.

We annotate the internal structure of these function terms as flat@m (example C-H), except for several exceptions (example A and B).

The first exception is the “所 (suǒ)” structure, in which “所 (suǒ)” is a special mark used before a verb, and represents the object receiving the action. As a very productive character, there exists a large paradigm of “所 (suǒ) + V” in modern Chinese. For this type, we annotate them as “comp@m” relation with “所 (suǒ)” as the head.

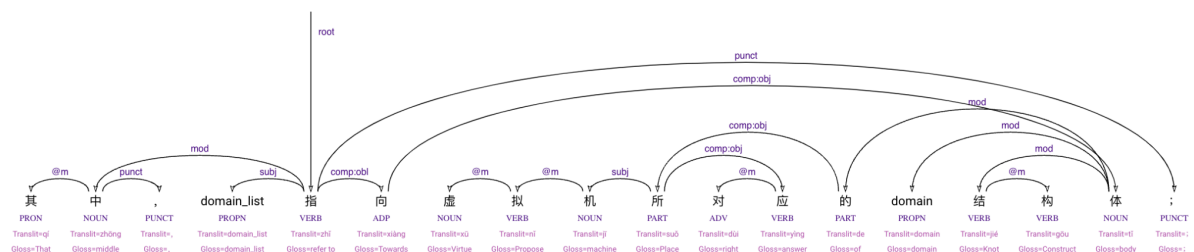
A. 所            述  
 suǒ            shù  
 PART            ‘discrcribe\_V’  
 ‘Described\_ADJ’



zh\_patentchar-sud-test\_95

Figure 3.10 (a)

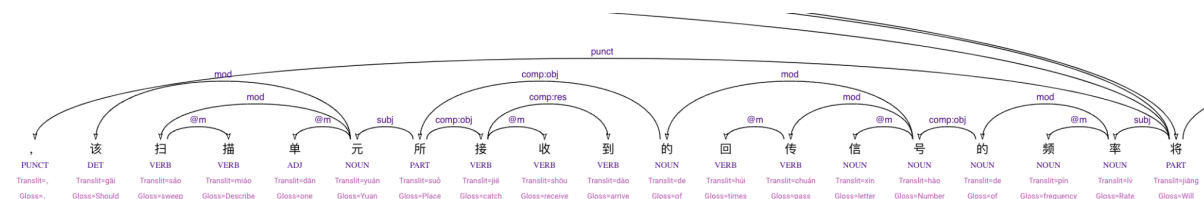
所            对            应  
 suǒ            duì            yīng  
 PART            ‘discrcribe\_V’  
 ‘Described\_ADJ’



zh\_patentchar-sud-test\_166

Figure 3.10 (b)

所            接            收            到  
 suǒ            duì            yīng            dào  
 PART            ‘discrcribe\_V’  
 ‘Described\_ADJ’

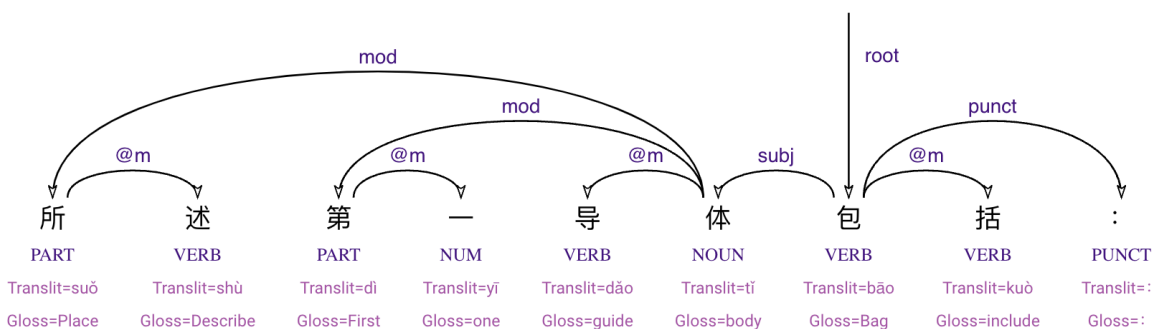


zh\_patentchar-sud-test\_88

Figure 3.10 (c)

The second exception is “第 (dì)”, which is a special mark that transforms the number into an ordinal number. We annotate them as “comp@m” in relation to “第 (dì)” as the head.

B. 第 一  
 dì yī  
 PART ‘one\_NUM’  
 ‘First\_NUM’

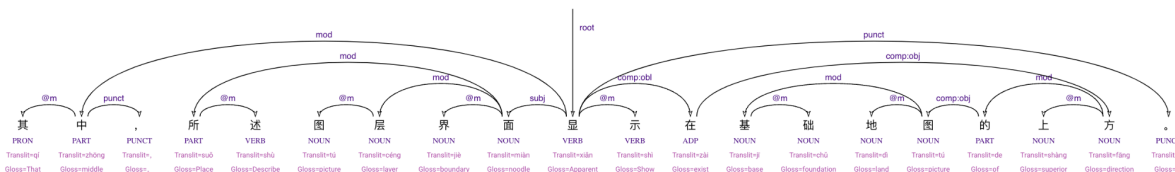


zh\_patentchar-sud-test\_4

Figure 3.10 (d)

There are also examples that we annotate as “flat@m”:

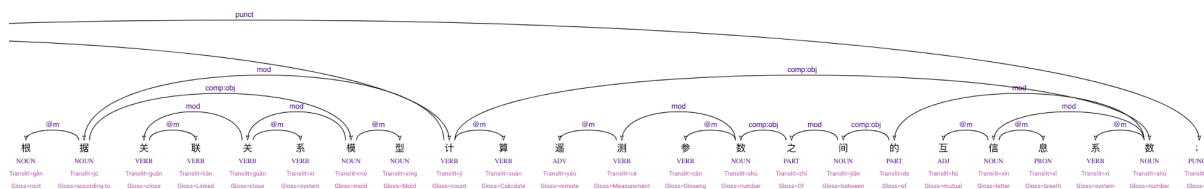
C. 其 中  
 qí zhōng  
 ‘Among (them)\_ADP’



zh\_patentchar-sud-test\_131

Figure 3.10 (e)

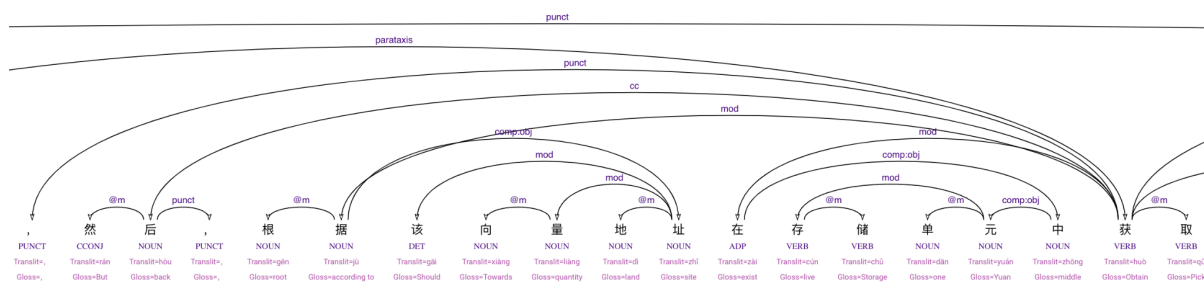
D. 之 间  
zhī jiān  
“Between\_ADP”



zh\_patentchar-sud-test\_101

Figure 3.10 (f)

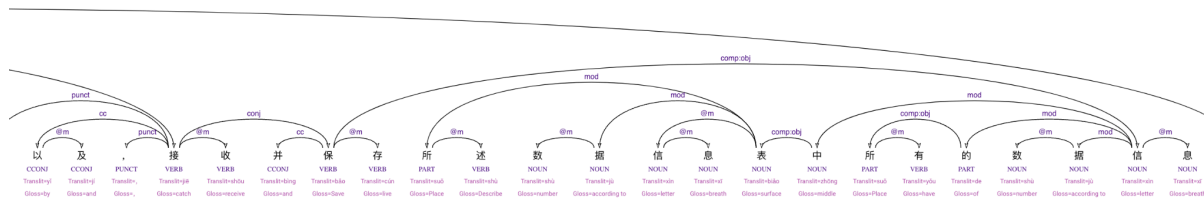
E. 然 后  
rán hòu  
“Then\_ADV”



zh\_patentchar-sud-test\_57

Figure 3.10 (g)

F. 以 及  
yǐ jí  
“And\_CONJ”



zh\_patentchar-sud-test\_74

Figure 3.10 (h)

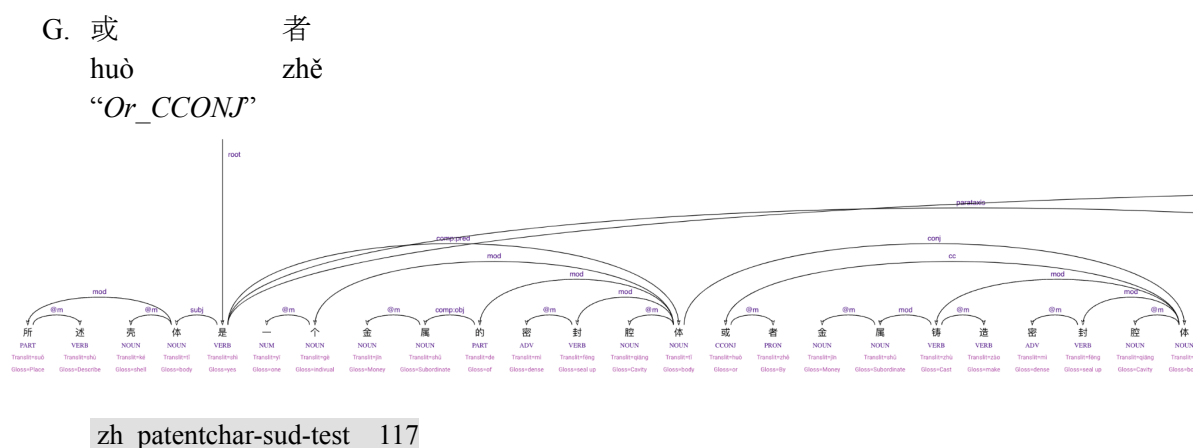


Figure 3.10 (i)

### 3.3.2.3 Choices involving the assignment of @m

In addition to annotating the internal structure of terms, there are instances where assigning “@m” is not straightforward. This hesitation arises primarily in two categories: Resultative/Directional complements and the “V + 于(yú)” structure, making it unclear whether these examples should be considered as internal relations or conventional syntactic relations.

#### A. Resultative/Directional complements

The first problematic structure is the resultative and directional complements, which is composed by a head verbal term (unigram or bigram) and a complement indicating the result or direction of the verb. The annotation can be regarded as a whole term or split then into two terms.

Here are a few examples:

#### - Resultative

- |     |              |             |             |             |
|-----|--------------|-------------|-------------|-------------|
| (1) | 洗            | 净           |             |             |
|     | xǐ           | jìng        |             |             |
|     | ‘wash_V’     | ‘clean_ADJ’ |             |             |
|     |              |             |             |             |
| (2) | 洗            | <u>干</u>    | <u>净</u>    |             |
|     | xǐ           | gān         | jìng        |             |
|     | ‘wash_V’     | ‘dry_ADJ’   | ‘clean_ADJ’ |             |
|     |              |             |             |             |
| (3) | <u>清</u>     | <u>洗</u>    | <u>干</u>    | <u>净</u>    |
|     | qīng         | xǐ          | gān         | jìng        |
|     | ‘clean up_V’ | ‘wash_V’    | ‘dry_ADJ’   | ‘clean_ADJ’ |

(4) 清 洗 完  
 qīng xǐ wán  
 ‘clean up\_V’ ‘wash\_V’ ‘finish\_V’

- Directional

(5) 输 出  
 shū chū  
 ‘transport\_V’ ‘get out\_V’  
 ‘Output\_N/V’

(6) 输 出 去  
 shū chū qù  
 ‘output\_V’ ‘go\_V’

Table 3. lists a group of the directional complement characters, which can further combine with the two character at right “来 (lái)” and “去 (qù)”<sup>110</sup>.

In observation, we can see that the resultative/directional complement structures may be in the form of 1+1, 1+2, 2+2 or 2+1.

We adopt the Syllable Count principle from “Wordhood in Chinese” by San Duanmu (1998) for these compliments. Only when the struction is in the form of 1+1, we annotate it as an internal relation, else for 1+2, 2+2 and 2+1, it would be a syntactic relation between the two segments.

Directional Complement Character	
上 (shàng) - up, above, on, go up 下 (xià) - down, below, go down 进 (jìn) - enter, go in, advance 出 (chū) - go out, exit, leave 回 (huí) - return, go back 过 (guò) - pass, cross, go over 起 (qǐ) - rise, get up, start	来 (lái) - come, arrive 去 (qù) - go, leave

**Table 3.6 - The combination of the Directional Complement Characters**

<sup>110</sup> The character “起 (qǐ)” can only combine with “来 (lái)” and not “去 (qù)”.

B. “V + 于(yú)” structure

When examining the five examples below, we observe a structure that combines a verb with the preposition “于(yú).” The verb itself can be either a single character (unigram) or two characters (bigram).

- (7) 基                    于  
 jī                      yú  
 ‘base\_V’              ‘on\_PREP’  
 ‘Based on\_V/PREP’
- (8) 介                    于  
 jiè                     yú  
 ‘between\_V’        ‘between\_PREP’  
 ‘Lie between\_V/PREP’
- (9) 位                    于  
 wèi                    yú  
 ‘locate\_V’          ‘at\_PREP’  
 ‘Be located at\_V’
- (10) 用                    于  
 yòng                   yú  
 ‘use\_V’                ‘for\_PREP’  
 ‘Use for\_V’
- (11) 放                    置                    于  
 fàng                    zhì                    yú  
 ‘put\_V’                ‘position\_V’        ‘at\_PREP’  
 ‘Be placed at\_V’

Samely as the resultative/directional complement structures, we apply the Syllable Count principle to this structure, which means when the verb is a unigram we annotate it as “@m”, and when the verb is bigram, we annotate it as a conventional syntactic relation..

3.3.2.4 Problematic syntactic structures

Up to here we have looked into cases where the word-internal structure is difficult to establish. In this remaining subsection we will discuss cases where the intra-word structure is challenging.

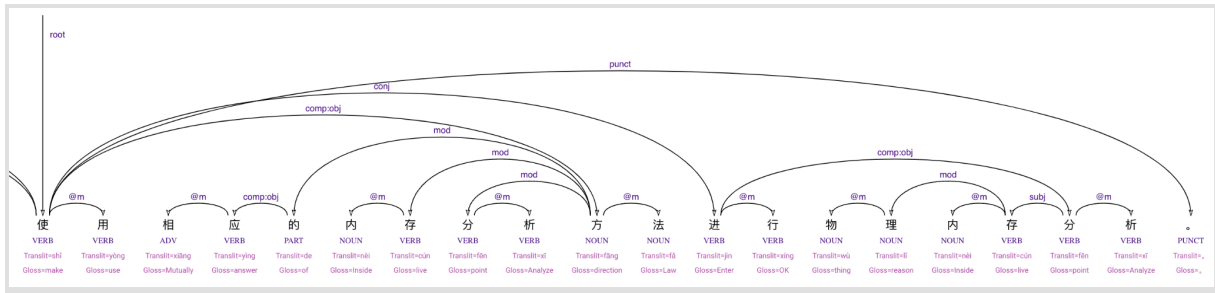
A. Serial Verb Constructions





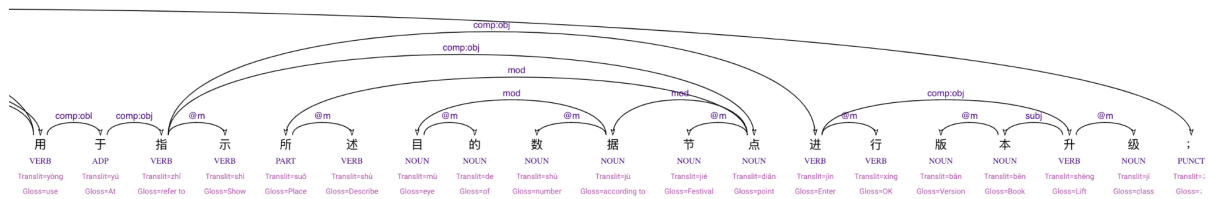






zh\_patentchar-sud-test\_173

Figure 3.11 (k)

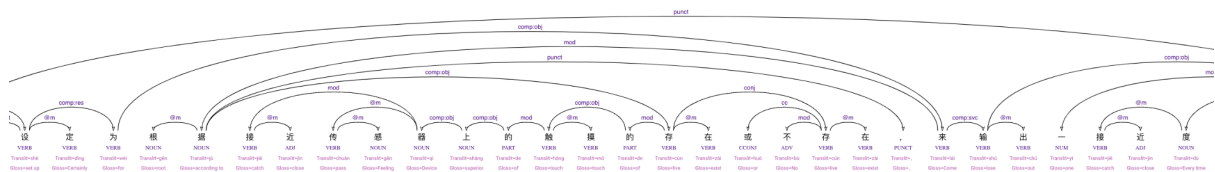


zh\_patentchar-sud-test\_63

Figure 3.11 (l)

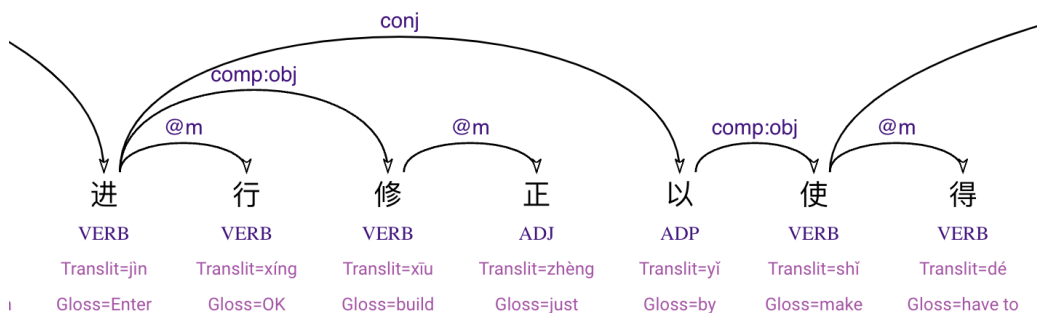
B. VERB + “来 (lái) / 以 (yǐ)” + VERB

In this structure, two verbs is connected by a character “来 (lái)” or “以 (yǐ)”, indicating ‘in order to’. In our schema we annotate this relation as “comp:obl”.



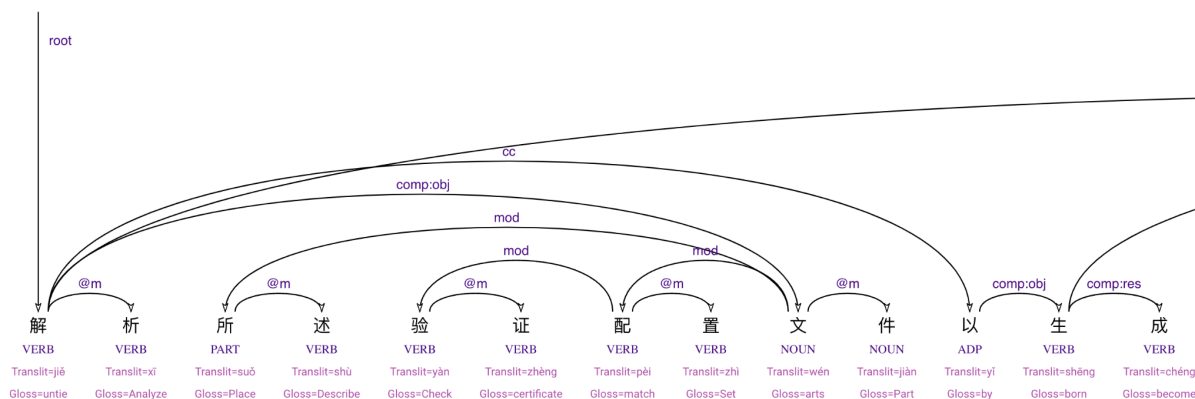
zh\_patentchar-sud-test\_50

Figure 3.11 (m)



zh\_patentchar-sud-test\_14

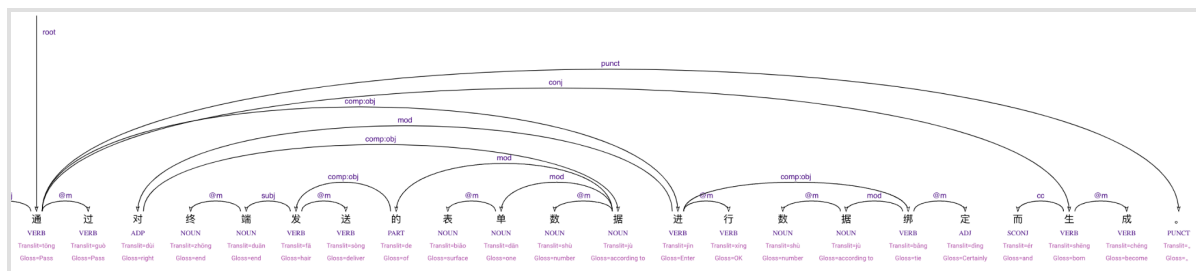
Figure 3.11 (n)



zh\_patentchar-sud-test\_22

Figure 3.11 (o)

Sentence 26 has a similar structure to sentence 22 above: “而生成 (ér shēng chéng)”, in which “而 (ér)” is a conjunction, and in this case the relation between the verbs is “conj”.



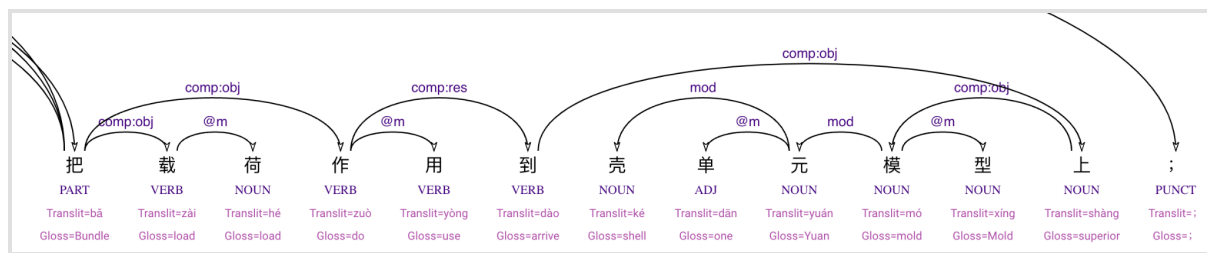
zh\_patentchar-sud-test\_26

Figure 3.11 (p)

### C. “把/将/被” Structure

From a linguistic perspective, “把 (bǎ)”, “被 (bèi)”, and “将 (jiāng)” are three very important syntactic structure particles in Chinese, and they play special roles in sentences:

- “把 (bǎ)”: It represents the receiver of the action in an active sense. The usage of “把 (bǎ)” involves placing the object before the verb, emphasizing that the doer of the action has done something to the object. For example, in “我把书读完了 (wǒ bǎ shū dú wán le, ‘I finished reading the book’)”, “书 (shū, ‘book’)” is the object, “我 (wǒ, ‘I’)” is the doer of the action, and “把” emphasizes that “I” completed the action of reading the “书 (shū, ‘book’)”.

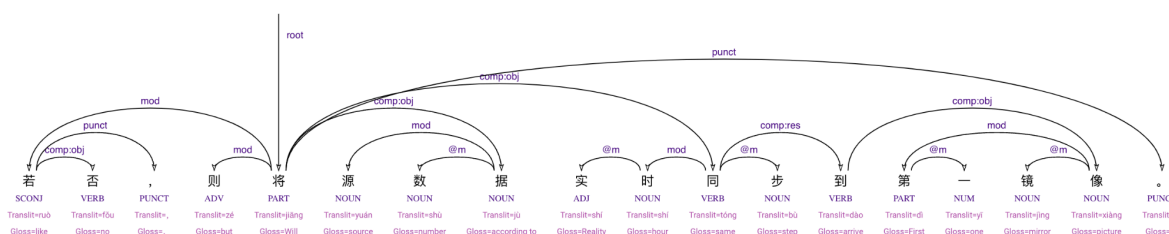


zh\_patentchar-sud-test\_\_193

Figure 3.11 (q)

In the example above, “把 (bǎ)” has two arguments in the form of subtrees: “荷载 (hè zài, ‘loads’)” and “作用到壳单元模型上 (zuò yòng dào ké dān yuán mó xíng shàng, ‘acting on shell cell models’)”.

- “将 (jiāng)”: Similar to “把,” it is also used to represent the receiver of the action in an active sense. This word is typically used to introduce the object in a sentence, placing the object before the verb and emphasizing that the doer of the action has done something to the object. For example, in “我将书读完了 (wǒ jiāng shū dú wán le, ‘I finished reading the book’),” “书 (shū, ‘book’)” is the object, “我 (wǒ, ‘I’)” is the doer of the action, and “将 (jiāng)” emphasizes that “我 (wǒ, ‘I’)” completed the action of reading the “书 (shū, ‘book’)”. “将 (jiāng)” and “把 (bǎ)” are both used to emphasize the doer of the action in an active sense, but “将 (jiāng)” is more commonly used to indicate the future tense or ongoing actions, while “把 (bǎ)” is used more frequently to indicate the past, present, and future tenses.

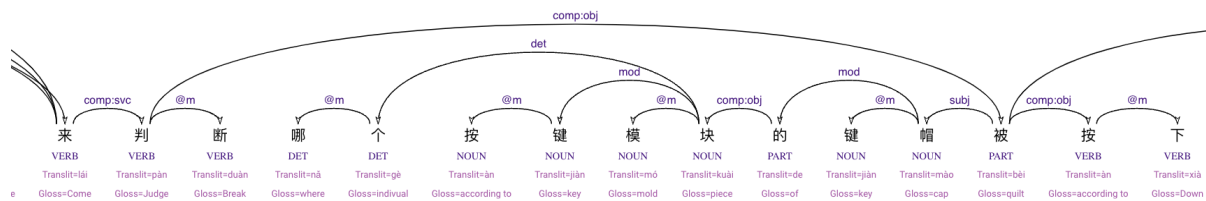


zh\_patentchar-sud-test\_\_182

Figure 3.11 (r)

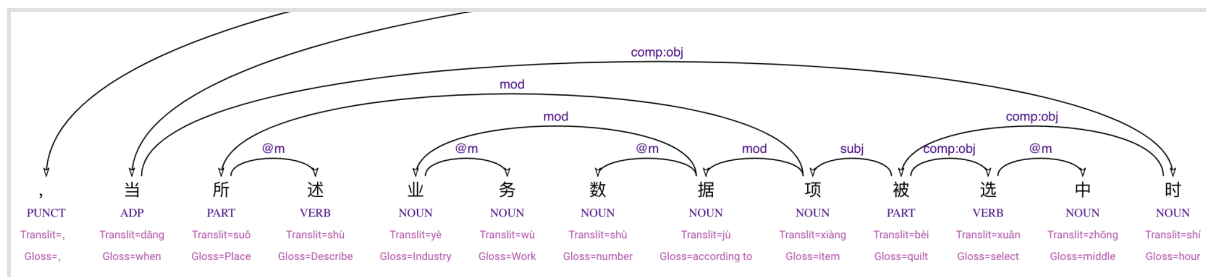
Like the example of “把 (bǎ),” “将 (jiāng)” also has two arguments: “源数据 (yuán shù jù, ‘source data’)” and “实时同步到第一镜像 (shí shí tóng bù dào dì yī jìng xiàng, ‘real-time synchronization to the first mirror’)”.

- “被 (bèi)” is used to emphasize the recipient of a passive action, unlike “把 (bǎ)” and “将 (jiāng)”. “被 (bèi)” is used to emphasize that the object is the receiver of the action, not the doer. For example, in the sentence “书被我读完了 (shū bèi wǒ dú wán le, ‘The book is read/finished by me’),” “书 (shū, ‘book’)” is the object, “我 (wǒ, ‘I’)” is the doer of the action, but “被 (bèi)” emphasizes that “书 (shū, ‘book’)” is the one being read by “我 (wǒ, ‘I’),” highlighting the passive nature of the action.



zh\_patentchar-sud-test\_87

Figure 3.11 (s)



zh\_patentchar-sud-test\_130

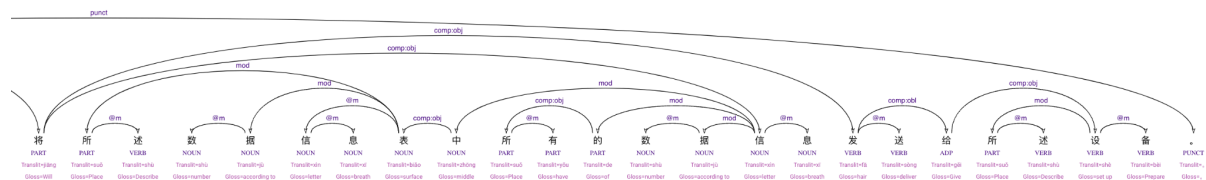
Figure 3.11 (t)

In the Chinese patent treebank, all of these three characters are annotated as a particle (“PART”) head with two arguments: an object (labelled as “comp:obj”) and an action (also “comp:obj”) for “把 (bǎ)” and “将 (jiāng)”; while for “被 (bèi)” is a subject (“subj”) and an action (“comp:obj”).

D. “给 (gěi)”

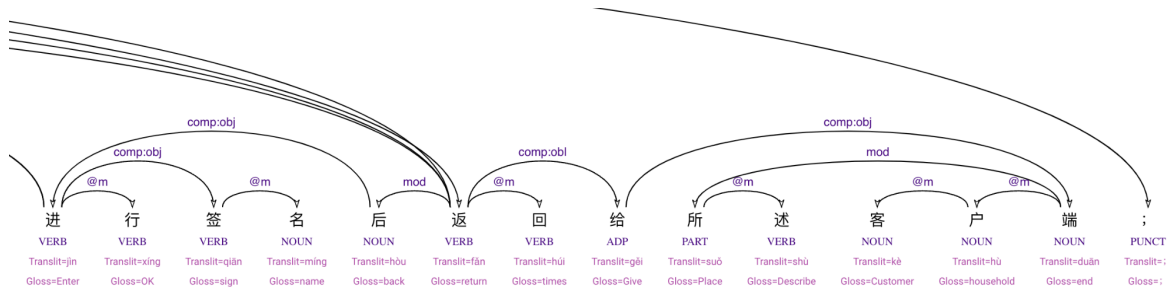
Another annotation challenge arises with the term “给 (gěi)”, making it difficult to determine whether it functions as a verb or a preposition. In the patent treebank, there are four instances of “给 (gěi)”, where the character “给 (gěi)” is part of the structures “发送给 (fā sòng gěi, ‘send to’)” or “返回给 (fǎn huí gěi, ‘reply to’)”.

In this context, we have categorized “给 (gěi)” as an adposition (“ADP”).



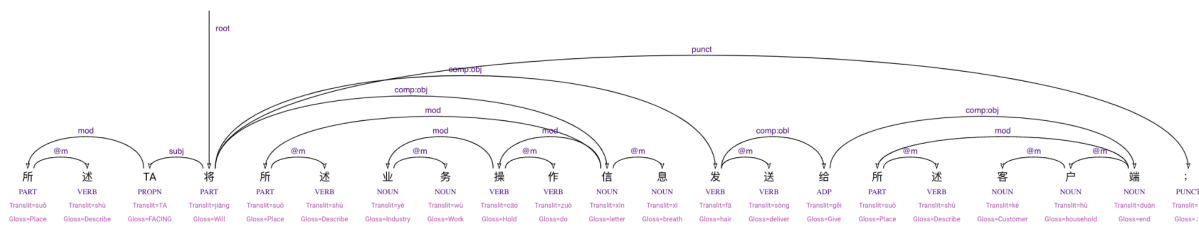
zh\_patentchar-sud-test\_75

Figure 3.11 (u)



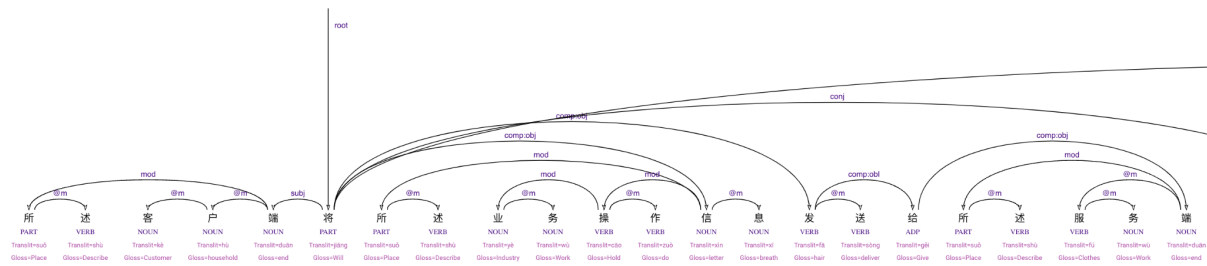
zh\_patentchar-sud-test\_90

Figure 3.11 (v)



zh\_patentchar-sud-test\_92

Figure 3.11 (w)

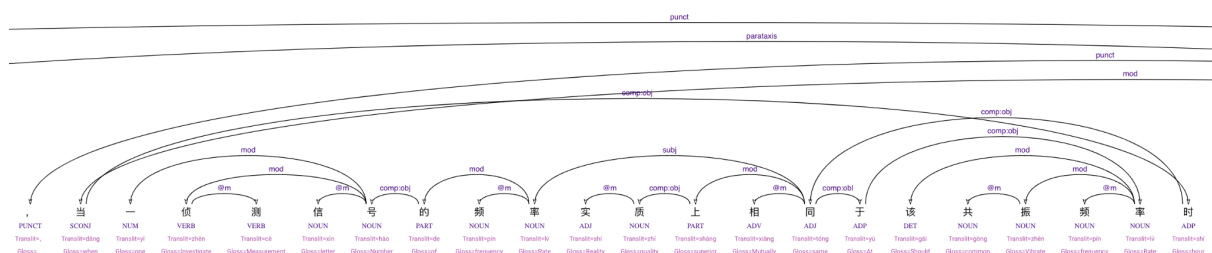


zh\_patentchar-sud-test\_93

Figure 3.11 (x)

E. The structure “(当) ... 时”

The structure “(当(dāng)) ... 时(shí)” in English can be translated as ‘when’ or ‘during’. It is used to indicate a specific point in time or a period during which an action or event takes place. In this structure, “当(dāng)” can be omitted, like in the second example (sentence 88).









### 3.4 The first treebank on Chinese patent claims at character-level

As the final result of the annotation, a first character-level treebank on Chinese patent claims composed of 200 sentences is released on GitHub as a part of the SUD project on the following address: [https://github.com/surfacesyntacticud/mSUD\\_Chinese-PatentChar/tree/main](https://github.com/surfacesyntacticud/mSUD_Chinese-PatentChar/tree/main).

The SUD treebank contains 200 sentences and 8,175 tokens. The number of each type of label at both character-level and word-level is presented in Tables 3.7.

<b>Label</b>	<b>Number</b>
mod@m	1,373
conj@m	1,048
comp@m	616
comp:obj@m	69
comp:res@m	41
flat@m	22
comp:aux@m	9
subj@m	8

**Table 3.7 (a) - The Number of Character-level Labels in the Annotated Treebank**

mod	2,190
comp:obj	987
punct	560
conj	139
parataxis	138
comp:res	39
comp:svc	21
comp:pred	6
det	6
comp:aux	4
comp:dir	1

**Table 3.7 (b) - The Number of Word-level Labels in the Annotated Treebank**

In order to examine the applicability of the annotation guideline, we calculate the inter-annotator score of 10 sentences with help of Wu Qishen, a PhD student from laboratory MoDyCo<sup>112</sup>. The inter-annotator score is shown in Table 3.7 below.

<b>UAS</b>	97.88%
<b>LUS</b>	96.54%
<b>LAS</b>	95.63%

**Table 3.8 - The Inter-annotator Score**

The treebank is also converted into conventional UD format with Grew-match<sup>113</sup>.

1. We first combine each “@m” relation and merge the characters into one token;
2. Then the ExtPos is used as UPOS for the terms that have been combined;
3. For every syntactic relation label in SUD, we find its correspondence in UD, with a script deciding automatically when there are multiple correspondences.



There is a correspondence between the labels used in SUD and the labels used in UD. Below is a table of comparison of these two formats<sup>114</sup>.

---

<sup>112</sup> <https://modyco.fr/welcome/>

<sup>113</sup> <https://match.grew.fr/>

<sup>114</sup> <https://surfacesyntacticud.github.io/conversions/>

	
nsubj	subj
csubj	
<b>aux</b>	comp:aux
<b>cop</b>	comp:pred
xcomp	comp:obj
<b>case</b>	
<b>mark</b>	
obj	
ccomp	
ccomp	comp:obl
obl	
iobj	
nmod	udep
obl, acl	mod
advcl	
advmod	
amod	
nummod	
fixed	
det	det
nummod	

**Table 3.12 - The Correspondence between UD (left column) and SUD (right column) Labels** <sup>115</sup>

This first version contains 100 sentences and 2160 tokens. The first version of the converted treebank is published at [https://github.com/UniversalDependencies/UD\\_Chinese-PatentChar](https://github.com/UniversalDependencies/UD_Chinese-PatentChar). And there is a new version containing 200 sentences that will be released soon.

The converted UD treebank uses 15 UPOS tags out of 17 possible: ADJ, ADP, ADV, AUX, CCONJ, DET, NOUN, NUM, PART, PRON, PROPN, PUNCT, SCONJ, VERB, X.

And it used 21 relation types out of 37 possible: acl, advcl, advmod, amod, appos, case, cc, ccomp, conj, csubj, dep, goeswith, mark, nmod, nsubj, nummod, obj, obl, parataxis, punct, root.

<sup>115</sup> <https://surfacesyntacticud.github.io/conversions/>

## Chapter 4 - Joint Segmentation/ Dependency Parsing of Chinese Patent Claims

The world of dependency parsing and word segmentation has advanced considerably since my first attempts to parse Chinese patents during my Master’s thesis in 2018. At the time we showed that both the traditional word segmentation and dependency parsing do not have acceptable results on patent claim texts.

Ever since the inception of the initial SIGHAN Bakeoff shared task on Chinese word segmentation back in 2003, numerous Chinese word segmentation standards have emerged over time. It wasn't until Bakeoff-4 in 2008, as documented by Jin and Chen, that the landscape saw the presence of seven distinct word segmentation conventions (Li & al., 2022).

Meng and al. (2019) argue that *“segmenting a chunk of text into words is usually the first step of processing Chinese text, but its necessity has rarely been explored.”* And they find that char-based models consistently outperform word-based models in four end-to-end NLP benchmark tasks.

In this section, we revisit the concept of “words” from a posterior standpoint. By training the dependency parser on a corpus at the character-level, we are able to automatically derive “words” through the assignment of the “@m” label. We present our work on developing a joint segmenter-parser based on our character-level treebanks annotated in Chapter 3.

In Section 4.1, we focus on the previous works on syntax parsing in Chinese, especially on the joint segmentation-parsing method (or the syntax parsing on character-level).

Then, in Section 4.2 we fine-tune a Bert-based dependency parser on our patent claim treebank. By combining all the “@m” relations, the dependency parser also serves as a word segmenter. Based on the evaluation scores on both syntactic and morphological levels, we manually analyse the errors during parsing at the end of the second section.

Section 4.3 explores the usefulness the dependency parsing for term extraction. After a brief presentation of the automatic term extraction (ATE), we present our method of the extraction of candidate terms. We also compare the results with other methods, such as the POS pattern extraction.

## 4.1 Automatic Syntax Analysis in Chinese

Language is a complex and dynamic system that serves as a medium for human communication. It consists of not only individual words and their meanings but also intricate rules governing their arrangement and relationships. Syntactic analysis, a fundamental branch of linguistics and natural language processing (NLP), plays a crucial role in deciphering the structural aspects of language. In this section, we delve into the world of syntactic analysis, exploring its importance, methods, and practical applications.

At its core, syntactic analysis aims to uncover the grammatical structure of sentences, shedding light on how words are combined to convey meaning. It focuses on identifying syntactic units within sentences, such as nouns, verbs, adjectives, and their interconnections. Moreover, syntactic analysis examines the hierarchical relationships between these units, paving the way for a more profound comprehension of language.

Syntactic analysis serves as a foundational tool for content-based text analysis. By identifying the structure of a sentence, it helps NLP systems understand the roles of various elements and their contributions to the overall meaning. This, in turn, enhances the accuracy of applications such as machine translation, sentiment analysis, and information extraction. Linguists use syntactic analysis to investigate the grammatical rules that govern languages. This research contributes to our understanding of the universality of certain syntactic structures and the uniqueness of others in specific languages or language families.

Syntactic analysis employs two primary approaches: rule-based methods and statistical methods.

- **Rule-Based Methods:** Rule-based approaches involve manually constructing grammatical rules to analyze sentences. While effective for simple sentences, they may fall short when dealing with the complexities of real-world text. Grammar rule coverage can be limited, and these methods are often less portable across different systems.
- **Statistical Methods:** Statistical models for syntactic analysis have gained prominence with the availability of large-scale treebanks (such as the Penn Treebank, the Tsinghua Treebank, and the Academia Sinica Treebank in Taiwan mentioned in Section 1.2.1.4) and advances in machine learning. These models assign scores to candidate syntactic structures and select the most likely structure as the result. They have achieved remarkable success in various NLP tasks.

Two predominant grammatical frameworks have garnered substantial attention in linguistic research. Context-free grammar (CFG), also recognized as constituent parsing or phrase-structure parsing, employs hierarchical phrase-structural trees to organize syntactic information at the sentence level. This approach has been researched intensively since very early (Chomsky, 1956).

On the other hand, dependency grammar represents another widely embraced framework for syntactic parsing. It establishes direct connections between words through dependency links, accompanied by labels denoting their syntactic relationships. Due to its compactness and straightforward annotation process, dependency parsing has garnered greater prominence compared to constituent parsing.

In the realm of constituent parsing, mainstream methods encompass chart-based (Zhou & al., 2018; Zan & al., 2020) and transition-based models (Watanabe et Sumita, 2015; Wang & al. 2018). Notably, contemporary neural models have achieved state-of-the-art performance in both of these approaches.

Parsing scores for Chinese tend to be lower in comparison to European languages. In Zhang Meishan’s study (2020), a range of state-of-the-art models were compared for both English and Chinese, employing both constituent grammar (Figure 4.1) and dependency grammar (Figure 4.2). The results showed that while phrase-level F1 scores for constituent parsers on the Penn Treebank (PTB) dataset typically ranged between 90% and 95%, the corresponding scores on the Chinese Treebank (CTB, see Section 1.2.1.4) dataset were between 85% and 90%. Additionally, similar differences were observed in terms of UAS (Unlabeled Attachment Score) for dependency parsers.

Model	Main Features	PTB	CTB
<b>Chart-based, Statistical Models</b>			
Collins (1997) [9]	head-lexicalization	88.2	N/A
Charniak (2000) [10]	max-entropy	89.5	80.8
McClosky et al. (2006) [11]	self-training	<b>92.3</b>	N/A
Petrov and Klein (2007) [12]	PCFG	90.1	<b>83.3</b>
Hall et al. (2014) [13]	CRF	89.9	N/A
<b>Transition-based, Statistical Models</b>			
Sagae and Lavie (2005) [14]	greedy	86.0	N/A
Zhu et al. (2013) [15]	global learning, beam	<b>91.3</b>	<b>85.6</b>
<b>Chart-based, Neural Models</b>			
Socher et al. (2013) [16]	recursive NN	90.4	N/A
Durrett and Klein (2015) [17]	CNN	91.1	N/A
Stern et al. (2017) [18]	LSTM, span	91.8	N/A
Kitaev and Klein (2018) [19] (a)	self-attentive	93.5	N/A
Kitaev and Klein (2018) [19] (b)	+ELMo	<b>95.1</b>	N/A
<b>Transition-based, Neural Models</b>			
Wang et al. (2015) [20]	neural+discrete	90.7	<b>86.6</b>
Watanabe and Sumita (2015) [21]	global learning, beam	90.7	N/A
Dyer et al. (2016) [22]	language modelling	92.4	82.7
Cross and Huang (2016) [23]	dynamic oracle	91.3	N/A
Liu and Zhang (2017) [24]	in-order	91.8	86.1
Fried and Klein (2018) [25]	policy gradient	92.6	86.0
Kitaev and Klein (2019) [26]	policy gradient	<b>95.4</b>	86.0
<b>Other Methods (report neural models only)</b>			
Shen et al. (2018) [27]	distance to tree	91.8	86.5
Teng and Zhang (2018) [28]	local classification	92.7	87.3
Vilares et al. (2019) [29]	sequence labeling	91.1	85.6
Zhou and Zhao (2019) [30]	HPSG grammar	<b>96.3</b>	<b>92.2</b>
Mrini et al. (2019) [31]	HPSG, improved attention	<b>96.3</b>	N/A

Table 4.1 - (Zhang, 2020)



Model	Main Features	PTB	CTB
<b>Graph-based, Statistical Models</b>			
McDonald et al. (2005) [57]	1-order	90.9	83.0
McDonald and Pereira (2006) [63]	2-order	91.5	85.2
Koo et al. (2008) [64]	word clusters	93.2	N/A
Chen et al. (2009) [65]	auto subtrees	93.2	86.7
Bohnet (2010) [66]	feature hashing	92.9	N/A
Koo and Collins (2010) [67]	3-order	93.0	86.0
Ma and Zhao (2012) [68]	4-order	<b>93.4</b>	<b>87.4</b>
<b>Transition-based, Statistical Models</b>			
Nivre (2008) [69] (a)	arc-standard	89.7	82.7
Nivre (2008) [69] (b)	arc-eager	89.9	80.3
Zhang and Clark (2008) [70]	global learning, beam	91.4	84.3
Zhang and Nivre (2011) [71]	rich non-local features	<b>92.9</b>	<b>86.0</b>
Goldberg and Nivre (2012) [42]	dynamic oracle	91.0	84.7
<b>Graph-based, Neural Models</b>			
Pei et al. (2015) [72]	feed-forward	93.3	N/A
Zhang et al. (2016) [73]	CNN	93.4	87.7
Wang and Chang (2016) [74]	2-layer LSTM	94.1	87.6
Kiperwasser and Goldberg (2016) [75]	2-layer LSTM	93.1	86.6
Dozat and Manning (2016) [76]	3-layer LSTM, biaffine	95.7	88.9
Li et al. (2019) [77] (a)	self-attentive	95.9	92.2
Li et al. (2019) [77] (b)	+ELMO	96.6	90.3
Li et al. (2019) [77] (c)	+BERT	<b>96.7</b>	<b>92.2</b>
Ji et al. (2019) [78]	GNN	96.0	N/A
<b>Transition-based, Neural Models</b>			
Chen and Manning (2014) [79]	feed-forward	91.8	83.9
Dyer et al. (2015) [80]	stack-LSTM	93.1	87.2
Zhou et al. (2015) [81]	global learning, beam	93.3	N/A
Andor et al. (2016) [82]	global learning, beam	94.6	N/A
Kiperwasser and Goldberg (2016) [75]	2-layer LSTM	93.9	87.6
Ballesteros et al. (2017) [83]	char, stack-LSTM	93.6	87.6
Ma et al. (2018) [84]	3-layer LSTM	<b>95.9</b>	<b>90.6</b>
<b>Other Methods (report neural models only)</b>			
Kiperwasser and Goldberg (2016) [85]	easy-first	93.0	87.1
Li et al. (2018) [61]	sequence-to-sequence	92.1	86.2
Strzyz et al. (2019) [86]	sequence labeling	93.7	N/A
Zhou and Zhao (2019) [30]	HPSG grammar	97.2	<b>91.2</b>
Mrini et al. (2019) [31]	HPSG, improved attention	<b>97.3</b>	N/A

Table 4.2 - (Zhang, 2020)

Additionally, Peng & al. (2022) compare different state-of-the-art dependency parsers for a wide range of UD languages and note that “*LAS takes more training data on Chinese than on other languages to reach comparable scores. Japanese and French, on the contrary, have above-average performance in LAS.*”

The Chinese have not only a lower parsing score, but also a slower learning speed.

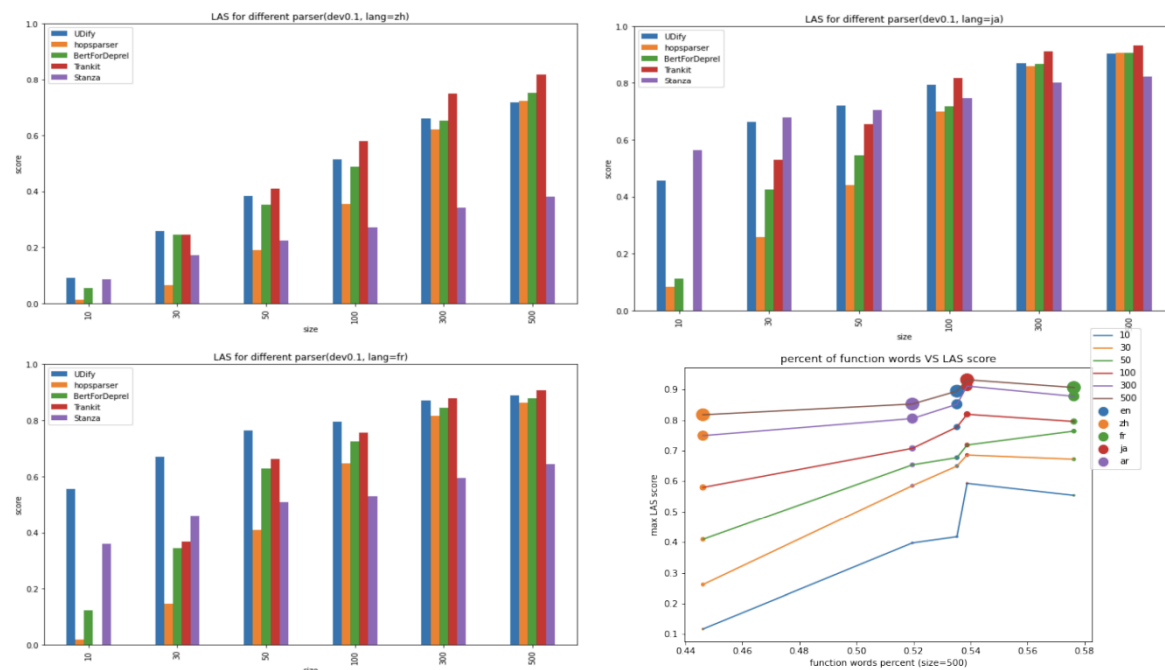


Figure 4.3 - (Peng et al., 2022)

As they showed in the last graph of Figure 4.3 on the right, the order of the languages by their percentage of function words is from 45% for Chinese (zh) to 57% for French (fr). The results demonstrate a general tendency of faster learning in languages with more function words, or put differently: The UAS and LAS scores give equal weight to relations between any words, even if the words are very frequent. It comes thus as no surprise that a language with more lexical words, which are by definition rarer, will be harder to parse.

Another reason for the low score of Chinese parsing may be the word segmentation step. Meng and al. (2019) show that “*char-based models consistently outperform word-based models*” in four end-to-end NLP benchmark tasks. This can be explained by the fact that word-based models are more vulnerable to data sparsity and the presence of out-of-vocabulary (OOV) words, and thus more prone to overfitting. This puts into question the whole procedure of word segmentation as a prior to Chinese NLP tasks.

A research by Huang (2006) compared the separate effects of data sparsity and the presence of out-of-vocabulary (OOV) word on segmentation results, and demonstrated that the segmentation errors caused by OOV are in general five times more important than word ambiguity on all four Bakeoff corpora.

Since Zhao (2009) proposed the first method for character-level dependencies parsing on the Chinese Penn Treebank, a series of research involving the character-based annotation (Li & Zhou 2012; Hatori & al. 2012; Zhang & al. 2014; Li & al. 2018) have already shown the usefulness of the

word-internal structures in Chinese syntactic parsing by obtaining limited but real improvements by means of adding character-level information to the parsing process (character POS, head character position and word internal dependency relation).

Zhao (2009) and Zhang & al. (2013) have annotated a large-scale word list of the Penn Treebank (PTB) and the Chinese constituent Treebank (CTB) on the morphological level. Each word had its own phrase structure tree. Other character-based constituent parsing attempts that we know of are generally based on these two annotated corpora.

Our work (Li & al. 2019) was the first to show that a character-based dependency treebank can be used to train a parser that gives state-of-the-art results on UD treebanks for both word segmentation and parsing on the word level. This first work was limited by the fact that the character-level dependency annotation was added to the existing word-level dependency annotation by means of automatic projection of a dependency-annotated dictionary and general rules, without human verification of the complete resulting structures.

All under-word level structures in Chinese have an internal relation belonging to one of the four following extended morphological syntactic relations in SUD, which largely correspond to its original SUD syntactic relation types:

1. m:mod label given to head-modifier relations
2. m:conj label given to coordinative relations
3. m:arg label given to subject-predicate and predicate-complement relations in which the complement is usually the result of the predicate
4. m:flat label given to unheaded word constructions and to unknown kinds of relations, usually transliterated directly from foreign languages

We finally annotated the 500 most frequent words in the Chinese SUD corpus, among which we count in total 71 left-headed words, 221 right-headed words and 198 coordinative words. For internal relations, we annotated 222 m:mod, 198 m:conj, 64 m:arg, and 16 m:flat relations. The degree of inter-annotator agreement over 100 words reached 88%. For the remaining words of our corpus, we provide an automatic character-based analysis by annotating them with the default left-right relation.

The final results for the parser are 81.72% for the UAS and 72.99% for the LAS.

In subsequent work, Yan & al. (2020) introduce a graph-based model that leverages Bert for the integration of Chinese word segmentation and dependency parsing. Li & al. (2022) manually annotated an SJTU Chinese Character Dependency Treebank (SCDT) based on the Chinese Penn Treebank (CTB-7.0) and combined constituent and dependency structure in Chinese character-level parsing using a joint span structure.

As discussed in Section 1.2.1.3, characters in Chinese correspond more or less to morphemes. In other languages, too, the word border is often arbitrary and syntax-like relations can exist inside of words between morphemes.

In field of linguistics, most corpora are annotated on the morpheme level, commonly with a transcription, a POS and, most importantly with a gloss, often following the Leipzig glossing rules. One reason for this is that the wordhood in the analysis of undocumented or under-resourced languages is a harder question than the establishment of morphemes, and words are often established based on global frequency-based analysis, once the morpheme-based annotation is done. The analysis of Beja is one example of this phenomenon (Kahane et al., 2021).

Recently, we have observed a rise in interest in combining morphological and syntactic analyses, see for example UDMorph<sup>116</sup>, a project that intends to provide a UD style annotation for morphological information within treebanks. There are also character-level annotations for other languages, such as Ruzsics et al. (2021), etc.

---

<sup>116</sup> <https://lindat.mff.cuni.cz/services/teitok-live/udmorph/>

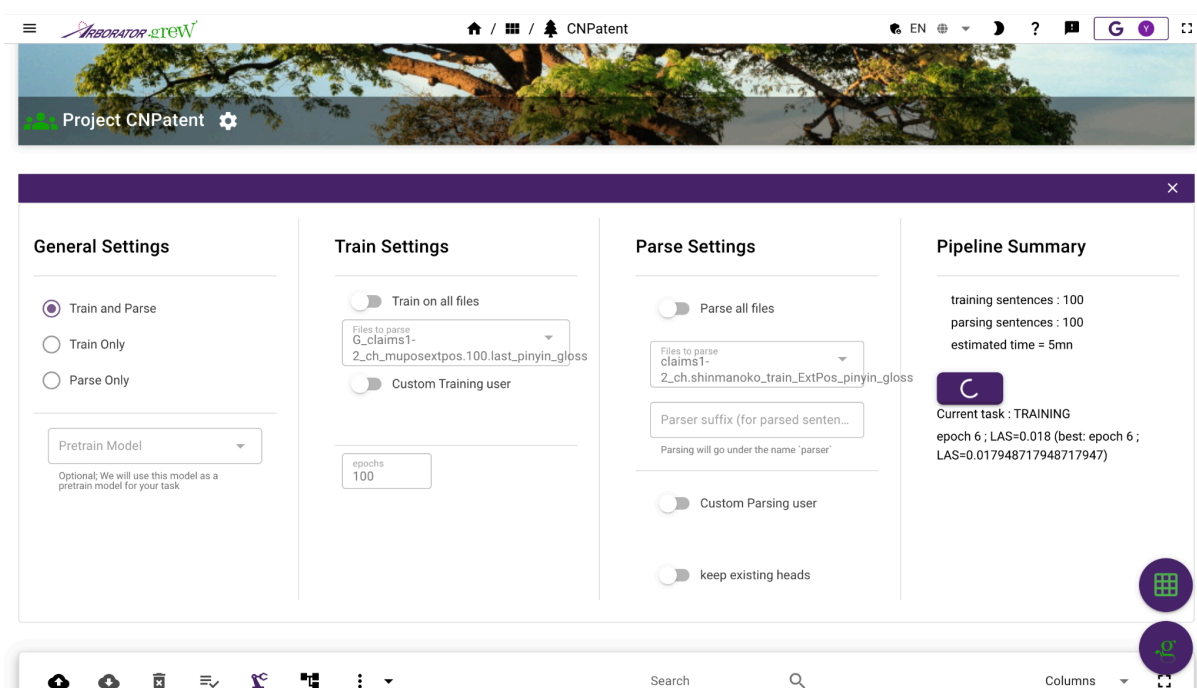
## 4.2 Word Segmentation by Parsing with “@m”

In this section, we delve into the initial application of dependency parsing as a word segmenter. This is achieved by merging the relations introduced at both the third and second levels, as outlined in Section 3.1.2.

In Section 4.2.1, we provide insights into the fine-tuning process of the Bert-based parser using our treebank, annotated using the methodology presented in Chapter 3. The subsequent Section 4.2.2 delves into the evaluation results concerning the quality of dependency parsing and its role as a word segmenter. Finally, in Section 4.2.3, we engage in a discussion of the results, complete with error analysis and suggestions for potential improvements.

### 4.2.1 Fine-tuning the Bert-based Parser

To perform syntactic dependency analysis, we fine-tune the parser proposed by Guiller (2020)<sup>117</sup>. In the beginning, I used this parser on a GPU-equipped server<sup>118</sup>. Then, my work was facilitated by the parser’s integration into the Arborator’s interface<sup>119</sup> and thus easily available for any linguist without technical training.



**Figure 4.4 - The Interface for Training the Parser**

<sup>117</sup> <https://github.com/kirianguiller/BertForDeprel/tree/master>

<sup>118</sup> I had access to the server of the University Paris-Nord, financed by the Labex EFL.

<sup>119</sup> <https://arboratorgrew.elizia.net/>

Guiller (2020) employs the architecture introduced by Dozat and Manning (2016), wherein the initial Bi-LSTM layers are replaced with a BERT model. *“The output of the final layer of the BERT transformer is connected to the input of a Multi-Layer Perceptron (MLP), designed to reduce the dimension of each contextualized vector in the sequence before applying the bi-affine transformation.”* This architecture can also be found in subsequent works such as (Kondratyuk and Straka 2019; Oh et al. 2020; Muller et al. 2020).

The standard assessment of a dependency parser encompasses three dimensions:

- Unlabelled Attached Score (UAS): We use this term to refer to the score of unlabeled attachment. This score evaluates the parser’s performance in finding the correct governors of tokens.
- Labelled Unattached Score (LUS)<sup>120</sup>: We use this term to refer to the score of labelled unattached dependency. This score assesses the parser’s performance in finding the correct syntactic relations between a token and its governor (regardless of the governor itself).
- Labelled Attached Score (LAS): We use this term to refer to the score of labelled attachment. This score evaluates the parser’s performance in finding both the correct syntactic functions and the correct attachments between tokens and their governors.

In this study, we are also interested in the accuracy of the assignment of “@m” and the POS tagging.

- Word Segmentation Score (WSS): We use this term to refer to the score of the assignment of “@m”. The score presents the accuracy of word segmentation by dependency parsing, while other word segmenters use a simple accuracy of inter-character splitting. It is calculated by the number of the correct assignments of “@m” and “head” at the same time divided by the total number of labels in the sentence. The score in Table 4.4 is the average of scores of all sentences. Note that in our annotation all dependents of “@m” dependents are also “@m” dependents, or put differently, subwords can not have full-words as dependents;
- POS Tagging Score (PTS): This term is used to refer to the accuracy of the assignment of parts-of-speech for each character. It is calculated by the number of the correct POS tagging divided by the length of the sentence.

Although my character-based patent treebank is quite small (200 sentences, 8,175 characters), it remains interesting to see how good a patent parser can become by training on this data.

As resources beyond the BERT model, we only have one other Chinese treebank which is annotated on the character level: Project chinese\_grammar\_wiki\_morphSUD which consists of thousands of sentences of Chinese learning corpus. This is a very different domain from patents, but we can expect some common relations to be present in both treebanks.

---

<sup>120</sup> The LUS in this context does not take into account the variations resulting from the assignment of “@m”. For example, if a relation “conj” is annotated as “conj@m”, it will be counted as correct label.

We employed the initial 100 sentences from Section 3.3 as our test dataset, while the subsequent 100 sentences were utilized as the training dataset.

We conducted a comparison of the results achieved with and without a pre-trained model using treebanks on a Chinese learning corpus that had been manually annotated at the character-level<sup>121</sup>.

	Without pre-trained model	With pre-trained model
<b>UAS</b>	69.41%	76.79%
<b>LUS</b>	82.73%	85.90%
<b>LAS</b>	64.76%	72.10%
<b>WSS</b>	86.51%	89.05%
<b>PTS</b>	89.30%	90.46%

**Table 4.3 - The Parsing Results with and without the Pre-trained Model**

- In the absence of the pre-trained model, the training process came to an automatic halt at the 166th epoch, achieving the highest LAS (Labeled Attached Score) of 0.81 on the training dataset. Table 4.3 displays the evaluation outcomes of the test data across various dimensions.
- When utilizing a pre-trained model, the training process automatically stopped at the 60th epoch, with the highest LAS (Labeled Attached Score) on the training dataset reaching 0.85. Table 4.3 presents the evaluation results of the test data across each dimension.

It is evident that the outcomes with and without a pre-trained model exhibit remarkable similarity, with the primary distinction lying in the learning speed, as depicted in Figure 4.3.

In the results, we observe that the parser displays a greater proficiency in recognizing relation tags compared to heads, as evidenced by a higher LUS score for both models. Moreover, with LAS scores ranging from 65% to 72%, the scores for word segmentation and POS tagging approach the 90% mark.

However, it's worth noting that these scores are relatively lower than conventional results, likely due to the limited size of the training dataset, consisting of only 100 sentences and 4,380 tokens. As demonstrated by Peng et al. (2022), Chinese parsers typically require a more substantial training dataset (more than 500 sentences) to achieve comparable accuracy levels to languages such as English, French, and Japanese.

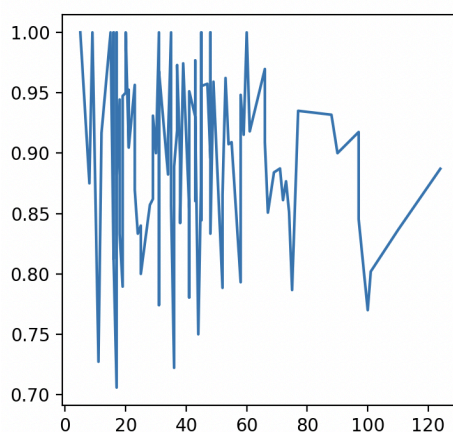
<sup>121</sup> [https://arboratorgrew.elizia.net/#/projects/chinese\\_grammar\\_wiki\\_morphSUD](https://arboratorgrew.elizia.net/#/projects/chinese_grammar_wiki_morphSUD)

### 4.2.2 Error Analysis

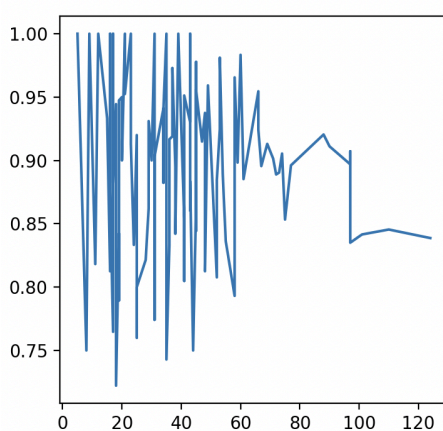
In this section, we aim to provide an explanation for the results presented in the preceding section, considering both quantitative and qualitative aspects. Our initial focus lies on examining errors related to word segmentation, particularly concerning the assignment of “@m”. Subsequently, we will shift our attention to analyzing errors associated with the dependency parsing process itself.

As stated before, the test gold treebank that we use consists of the first 100 sentences manually annotated in Section 3.3.

To determine whether the length of sentences affects word segmentation results, we computed the correlation between the Word Segmentation Score (WSS) and sentence length (measured by the number of characters). For the training without the pre-trained model, the Pearson correlation score was -0.07901659 (Figure 4.5 (a)), and for the training with the pre-trained model, it was -0.01038436 (Figure 4.5 (b)). These results indicate that there is no significant influence of the sentence length on the word segmentation quality.



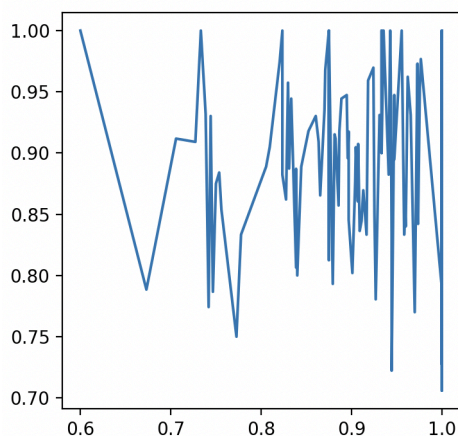
**Figure 4.5 (a) - The Length of the Sentence and its WSS (without pre-trained modal)**



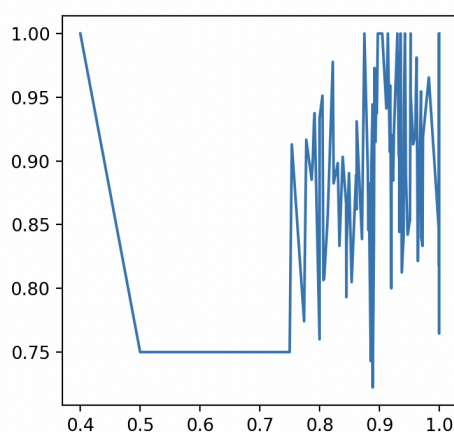
**Figure 4.5 (b) - The Length of the Sentence and its WSS (with modal)**



Additionally, we investigated the correlation between the WSS and Part-of-Speech Tagging Score (PTS) to assess whether word segmentation is influenced by POS tagging. The correlation score was 0.03876942 (Figure 4.6 (a)) in the case of training without the pre-trained model, and 0.19258711 (Figure 4.6 (b)) in the case of training with the pre-trained model, suggesting that there is no correlation of POS tagging quality and word segmentation quality in the experiment.



**Figure 4.6 (a) - WSS and PTS (without modal)**



**Figure 4.6 (b) - WSS and PTS (with modal)**

We also compared our dependency-based word segmenter with two standard tools: Jieba<sup>122</sup> and NLPIR/ICTCLAS segmenter<sup>123</sup>.

- Jieba<sup>124</sup> is a widely used Chinese word segmentation tool renowned for its efficiency and accuracy. Its algorithm relies on a prefix dictionary structure, enabling rapid word graph scanning. One of its key features is the construction of a Directed Acyclic Graph (DAG) encompassing all potential word combinations. To determine the most probable word

<sup>122</sup> <https://github.com/fxsjy/jieba>

<sup>123</sup> <https://github.com/tsroten/pynlpir>

<sup>124</sup> The word “jieba” in Chinese means “to stutter”.

segmentation, Jieba employs dynamic programming techniques, considering word frequencies to optimize the results. Additionally, for handling unknown words, Jieba utilizes a Hidden Markov Model (HMM) combined with the Viterbi algorithm, enhancing its ability to accurately segment text, even in challenging scenarios where word boundaries are less clear.

- PyNLPIR is a Python wrapper around the NLPIR/ICTCLAS Chinese segmentation software<sup>125</sup>. Developed by the Institute of Computing Technology, ICTCLAS is a Chinese lexical analysis system, using an approach based on multi-layer HMM. ICTCLAS comprises three key components: word segmentation, Part-Of-Speech tagging, and unknown word recognition. Notably, its word segmentation precision has been measured at an impressive 97.58%, as per the latest official evaluation conducted within the national 973 project<sup>126</sup>. Additionally, when it comes to the recognition of unknown words through role tagging, ICTCLAS boasts an impressive recall rate of over 90%.

Segmenter	Accuracy
Parsed	91.71%
Parsed with the pre-trained model	92.50%
Jieba	92.64%
PyNLPIR (ICTCLAS)	88.88%

**Table 4.5 - Accuracy of Different Segmentations Compared to the Manually Annotated Test Gold Treebank**

Among all four systems, our parser with the pre-trained model and the Jieba segmenter perform the best. We delve into the errors present in the segmented text by Jieba. These errors can be categorized into three systematic discrepancies.

- The first pertains to the segmentation of the serial numbers, which is segmented by Jieba but remains unsegmented in the gold treebank.

Gold Segmentation	Jieba's Segmentation
1./-/种/压力/传感器/, /所述/压力/传感器/包括/:	1./-/一种/压力/传感器/, /所述/压力/传感器/包括/:

**Table 4.6 (a) - Segmentation Discrepancies Related to Serial Numbers**

<sup>125</sup> <http://sewm.pku.edu.cn/QA/reference/ICTCLAS/FreeICTCLAS/English.html>

<sup>126</sup> The National Basic Research Programme (973 Programme) is an on-going (from 1997) keystone research programme supported by the Ministry of Science and Technology of China, aimed to develop basic research, innovation and technologies in line with national targets for economic and social development. Its strategic objectives are to stimulate original innovations and address scientific issues important for national socio-economic growth.

- The second type of error pertains to determinants like ‘一种 (yī zhǒng, ‘a kind/ a type’), which is a frequent expression in patent claims and has occurred 14 times in the segmented test text.

Gold Segmentation	Jieba’s Segmentation
一/种/压力/传感器 1/kind/pressure/sensor	一种/压力/传感器 a/pressure/sensor
所述/基座/包括/多/个/嵌槽/; The /base/ includes/multiple/embedded slots/;	所述/基座/包括/多/个/嵌槽/; The /base/ includes/multi/ple/embedded slots/;

**Table 4.6 (b) - Segmentation Discrepancies Related to Determinants**

- The last and most important type of errors come from the process of polysyllabic terms.

Gold Segmentation	Jieba’s Segmentation
根据/所述/数据/复制/日志/获取/所述/源数据/节点/ 的/版本/, According to/the/data/replication/log/ get/the/ source-data/node/version/,	根据/所述/数据/复制/日志/获取/所述/源/数据/节点/ 的/版本/, According to/the/data/replication/log/ get/the/ source-/data/node/version/,
一/种/基/于/卡片点/的/数学/图形/认知/思维/可视化/ /系统/, A/kind/based/on/card point/ mathematical/ graphic/cognitive/thinking/visualization/ system/,	一种/基于/卡片/点/的/数学/图形/认知/思维/可视化/ /系统/, A/kind/based/on/card /point/ mathematical/ graphic/cognitive/thinking/visualization/ system/,
一/种/向量/四则/运算/装置/, One/kind/vector/four arithmetic/operation/device/,	一种/向量/四则运算/装置/, One/kind/vector/four /arithmetic/operation/device/,
步骤/1/, /根据/区块/数据/建立/地质/模型/; Step /1/, /build /geological/model/ based on /block/ data/; 步骤/2/, /根据/生产/资料/进行/历史/数据/历史/拟 合/; Step /2/, /based on / production / data / for / history / data / history / fitting /;	步骤/1/, /根据/区块/数据/建立/地质模型/; Step /1/, /build /geological/model/ based on /block data/; 步骤/2/, /根据/生产资料/进行/历史数据/历史拟合/ ; Step /2/, /based on / production data / for / history data / history fitting /;

**Table 4.6 (c) - Segmentation Discrepancies Related to Polysyllabic Terms (Jieba)**

To elucidate the relatively lower accuracy of the ICTCLAS segmenter, while there are no discrepancies on the serial numbers and the determinants, the difference on polysyllabic terms also exists, which is especially problematic on the segmentation of trisyllabic terms (e.g. second to fourth examples in Table 4.6 (d)). In addition, another important difference is that the very frequent term “所述 (suǒ shù, ‘said’)” is always segmented (Table 4.6 (e)).

Gold Segmentation	ICTCLAS's Segmentation
在/栅控/器件/应用/电路/中/ in /gate/devices/applications/circuit/in/	在/栅/控/器件/应/用/电/路/中/ In /gate/control/device/appli/cation/electrical/circuit/
一/种/触控笔/, /其/特征/在/于/ A/kind of/stylus/,/its/characteristics/in/in/	一/种/触/控/笔/, /其/特征/在于/ A/kind of/touch/control/pen/,/its/characteristic/lies in/
一/桶状部/; 1/barrel/;	一/桶/状/部/; One/barrel/shaped/part/;
一/处理/电路/, /具有/一/通信/子电路/。 a /processing/circuit/, /having/a/ communication/ subcircuit/.	一/处理/电路/, /具有/一/通信/子/电/路/。 A/process/circuit/,/have/a/communication/sub/electric al/circuit/.

Table 4.6 (d) - Segmentation Discrepancies Related to Polysyllabic Terms (ICTCLAS)

Gold Segmentation	ICTCLAS's Segmentation
所述/第一/导体/包括/: Said/first/conductor/includes/:	所/述/第/一/导/体/包/括/: The/said/first/conductor/includes/:

Table 4.6 (e) - Segmentation Discrepancies Related to the Term “所述”

Finally, we show some examples of errors in our segmenter in Table 4.6 (f) for the parser-segmenter with the pretrained model, and in Table 4.6 (g) the parser-segmenter without the pretrained model.

Gold Segmentation	parser-segmenter's Segmentation	Category of errors
一条 one piece	一/条 one / piece	Determinants
一种 one kind	一/种 one / kind	Determinants
导电/部分 conductive/part	导电部/分 conductive part/min	Polysyllabic Terms
面积比 area ratio	面/积/比 area/area/ratio	Polysyllabic Terms
间隔/开 space/open	间/隔/开 space/separate	Polysyllabic Terms
触控笔 stylus	触/控/笔 touch control/pen	Polysyllabic Terms
一/桶状部 a barrel	一/桶/状/部 a barrel/shape/part	Polysyllabic Terms & Determinants
不同 not/ same	不/同 not same	Negative modifier “不 (bù)”

Table 4.6 (f) - Segmentation Discrepancies of the parser-segmenter with the pretrained model

We can observe in the Table 4.6 (f), that the most of segmentation errors by the parser-segmenter with the pretrained model can be categorized into discrepancies related to determinants and discrepancies related to polysyllabic terms, which are the same of Jieba. Except these typical errors, the Table 4.6

(g) lists some more specific errors made by the parser-segmenter without the pretrained model, involving disyllabic terms, the possessive marker “的 (de)”, the determinants and even punctuation.

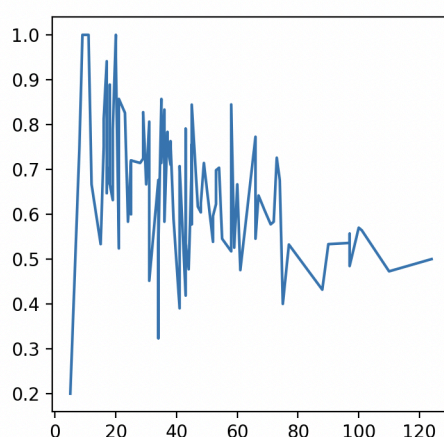
Gold Segmentation	parser-segmenter's Segmentation	Category of errors
汇聚 gather	汇/聚 converge/gather	Disyllabic Terms
表征 symptoms	表/征 symptoms/symptoms	Disyllabic Terms
出现/的 appear/ed	出/现的 app/eared	Mix
一/条/第-一/类/参考/信息/ 1/item/first/category/reference/ information/	一条第-一/类/参考/信息/ a first/category/reference/information/	Mix
用户/数量/, /选取/得到 User/number/,/select/get	用户/数量, 选取/得到 User/number,select/get	Punctuation

**Table 4.6 (g) - Segmentation Discrepancies of the parser-segmenter without the pretrained model**

Moving on to the syntactic analysis, to see what influences most the performance of the parser we conducted correlations between the Labeled Attachment Score (LAS), sentence length (measured in characters), Word Segmentation Score (WSS), and Part-of-Speech Tagging Score (PTS). Here are the results:

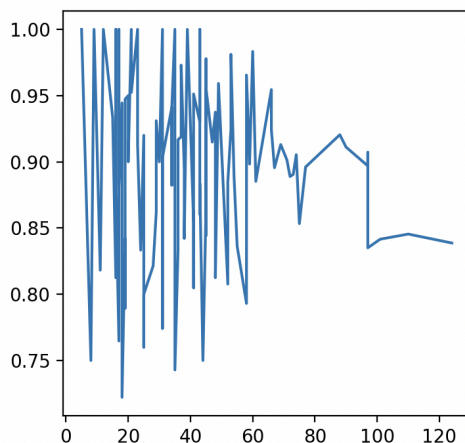
1. LAS vs. Sentence Length:

- Without the pre-trained model (Figure 4.7 (a)): Correlation score of -0.29089937



**Figure 4.7 (a) - LAS and length (without modal)**

- With the pre-trained model (Figure 4.7 (b)): Correlation score of -0.43473626



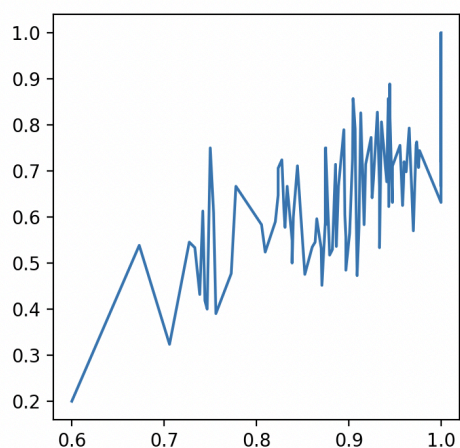
**Figure 4.7 (b) - LAS and length (with modal)**

In both cases, we observed only a weak correlation between LAS and sentence length. This result suggests that the length of the parsed sentence is not a vital factor in the quality of dependency parsing on the patent claims, but still, it has a certain influence on the dependency parsing quality, especially in the case of training with the pre-trained model.

One difficulty of parsing long sentences is that longer sentences may contain a longer dependency distance, which is widely accepted as a factor that may increase the syntactic complexity. The influence of the dependency distance will be discussed later in the section.

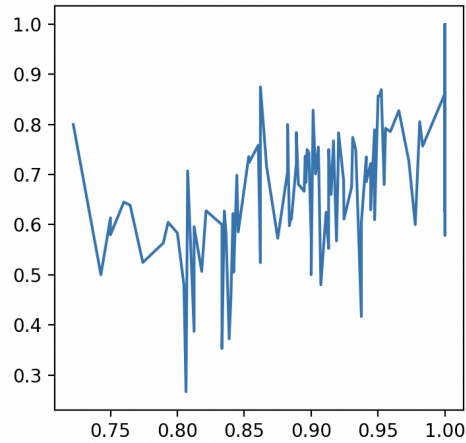
2. LAS vs. WSS:

- Without the pre-trained model (Figure 4.8 (a)): Correlation score of 0.70104524



**Figure 4.8 (a) - LAS and WSS (without modal)**

- With the pre-trained model (Figure 4.8 (b)): Correlation score of 0.42898856

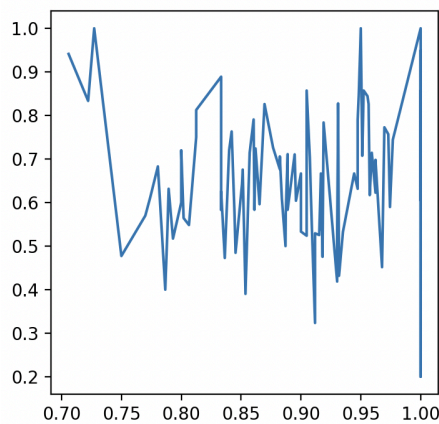


**Figure 4.8 (b) - LAS and WSS (with modal)**

Among all, the correlation between LAS and WSS is the strongest one. This indicates that the accuracy of syntactic-level dependency parsing is significantly influenced by the quality of word segmentation, and this is especially true in the case of training without the pre-trained model.

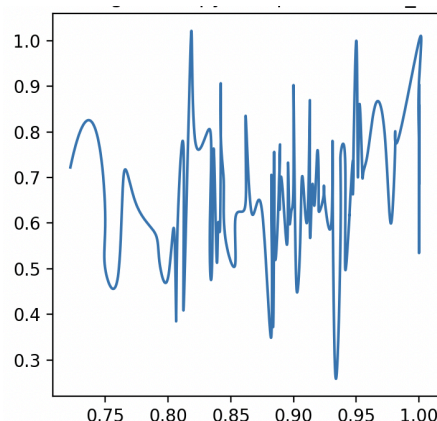
3. LAS vs. PTS:

- Without the pre-trained model (Figure 4.9 (a)): Correlation score of 0.07825408



**Figure 4.9 (a) - LAS and PTS (without modal)**

- With the pre-trained model (Figure 4.9 (b)): Correlation score of 0.35256026



**Figure 4.9 (b) - LAS and PTS (with modal)**

Compared to the segmentation, there is minimal correlation between the quality of dependency parsing and that of the POS tagging.

In summary, while a weak correlation exists between LAS and sentence length, the most substantial correlation is between LAS and WSS, highlighting the crucial role of accurate word segmentation in achieving high-quality dependency parsing. On the other hand, there is minimal correlation between LAS and PTS, suggesting that the quality of POS tagging has a limited impact on dependency parsing accuracy.

Another important factor that may affect the performance of the parser is the level of syntactic complexity, which is usually represented by the dependency distance that is already mentioned above. (Gibson, 2000; Gildea and Temperley, 2010; Grodner and Gibson, 2005; Liu 2007, 2008; Liu et al., 2017; Oya, 2013)

The dependency distance (DD) is a metric that quantifies the linear distance between two words connected by a dependency relationship within the same sentence. Wang and Liu (2017) have highlighted that several factors can influence DD, including language type, sentence length, chunking (Lu et al., 2016), the chosen annotation scheme, genre, and grammatical structure.

Liu and al. (2017) calculate the mean dependency distance (MDD)<sup>127</sup> of a sentence or a text by the following formulas:

$$MDD(\text{the sentence}) = \frac{1}{n-1} \sum_{i=1}^n |DD_i|, \quad (1)$$

$$MDD(\text{the text}) = \frac{1}{n-s} \sum_{i=1}^n |DD_i|, \quad (2)$$

**Figure 4.10 - (Liu et al., 2017)**

<sup>127</sup> The mean dependency distance (MDD) is also called average dependency distances (ADD). (Oya, 2021)



The mean dependency distance (MDD) of the patent claim treebank composed of all 200 sentences is 3.942563381098799. Compared to the MDD of the HIT Chinese treebank of 1.89 (Liu et al., 2009), the distance between the dependent and the head is longer in patent corpus.

To see the effect of MDD on the parser's performance, we also calculate the correlation of dependency distance and the correctness of head and deprel prediction:

- The correlation of correct prediction of the head position and dependency distance is: -0.15736031 of the case without the pretrained model and -0.12642214 of the case with the pretrained model.
- The correlation of correct prediction of the deprel and dependency distance is: -0.05250661 of the case without the pretrained model and -0.01924076 of the case with the pretrained model.

The effect of the dependency distance and the parsing accuracy is not obvious in this case.

In addition to the general accuracy of the dependency parsing, we are also interested in the accuracy of each individual relation. In Table 4.7 below, we list the accuracy of parsing of each dependency relation label with pretrained model and the quantity of each deprel label used in the gold treebank. The accuracy is calculated by deviding the number of correctly annotated labels by the total number of that label.

Deprel	Accuracy	Quantity
punct	1.0	304
comp:obl	0.9066666666666666	75
mod@m	0.8620689655172413	667
mod	0.8577476714648603	1,181
comp@m	0.8531468531468531	286
cc	0.84	50
root	0.82	100
conj@m	0.813614262560778	617
parataxis	0.8024691358024691	81
comp:obj	0.7700729927007299	548
subj	0.7380952380952381	126

conj	0.46969696969697	66
comp:res	0.4444444444444444	18
appos	0.4	5
comp:pred	0.3333333333333333	3
comp:res@m	0.22580645161290322	31
flat@m	0.14285714285714285	14
comp:svc	0.07692307692307693	13
comp:obl@m	0.0	6
comp:aux	0.0	4
comp:aux@m	0.0	9
comp:obj@m	0.0	70
subj@m	0.0	7
det	0.0	5

**Table 4.7 - The Parsing Accuracy of Each Type of Relations in The Patent Treebank**

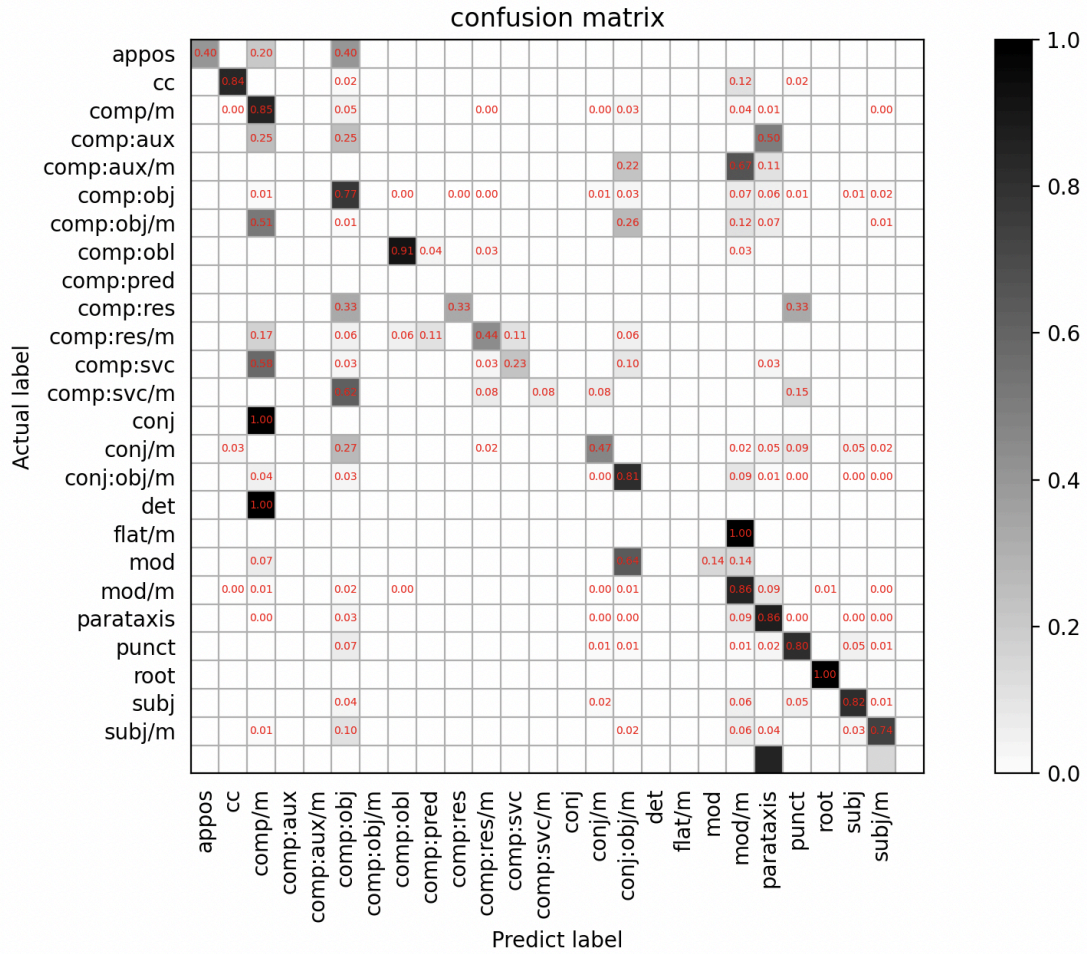


Figure 4.10 - Confusion Matrix for the Predicted Deprel Labels

## 4.3 Dependency Parser-based Term Recognition for Chinese Patents

Elevating our focus beyond individual word recognition, we now explore the concept of terms, which can encompass multiple words.

In this section, we delve into the second application of the dependency parser, which involves using it as a term recognizer on the patent corpus. This is accomplished by amalgamating the “mod” relations introduced at the first level of annotation in Section 3.1.2. As discussed in Section 4.3.2, we consider that term-internal relations should all be tagged as “mod” relations. All other relations remain outside of the term on the actual syntactic level.

The work is an unpublished collaboration with Liu Lufei from the Qatent team.

After providing a concise overview of the term recognition methods in Section 4.3.1, Section 4.3.2 outlines our approach to extracting technical terms from patent claims, focusing specifically on dependency syntax parsing as opposed to the conventional ATE (Automatic Term Extraction) based on parts-of-speech. Additionally, we will perform a comparative analysis of our results vis-à-vis other existing techniques, such as the POS pattern extraction, which will be detailed in Section 4.3.3.

### 4.3.1 A Brief Presentation of Term Recognition

Only a few ATE tasks are applied on or for patent texts. Zeng Zhen, Lü Xueqiang and Li Zhuo (2016) present a method that begins by utilizing part-of-speech rule templates to obtain a set of candidate single-word and multi-word terms. It then calculates lexical density weight parameters to extract single-word terms. Finally, it combines balanced corpora to automatically generate a filtering dictionary. The method involves using this filtering dictionary and the term factors of each word composing the long term to select the final long terms. In a study titled “Research on Chinese Patent Candidate Term Selection Based on Dependency Syntax Analysis” by Yu Yan, Chen Lei, Jiang Jinde, and Zhao Naidu (2019), a method for selecting Chinese patent candidate terms based on dependency syntax analysis is proposed.

The Automatic Term Extraction (ATE) systems usually employ a two-step procedure: (1) extracting a list of candidate terms, and (2) determining which candidate terms are correct using supervised or unsupervised techniques. (Tran et al., 2023)

In summary, term extraction research is mainly focused on improving candidate term ranking algorithms (step 2), with relatively little attention given to candidate term selection studies (step 1). In particular, it is difficult to establish general pattern-matching rules for candidate term selection methods. This results in the formulation of idiosyncratic pattern-matching rules for different datasets, making it difficult to compare results across datasets and genres. (Yu et al. 2019)

Here we are primarily interested in the first step of the construction of the term candidate list.

Yu et al. (2019) classified candidate term selection methods into three categories: n-gram filtering, noun phrase chunking, and POS tag pattern matching.

- N-gram filtering

Numerous studies (Hu et al., 2013; Liu et al., 2014; Ding et al., 2015) have investigated N-gram filtering as an integral aspect of automated term extraction. N-gram filtering fundamentally entails the meticulous removal of semantically insignificant elements, such as stop words, particles, or interjections, from a text fragment. Following this initial filtering phase, the text fragment undergoes scrutiny to identify continuous n-word sequences, where 'n' signifies the number of words in each sequence. These sequences are then subjected to specific rules and criteria for the purpose of selecting pertinent multi-word phrases. This selection process may prioritize high-frequency words to ensure the chosen phrases are both informative and representative of the text's content.

The method is well-regarded for its simplicity. Meantime, a primary concern is the potential persistence of non-terminological word sequences even after filtering, which, if not managed effectively, can introduce noise and adversely affect the precision of term extraction, posing a substantial obstacle in the pursuit of accurate and meaningful text analysis.

- Noun phrase chunking

Noun phrase chunking is a crucial approach in the process of term extraction (Frantzi et al., 2000), primarily because terms are typically represented as noun phrases within text sequences that have been subjected to part-of-speech tagging. The task at hand involves the identification of noun phrases following specific syntactic patterns, with one common pattern being the "ADJ + N" structure.

This method, characterized by its simplicity and efficiency, has found widespread use in the realm of term extraction, particularly in English language contexts.

- POS tag pattern matching

The part-of-speech (POS) tag pattern matching serves as a prominent technique, sharing a common foundational principle with noun phrase chunking. Both approaches operate under the assumption that terms within a text exhibit discernible patterns based on their part-of-speech sequences. Nevertheless, the distinguishing factor between the two methodologies lies in the complexity of the matching patterns employed, as elucidated by Xu et al. (2014), Zeng et al. (2016), and Yang et al. (2016).

The strength of part-of-speech pattern matching lies in its capacity to delineate and specify intricate matching rules that are tailored to the nuances of Chinese text. This adaptability makes it a cornerstone method for term extraction within the realm of the Chinese language, by crafting rules that align with the unique characteristics of Chinese linguistic structures and the diverse linguistic contexts and datasets. However, one noteworthy challenge is the manual definition of distinct matching rules for different Chinese datasets, which can be a time-consuming and labour-intensive process.

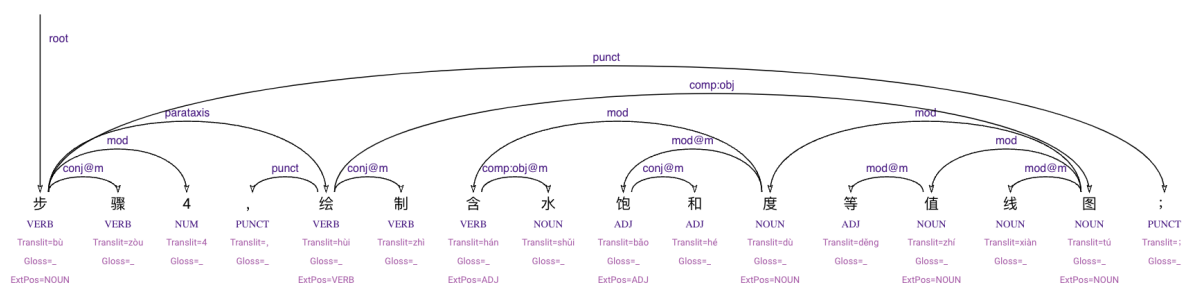
In Section 4.3.3 we will compare the conventional POS-based system with our dependency-based approach for the term recognition.

## 4.3.2 Term Recognizer based on Dependency Relation Information

This section focuses on term extraction through dependency parsing. We delve into the criteria for extraction and provide an overview of the extraction results.

In a recent work, Yu et al. (2019) introduced the application of dependency syntactic analysis to term recognition. They highlighted that this approach uncovers semantic modifier relationships among words within sentences by examining dependency relationships. This methodology enhances the comprehension of semantics and effectively overcomes the constraints associated with solely relying on part-of-speech methods, which may encounter difficulties in capturing intricate semantic connections.

Utilizing our character-level dependency parser, we employ dependency relations to extract terms from patent claims. Initially, we aggregate all internal relations marked with “@m”, followed by the extraction of tokens linked by the “mod” relation, denoting the attributive compound relationship. One example is demonstrated in Figure 4.11, in which the term “含水饱和等值线图 (hán shuǐ bǎo hé děng zhí xiàn tú, ‘water saturation contour map’)” is connected by a series of “mod” or “xxx@m” relations.



**Figure 4.11 - Example sentence from my PatentChar SUD treebank (zh\_patentchar-sud-test\_80)**

The question of why we utilize the “mod” relationship to identify terms can be answered by considering the common practice in POS tagging methods. Typically, when dealing with the combination of nouns (N) in Chinese text, denoted as “N+N”, the dependency relation that frequently arises is “mod”. In other words, this choice of extracting the “mod” relation is based on the prevalent linguistic pattern where nouns are commonly paired together, and the “mod” dependency relation effectively captures the attributive or modifier role played by one noun toward the other. As a result, by leveraging the “mod” relationship, we can effectively identify and extract terms or noun phrases that are commonly structured in this manner in Chinese language text.

This is the same for N+N+N and N+N+N+N, with two or three “mod” relations.

We conducted separate experiments on both the manually annotated treebank and the automatically parsed treebank, utilizing a stopwords list<sup>128</sup>, some of which is displayed in Table 4.8.

<sup>128</sup> The stopwords list is composed of two parts: the manually selected words and the stop word list downloaded from HIT at <https://github.com/YueYongDev/stopwords/blob/master/哈工大停用词表.txt>, which is presented in the Annex.

Some Stop Words	Pinyin	Translation
权利要求	quán lì yāo qiú	Claim(s)
权利	quán lì	Right
种	zhǒng	Type
一种	yī zhǒng	One type / A kind
至少	zhì shǎo	At least
一个	yī gè	One
要求	yāo qiú	Requirement / Claim
所述	suǒ shù	Aforementioned
上述	shàng shù	Above-mentioned
以上	yǐ shàng	Above
如图所示	rú tú suǒ shì	As shown in the figure
如图	rú tú	As shown in the figure
所示	suǒ shì	As shown
发明	fā míng	Invention
内容	nèi róng	Content
发明内容	fā míng nèi róng	Invention content
实施例	shí shī lì	Embodiment
技术方案	jìshù fāng àn	Technical solution
...	...	...

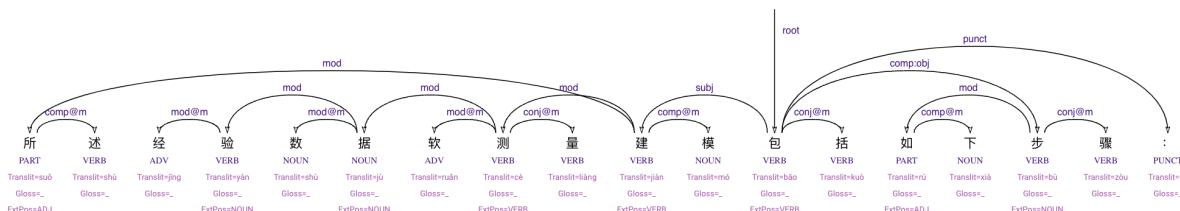
Table 4.8 - A Part of the Stopword List

Manually Annotated Only	Both	Automatically Parsed Only
108	220	303
可视化系统 visualization system 卫星状态 satellite status 交换 exchange 图形思维 graphic thinking 目的节点 destination node 目的数据 destination data 测量建模 measurement modeling 数据服务器 data server 管理CPU manage CPU 解析软测量 analyzing soft measurements 数学图形 mathematical graphics 块接入卡 block access card 数据收集 data collection 在轨卫星 satellite in orbit 要素页面 feature page 含油面积 oil-bearing area 测量接口 measurement interface 压力传感器 pressure sensor 按键模块 button module 分布式数据库 distributed database ...	面积 area 电容 capacitance 力量 strength 识别电路 identification circuit 分布式 distributed 含水饱和度 water saturation 栅控器件 gate control device 信息表 information sheet 业务 business 测试数据 test data 基准事件 base event 实时采集 real-time collection 设备 equipment 触控笔 stylus 信息系数 information coefficient 操作指令 operation instructions 用户地址 user address 用户 user 信号 signal 消息 information ...	键参数 key parameter 硬软件 hardware and software 数据指令 data instructions 接收用户 receive user 向量指令 vector instructions 复日志 (re)ply log 数据数 number of data 目标请求 target request 待验数据 data to be verified 建模数据 modeling data 统模型 (sys)tem model 业务素页 business feat(ure) page 散点关联 scatter correlation 响应判断 response judgment 思系统 thinking system 向量存储 vector storage 量向量 quantity vector 发送携带有 send with 件标 part mark 级指令 level instructions ...

Table 4.9 - Examples of Extracted Terms from Different Treebaks

Within the category of stopwords, it's worth noting that terms like “权利要求 (quán lì yāo qiú,

‘claim’)” and “实施例 (shí shī lì, ‘example of implementation’)” are indeed legal terms commonly used in patent drafting. However, they appear in a wide range of IPC classes and therefore do not fall under the category of technical terms associated with the specific technical domains we are focusing on here, see Section 1.3.2 for a discussion of this issue notes that many of the stopwords of Table 4.8 are not nouns, such as “上述 (shàng shù, ‘aforementioned’)” or “所述 (suǒ shù, ‘said’)”. They have to be included in the stopword list as they may appear related to a noun phrase by a “mod” relation (see example in Figure 4.12), while we want to exclude them from the complete term.



**Figure 4.12 - Example sentence from my PatentChar SUD treebank (zh\_patentchar-sud-test\_38)**

As results, we extracted 328 candidate terms from the manually annotated treebank, and 523 candidate terms from the automatically parsed treebank. There are 220 candidate terms in common among the two lists (Table 4.9). As evident in the third column, the presence of errors introduced by dependency parsing leads to a higher number of “broken” candidate terms (in red) within the list labeled as “Automatically Parsed Only”.

To further assess the term recognizer’s quality, our initial focus should be on defining what constitutes a “good term” within the context of this study. The ultimate objective here is to facilitate lexical variation for augmented inventing (Lee 2020), rather than solely acquiring a technical lexicon tailored to a specific domain.

Drawing from the definition of a term in Section 1.3, lacking of manually annotated gold test files, we establish our own set of criteria for extracting terms from patent claims. In real patent claims, it’s noteworthy that a substantial number of patents receive important technical indexing, as indicated in Table 4.10 (in yellow). We can regard these indexed terms as valuable candidates for substitution. Consequently, we assess the performance of the dependency-based term recognizer by determining how many of these indexed terms it successfully identifies.

Chinese	English
一种新型可加热分离型升降火锅	Novel heatable separation type lift chafing dish
1.一种新型可加热分离型升降火锅, 它包含锅体(1)、锅盖(2), 其特征在于: 它还包含锅胆(3)、升降机构(4)、马达(5)、升降盘(6), 所述的锅体(1)由锅身(11)、锅底座(12)、发热盘(13)组成, 锅底座(12)上部设置有马达座(121), 马达(5)固定连接于马达座(121), 发热盘(13)设置在锅底座(12)正上方且位于锅身(11)内部, 升降机构(4)设置在锅底	1. The utility model provides a novel heatable separation type lift chafing dish, it contains the pot body (1), pot cover (2), its characterized in that: it still contains pot courage (3), elevating system (4), motor (5), lifting disk (6), the pot body (1) constitute by pot body (11), pot base (12), dish (13) generate heat, pot base (12) upper portion is provided with motor seat (121), motor (5) fixed connection is in motor seat (121), dish generate heat (13) set up directly over pot base (12) and be located inside pot body (11), elevating system (4) set up in pot base (12) inside middle part and be



<p>座(12)内部中间部位并与马达(5)连接, 锅胆(3)套在升降机构(4)上并在发热盘(13)的正上部, 升降盘(6)套在升降机构(4)的上部, 锅盖(2)盖在锅胆(3)上部。</p> <p>2. 根据权利要求1所述的一种新型可加热分离型升降火锅, 其特征在于: 所述的发热盘(13)为圆形中空结构, 发热盘(13)中空部位设置有底座(131)。</p> <p>3. 根据权利要求1所述的一种新型可加热分离型升降火锅, 其特征在于: 所述的锅胆(3)包含内胆(31)和手柄(32), 内胆(31)中心设置有锅芯(311), 锅芯(311)为圆柱形中空结构, 锅芯(311)在内胆(31)正上方并垂直连接于内胆(31), 两个手柄(32)设置在内胆(31)上方的左右两侧, 两个手柄(32)上设置有锁紧耳扣(321)。</p> <p>4. 根据权利要求1所述的一种新型可加热分离型升降火锅, 其特征在于: 所述的升降盘(6)包含圆盘(61)和凸台(62), 凸台(62)为顶部封闭的圆柱形中空结构, 圆盘(61)表面设有若干圆孔(611), 圆盘(61)与凸台(62)底部连接。</p> <p>5. 根据权利要求1所述的一种新型可加热分离型升降火锅, 其特征在于: 所述的升降机构(4)包含传送螺杆(41)、升降螺杆(42)、升降杆(43), 传送螺杆(41)位于发热盘(13)的正下方, 传送螺杆(41)的一端固连于马达(5), 升降螺杆(42)设置在传送螺杆(41)正上方且垂直于传送螺杆(41), 升降螺杆(42)通过底座(131)固定在发热盘(13)上并与传送螺杆(41)传动连接, 升降杆(43)内部为螺纹结构并旋转连接于升降螺杆(42), 升降杆(43)上部两侧设有限位台(431), 两个限位台(431)位于锅芯(311)正上方。</p> <p>6. 根据权利要求1所述的一种新型可加热分离型升降火锅, 其特征在于: 所述的锅身(11)上部设置有两个锁紧扣位(111)。</p>	<p>connected with motor (5), pot courage (3) cover is on elevating system (4) and on the positive upper portion of dish (13) generate heat, lifting disk (6) cover is on the upper portion of elevating system (4), pot lid (2) cover is on pot courage (3) upper portion.</p> <p>2. A novel heatable separation type lifting hot pot as claimed in claim 1, characterized in that: the heating plate (13) is of a circular hollow structure, and a base (131) is arranged in the hollow part of the heating plate (13).</p> <p>3. A novel heatable separation type lifting hot pot as claimed in claim 1, characterized in that: the cooker liner (3) comprises an inner liner (31) and handles (32), a cooker core (311) is arranged at the center of the inner liner (31), the cooker core (311) is of a cylindrical hollow structure, the cooker core (311) is arranged right above the inner liner (31) and is vertically connected to the inner liner (31), the two handles (32) are arranged on the left side and the right side above the inner liner (31), and locking lugs (321) are arranged on the two handles (32).</p> <p>4. A novel heatable separation type lifting hot pot as claimed in claim 1, characterized in that: the lifting disc (6) comprises a disc (61) and a boss (62), the boss (62) is of a cylindrical hollow structure with a closed top, a plurality of round holes (611) are formed in the surface of the disc (61), and the disc (61) is connected with the bottom of the boss (62).</p> <p>5. A novel heatable separation type lifting hot pot as claimed in claim 1, characterized in that: elevating system (4) contain transfer screw (41), lifting screw (42), lifter (43), transfer screw (41) are located heating plate (13) under, the one end of transfer screw (41) links firmly in motor (5), lifting screw (42) set up directly over transfer screw (41) and perpendicular to transfer screw (41), lifting screw (42) are fixed on heating plate (13) and are connected with transfer screw (41) transmission through base (131), inside being helicitic texture and swivelling joint in lifting screw (42) of lifter (43), lifter (43) upper portion both sides are equipped with spacing platform (431), two spacing platform (431) are located directly over pot core (311).</p> <p>6. A novel heatable separation type lifting hot pot as claimed in claim 1, characterized in that: two locking buckle positions (111) are arranged at the upper part of the pot body (11).</p>
---	---

Table 4.10 - Example of Indexed Terms in One Patent Claims

Table 4.11 presents all 22 indexed terms in the claim above.

The extraction results for the mentioned claims are displayed in Figure 4.11 below. The algorithm successfully identified a total of 89 terms. Among these, 14 terms are indexed terms, indicated in blue in Table 4.11. Additionally, there is one candidate term, “火锅 (huǒ guō, ‘hot pot’)”, that is also considered a good term.

Indexed Terms	Pinyin	Translation
锅体	guō tǐ	pot body
锅盖	guō gài	pot cover
锅胆	guō dǎn	pot courage
升降机构	shēng jiàng jī gòu	elevating system
马达	mǎ dá	motor
升降盘	shēng jiàng pán	lifting disk
锅身	guō shēn	pot body
锅底座	guō dǐ zuò	pot base
发热盘	fā rè pán	heating plate
马达座	mǎ dá zuò	motor base
底座	dǐ zuò	base
内胆	nèi dǎn	cooker liner
手柄	shǒu bǐng	handles
锅芯	guō xīn	pot core
锁紧耳扣	suǒ jǐn ěr kòu	locking lugs
圆盘	yuán pán	disc
凸台	tū tái	boss
圆孔	yuán kǒng	round holes
传送螺杆	chuán sòng luó gǎn	transfer screw
升降螺杆	shēng jiàng luó gǎn	lifting screw
升降杆	shēng jiàng gān	lifter
限位台	xiàn wèi tái	spacing platform

Table 4.11 - List of Indexed Terms in the Patent Claims Above

At this stage, prior to the application of any ranking method, the precision stands at 16.30%, while the recall is 63.64%.

In the subsequent section, we will conduct a more comprehensive evaluation of the extraction results using the C-value metric, comparing it with a conventional POS tagging-based term recognizer.

Finally, it should be emphasized that this method, relying on the synthesis of syntactic units, is not suited for isolated terms, which are commonly encountered in chemistry, such as individual chemical elements, e.g. “铀 (yóu, uranium)”.

- 1.一种新型可加热分离型升降火锅，它包含锅体(1)、锅盖(2)，其特征在于：它还包含锅胆(3)、升降机构(4)、马达(5)、升降盘(6)，所述的锅体(1)由锅身(11)、锅底座(12)、发热盘(13)组成，锅底座(12)上部设置有马达座(121)，马达(5)固定连接于马达座(121)，发热盘(13)设置在锅底座(12)正上方且位于锅身(11)内部，升降机构(4)设置在锅底座(12)内部中间部位并与马达(5)连接，锅胆(3)套在升降机构(4)上并在发热盘(13)的正上部，升降盘(6)套在升降机构(4)的上部，锅盖(2)盖在锅胆(3)上部。
- 2.根据权利要求1所述的一种新型可加热分离型升降火锅，其特征在于：所述的发热盘(13)为圆形中空结构，发热盘(13)中空部位设置有底座(131)。
- 3.根据权利要求1所述的一种新型可加热分离型升降火锅，其特征在于：所述的锅胆(3)包含内胆(31)和手柄(32)，内胆(31)中心设置有锅芯(311)，锅芯(311)为圆柱形中空结构，锅芯(311)在内胆(31)正上方并垂直连接于内胆(31)，两个手柄(32)设置在内胆(31)上方的左右两侧，两个手柄(32)上设置有锁紧耳扣(321)。
- 4.根据权利要求1所述的一种新型可加热分离型升降火锅，其特征在于：所述的升降盘(6)包含圆盘(61)和凸台(62)，凸台(62)为顶部封闭的圆柱形中空结构，圆盘(61)表面设有若干圆孔(611)，圆盘(61)与凸台(62)底部连接。
- 5.根据权利要求1所述的一种新型可加热分离型升降火锅，其特征在于：所述的升降机构(4)包含传送螺杆(41)、升降螺杆(42)、升降杆(43)，传送螺杆(41)位于发热盘(13)的正下方，传送螺杆(41)的一端固连于马达(5)，升降螺杆(42)设置在传送螺杆(41)正上方且垂直于传送螺杆(41)，升降螺杆(42)通过底座(131)固定在发热盘(13)上并与传送螺杆(41)传动连接，升降杆(43)内部为螺纹结构并旋转连接于升降螺杆(42)，升降杆(43)上部两侧设有限位台(431)，两个限位台(431)位于锅芯(311)正上方。
- 6.根据权利要求1所述的一种新型可加热分离型升降火锅，其特征在于：所述的锅身(11)上部设置有两个锁紧扣位(111)。

Figure 4.13 - The Extraction Results

### 4.3.3 Comparison with the POS Matcher

To assess the term recognizer’s outcomes, we conduct a comparison with the SpaCy matcher, which operates based on POS tagging for term recognition.

In a prior study on the patent abstracts in the field of new energy vehicles by Zeng et al. (2016), they introduced POS combinations ranging from bigrams to hexagrams (as depicted in Table 4.14). We adopt a similar approach, but restrict the combinations to "n + n", "v + n", and "n + n + n" (outlined in Table 4.13) due to the absence of annotations equivalent to "b," which signifies a "distinguishing word".

n + n	26 245	v + n	23 457
n + n + n	19 743	n + b + n	18 209
b + m + n + n	8267	b + n + n + n	6246
n + n + n + vn + n	1769	m + n + b + vn + n	2021
b + n + b + n + n + n	321	n + n + u + b + vn + n	145

**Figure 4.14 - The POS patterns for term recognition (Zeng et al. 2016), in which “n” represents “noun”, “v” represents “verb”, “vn” represents “gerund”, “b” represents “distinguishing word”, “m” represents “number” and “u” represents “auxiliary”.**

n
n + n
v + n
n + n + n

**Table 4.13 - Combination of POS Tags Used in the SpaCy Matcher**

As a result, we have identified a total of 84 terms extracted from the patent claims text discussed in Section 4.3.2. To facilitate a comparison of the results, we employ the C-value (Frantzi et al. 2000) ranking method. This method is commonly utilized for ranking extracted terms in Automatic Term Extraction (ATE) scenarios.

As depicted in Figure 4.15, the calculation of the C-value consists of two components: the frequency of the term itself and the number of occurrences when it is nested within other longer terms. In other words, the C-value assigns value to words that are not nested within others.

$$C\text{-value}(a) = \begin{cases} \log_2 |a| \cdot f(a) & a \text{ is not nested,} \\ \log_2 |a| (f(a) - \frac{1}{P(T_a)} \sum_{b \in T_a} f(b)) & \text{otherwise} \end{cases}$$

Figure 4.15 - (Frantzi et al. 2000)<sup>129</sup>

Table 4.14 (a) displays the top 30 terms extracted through POS tagging along with their frequency and C-value, while Table 4.14 (b) presents the top 30 terms extracted via dependency parsing based on their C-value.

All indexed terms are underlined. The errors (including broken terms such as “型锅 (xíng guō, type pot)” in Table 4.14 (b), incorrect combinations such as “包含锅胆 (bāo hán guō dǎn, containing pot courage)” in Table 4.14 (a), and errors involving parentheses such as “发热盘 ( fā rè pán, heating plate)” in Table 4.14 (b)) of extraction are marked in red, and the general patent-related terms are marked in green.

Candidate Term	Pinyin	Translation	Frequency	C-value
加热分离型	jiā rè fēn lí xíng	‘heating separation type’	6	13.9
升降火锅	shēng jiàng huǒ guō	‘elevated hot pot’	6	12.0
<u>传送螺杆</u>	chuán sòng luó gān	‘transfer screw’	6	12.0
特征	tè zhēng	‘character’	6	6.0
圆柱形中空结构	yuán zhù xíng zhōng kōng jié gòu	‘cylindrical hollow structure’	2	5.6
上方	shàng fāng	‘above’	5	5.0
权利	quán lì	‘right’	5	5.0
所述	suǒ shù	‘said’	5	5.0
<u>发热盘</u>	fā rè pán	‘heating plate’	4	4.8
上部	shàng bù	‘upper part’	5	4.0
马达座(	mǎ dá zuò (	‘motor base (’	2	4.0
<u>升降机构</u>	shēng jiàng jī gòu	‘elevating system’	2	4.0
机构	jī gòu	‘system’	5	3.0
凸台	tū tái	‘boss’	3	3.0
内部中间部位	nèi bù zhōng jiān bù wèi	‘internal middle part’	1	2.6
垂直连接于	chuí zhí lián jiē yú	‘vertically connected to’	1	2.3

<sup>129</sup> The formula f(a) stands for the frequency of the word a.

包含锅胆	bāo hán guō dǎn	‘includes pot courage’	1	2.0
锅胆	guō dǎn	‘pot courage’	3	2.0
锅身	guō shēn	‘pot body’	3	2.0
位于锅身	wèi yú guō shēn	‘located on the pot body’	1	2.0
内部	nèi bù	‘inside part’	3	2.0
锅盖(2	guō gài (2	‘pot cover (2’	1	2.0
发热盘(	fā rè pán (	‘heating plate (’	1	2.0
中空结构	zhōng kōng jié gòu	‘hollow structure’	3	2.0
结构	jié gòu	‘structure’	4	2.0
内胆	nèi dǎn	‘inner liner’	2	2.0
锅芯	guō xīn	‘pot core’	3	2.0
为圆柱形	wèi yuán zhù xíng	‘is cylindrical’	1	2.0
左右两侧	zuǒ yòu liǎng cè	‘left and right sides’	1	2.0
升降盘(	shēng jiàng pán (	‘lifting disc’	1	2.0

Table 4.14 (a) - Terms Extracted by POS Tagging

Candidate Term	Pinyin	Translation	Frequency	C-value
传送螺杆	chuán sòng luó gān	‘transfer screw’	5	10.0
可加型	kě jiā xíng	‘agglutinable type’	6	9.5
型锅	xíng guō	‘type pot’	7	7.0
新型	xīn xíng	‘new type’	6	6.0
升锅	shēng guō	‘rise pot’	6	6.0
火锅	huǒ guō	‘hot pot’	6	6.0
底座	dǐ zuò	‘base’	6	6.0
发盘	fā pán	‘distribute plate’	6	6.0
螺杆	luó gān	‘screw’	8	5.8
型升	xíng shēng	‘type rise’	5	5.0
螺杆上方	luó gān shàng fāng	‘top of screw’	2	4.0
升螺杆	shēng luó gān	‘rise screw’	2	3.2
锅胆	guō dǎn	‘pot courage’	4	3.0
锅身	guō shēn	‘pot body’	3	3.0
锅座	guō zuò	‘pot base’	3	3.0
锅芯	guō xīn	‘pot core’	4	3.0

圆盘	yuán pán	‘disc’	3	3.0
凸台	tū tái	‘boss’	3	3.0
升盘	shēng pán	‘rise plate’	2	2.0
身部	shēn bù	‘body part’	2	2.0
升机	shēng jī	‘rise machine’	3	2.0
马胆	mǎ dǎn	‘horse courage’	2	2.0
中空部位	zhōng kōng bù wèi	‘hollow part’	1	2.0
手柄	shǒu bǐng	‘handles’	3	2.0
升降机构	shēng jiàng jī gòu	‘elevating system’	1	2.0
机构升降	jī gòu shēng jiàng	‘system elevating’	1	2.0
升杆	shēng gān	‘rise lever’	2	2.0
发热螺杆	fā rè luó gān	‘heating screw’	1	2.0
锅芯上方	guō xīn shàng fāng	‘top of pot core’	1	2.0
锅体锅	guō tǐ guō	‘pot body pot’	1	1.6

**Table 4.14 (b) - Terms Extracted by Character-level Dependency Parsing**

To assess the quality of term extraction, we conducted a manual evaluation of precision for each of the 30 terms. Our evaluation considered not only the presence of terms in the index list (#precision 1) but also the percentage of correctly extracted terms, especially those appear repetitively in the patent texts (#precision 2).

In Tables 4.14 (a) and 4.14 (b), we can see that 7 terms are correctly recognized by POS tagging, while 10 indexed terms and 11 correct terms (including “火锅 (huǒ guō, ‘hot pot’)”) recognized by the character-level dependency parsing.

As result, for the extraction by POS tagging:

$$\#precision\ 1 = \#precision\ 2 = 23.33\%$$

$$\#recall = 31.81\%$$

And for the extraction by dependency parsing:

$$\#precision\ 1 = 33.33\%$$

$$\#precision\ 2 = 36.67\%$$

$$\#recall = 45.45\%$$

## Chapter 5 - Lexical Variation with the Construction of a Patent-related Taxonomy

This chapter delves into the innovative approach of utilizing bilingual taxonomies to hierarchize technical terms. The focal point of this exploration is the construction, enhancement, and assessment of a comprehensive, domain-specific technical taxonomy rooted in the International Patent Classification (IPC) system.

Two primary reasons propelled the creation of this novel taxonomy. Firstly, the absence of an open-resource patent-related taxonomy prompted the need for its development. Existing taxonomies proved insufficient, prompting the necessity for a taxonomy tailored specifically for patents.

Additionally, as demonstrated in Section 1.1.1.2, the research has highlighted the limited availability of suitable taxonomies for patent-related applications. In the domain of patent drafting, hypernyms assume a crucial role. Currently, most prior efforts in lexical variation and term substitution rely on word similarity within distributional spaces to execute replacements (Zhou et al., 2019). However, this approach can only offer a list of semantically related terms, which fails to meet our requirement for the generalization of terms in patent claims.

The taxonomy's construction rests upon dual objectives. Firstly, it is crafted to serve practical applications by offering a structured repository of patent-related terms. Secondly, it serves linguistic interests by enabling in-depth taxonomy analysis.

Redefining the conventional notion of a “taxonomy”, we adopt the perspective that taxonomies are design science artefacts employed by researchers and practitioners to categorize and elucidate objects within a given domain. Our focus is on patent-specific terms that cater to the intricacies of this specialized field.

A technical taxonomy can provide attorneys with a standardized vocabulary and a clear framework for organizing and describing the components and functions of an invention, which can facilitate the drafting process and increase the accuracy and effectiveness of patent applications. Furthermore, the use of a technical taxonomy can help attorneys identify potential gaps or weaknesses in their clients' inventions and suggest improvements or modifications to enhance the patentability and marketability of the invention.

Our methodology entails identifying relevant IPC classes pertinent to a particular domain, like IPC section A (Human Necessity), and extracting terminological lists in a hierarchical structure from their class titles (Section 5.2). To augment the taxonomy, we incorporate additional terms using a pre-trained hypernym generator detailed in Section 5.3.1, creating a more comprehensive resource, with the objective of performing an oriented and dynamic lexical variation in Section 5.3.2.

Central to this endeavour is the application of our taxonomy to patent tech-mining. The taxonomy provides a valuable list of hypernym and hyponym terms, essential for enhancing lexical substitution capabilities.



This chapter is structured as follows: Section 5.1 provides an introduction to the International Patent Classification schema and the rationale behind our new IPC-based technical taxonomy. Section 5.2.1 delves into the process of transforming IPC titles into comprehensive technical terms stored within tree structures and the subsequent refinement process. Section 5.2.2 evaluates the quality and usefulness of the taxonomy. Lastly, Section 5.3 outlines the taxonomies-based oriented dynamic lexical variation.

This research has been done in collaboration with Zuo You from team Almanach from Inria, who first built the English CPC version of the described technical taxonomies and trained the hypernym generation models.

## 5.1 Leveraging Technical Knowledge in IPC Titles

In this section, we aim to address several important questions regarding the construction of the new IPC-based technical taxonomy. Based on the introduction of the original format of the IPC titles, we analyze the characteristics of potential technical terms within the original IPC titles to provide insights and answers to these questions. Specifically, we explore the following aspects:

- Characteristics of potential technical terms: We examine the composition and structure of IPC titles to identify common patterns and characteristics that indicate potential technical terms.
- Justification for basing the taxonomy on IPC: We explain why using the IPC as the foundation for the technical taxonomy is advantageous compared to other existing resources. We highlight the comprehensiveness and standardized nature of the IPC and its suitability for capturing domain-specific technical terms.
- Expected processing results: We discuss the anticipated outcomes of the processing techniques applied to extract technical terms from the IPC titles. This includes the identification of relevant hypernymy and hyponymy relations and the construction of a comprehensive technical taxonomy.
- Use case and practical utility: We demonstrate the usefulness of the IPC-based technical taxonomy in various applications, such as patent classification, passage retrieval, and keyword extraction. We emphasize how the taxonomy enhances the efficiency and accuracy of these tasks by providing a structured and domain-specific knowledge framework.
- Decision on the form of the taxonomy: We discuss the linguistic and practical considerations involved in representing the extracted technical terms. This decision encompasses issues such as the representation format (e.g., as individual terms or compound terms) and the organization of the taxonomy for optimal usability.
- Methodology for term extraction: We outline the approach and techniques employed for extracting technical terms from the IPC titles. This involves leveraging linguistic patterns, syntactic analysis, and semantic relationships to identify and extract relevant terms.

### 5.1.1 The Original Format of the IPC Titles

Our study uses the parallel IPC (edition 2016) in English and Chinese, which represents the majority of the world's patent applications<sup>130</sup> and can be obtained from the World Intellectual Property Organization (WIPO)<sup>131</sup> and the official website of the China National Intellectual Property Administration (CNIPA)<sup>132</sup>, respectively.

---

<sup>130</sup> According to the World Intellectual Property Organization (WIPO) 2020 report ([https://www.wipo.int/edocs/pubdocs/en/wipo\\_pub\\_941\\_2020.pdf](https://www.wipo.int/edocs/pubdocs/en/wipo_pub_941_2020.pdf)), out of a total of 275,900 PCT (Patent Cooperation Treaty) applications filed, English was the most commonly used language, accounting for 62.4% of all applications, while Chinese accounted for 23.8% of all applications. These two languages combined make up more than 86% of all PCT applications filed in 2020.

<sup>131</sup> <https://www.wipo.int/classifications/ipc/en/ITsupport/Version20160101/index.html>

<sup>132</sup> [https://www.cnipa.gov.cn/art/2016/8/31/art\\_2152\\_152142.html](https://www.cnipa.gov.cn/art/2016/8/31/art_2152_152142.html)

The decision to use the International Patent Classification (IPC) as the source of our taxonomy was based on several factors. Firstly, the IPC is an internationally recognized and widely used patent classification system. It provides a comprehensive and standardized framework for organizing and categorizing patent documents. The titles of each class within the IPC contain valuable information about the subject matter and technical domains covered by the patents.

The Chinese version of the IPC is an official translation of the original English version provided by the World Intellectual Property Organization (WIPO). By aligning the Chinese titles with their corresponding English titles, we were able to create a parallel corpus that facilitated the construction of our taxonomy.

The IPC is organized into eight sections as presented in Section 1.1.2, each covering a specific domain of technology. These sections include

- A. Human necessities
- B. Performing operations; transporting
- C. Chemistry; metallurgy
- D. Textiles; paper
- E. Fixed constructions
- F. Mechanical engineering; lighting; heating; weapons; blasting engines or pumps
- G. Physics
- H. Electricity

Within each section, there are classes, subclasses, groups, and subgroups that further refine the classification. The titles associated with these classifications are expressed in the form of noun phrases, participle phrases, or prepositional phrases, providing descriptive information about the subject matter of the patents. The subgroups in the IPC version of 2016 comprise approximately 70,000 classification entries.

In our research, we primarily rely on the International Patent Classification (IPC) as the main patent classification system (more details in Section 1.1.2). However, it is worth acknowledging the existence of alternative systems, such as the Cooperative Patent Classification (CPC). The CPC is a comprehensive classification system jointly managed by the European Patent Office (EPO) and the United States Patent and Trademark Office (USPTO). The Chinese Patent Office joined this initiative in 2016, expanding its scope and coverage. With over 300,000 entries, the CPC offers a larger collection of classifications compared to the IPC. However, during the course of our study, we encountered limitations in accessing complete translations of the CPC into Chinese. Consequently, we made the decision to prioritize the use of the IPC, which provided a more comprehensive and readily available official Chinese translation.

The following Table 5.1 gives an example of the original IPC titles in Section A - HUMAN NECESSITIES. As shown in Table, the IPC titles are organized in a hierarchic structure, containing different levels, such as “Sections” (e.g. A), “Classes” (e.g. A01), “Subclasses” (e.g. A01B), and “Groups”, which are subdivided into “Main Groups” (marked as “level 0”, e.g. A01B 1/00) and “Subgroups” (marked as “level 1” and so on, until “level 9”<sup>133</sup> in edition 2016.01, e.g. A01B 1/02 and A01B 1/04).

---

<sup>133</sup> In the 2016.01 version, the only subgroup on level 9 is “G09G3/325 the data current flowing through the driving transistor during a setting phase, e.g. by using a switch for connecting the driving transistor to the data driver”.

A	-3	HUMAN NECESSITIES	人类生活必需
A01	-2	AGRICULTURE; FORESTRY; ANIMAL HUSBANDRY; HUNTING; TRAPPING; FISHING	农业;林业;畜牧业;狩猎;诱捕;捕鱼
A01B	-1	SOIL WORKING IN AGRICULTURE OR FORESTRY; PARTS, DETAILS, OR ACCESSORIES OF AGRICULTURAL MACHINES OR IMPLEMENTS, IN GENERAL	农业或林业的整地;一般农业机械或农具的部件、零件或附件
A01B 1/00	0	Hand tools	手动工具
A01B 1/02	1	Spades; Shovels	锹;铲
A01B 1/04	2	with teeth	带齿的

**Table 5.1 - Example of the general organization of the original IPC titles from Section A - HUMAN NECESSITIES<sup>134</sup>**

In the English version of the IPC, all three top levels (section titles, class titles, and subclass titles) are written in capital letters. However, for group titles, the capitalization may vary depending on the phrase types used in the title. This characteristic, unique to the English IPC, proves to be highly valuable for extracting terms in subsequent sections.

Additionally, besides capitalization information, there are several other language-specific features in English that can provide useful information for processing Chinese titles. It is for this reason that we construct our taxonomy using parallel bilingual titles, allowing us to leverage these English language-specific features for effective processing for Chinese.

Section	No. of classes	No. of subclasses	No. of main groups	No. of subgroups	Total no. of groups
A	16	84	1'132	7'763	8'895
B	38	169	1'992	14'930	16'922
C	21	87	1'321	13'187	14'508
D	9	39	350	2'700	3'050
E	8	31	323	3'115	3'438
F	18	97	1'072	7'705	8'777
G	14	81	696	7'426	8'122
H	6	51	548	8'326	8'874
Total	130	639	7'434	65'152	72'586

**Table 5.2 - Number of Titles in Each Section at Each Level of IPC (Edition 2016.01)**

<sup>134</sup> As the level code starts from 0 from the fourth level, the first three levels do not have a level code. We call them here from the top “level -3”, “level -2” and “level -1”.

On the official website of the World Intellectual Property Organization (WIPO), detailed statistics regarding the exact number of titles at each level of each section in the International Patent Classification (IPC) can be found. These statistics reveal an observed phenomenon of imbalance among the sections. It is specifically worth noting that Sections B and C contain approximately five times the number of group titles compared to Sections D and E. This discrepancy in the number of group titles suggests that there is a higher level of granularity and subdivision within the subject matters covered by Sections B and C.

### 5.1.2 Alignment of English-Chinese IPC Titles

To establish alignment between the Chinese version of IPC and the English version, a two-step process is employed.

- In the first step, the IPC text files for each of the eight sections are converted into a structured tree object. This conversion involves organizing the IPC titles/headings based on their IPC code level, resulting in the creation of a hierarchical structure. The hierarchical structure is the foundation for the subsequent stages of taxonomy construction.
- In the second step, the Chinese titles are aligned with their corresponding English titles that share the same IPC code. Par example, in the Table 5.3 the English term “from material of animal origin” and the Chinese term “从动物来源的材料(由微生物或生物化学获得的动物饲料入A23K\_10/10)” share the same IPC code “A23K 10/20”. This alignment enables the establishment of a clear correspondence between the Chinese and English versions of the IPC. This correspondence is crucial for the subsequent analysis and taxonomy construction, as it allows for leveraging the processing of the English titles to facilitate further steps involving the Chinese language.

Although the Chinese version of IPC titles is a direct translation from the original English file by the Patent Literature Department of the China National Intellectual Property Administration, the titles in Chinese do not always correspond to the English ones in a strict way. Table 5.3 shows some examples of this non-correspondence.

A23K 10/20	1	from material of animal origin	从动物来源的材料(由微生物或生物化学获得的动物饲料入A23K_10/10)
A23G 9/06	2	characterised by using carbon dioxide or carbon dioxide snow as cooling medium	组优先于A23G_09/08到A23G_09/14各组。
A01N63/00	0	Biocides, pest repellants or attractants, or plant growth regulators containing micro-organisms, viruses, microbial fungi, animals, e.g. nematodes, or substances produced by, or obtained from micro-organisms, viruses, microbial fungi or animals, e.g. enzymes or fermentates	含有微生物、病毒、微生物真菌、动物
A61Q1/02	1	Preparations containing skin colorants, e.g. pigments	含皮肤色料

A23K 10/20	1	from material of animal origin	从动物来源的材料(由微生物或生物化学获得的动物饲料入A23K_10/10)
B01D24/00	0	Filters comprising loose filtering material, i.e. filtering material without any binder between the individual particles or fibres thereof	含有疏松材料的过滤器, 例如, 在单个粒子或纤维间没有黏合剂的过滤材料
A61B17/3207	3	Atherectomy devices	粥样斑块环切(Atherectomy)装置
A47C1/14	1	Beach chairs	沙(海)滩用椅
A47B77/08	2	for incorporating apparatus operated by power, including water power; for incorporating apparatus for cooking, cooling, or laundry purposes	与用动力(包括水力)操作的装置相结合的;与烹调、冷却或洗涤装置相结合的

**Table 5.3 - Non-correspondence between the original English IPC titles and the official Chinese translation**

As shown above, the types of non-correspondence vary from minor issues during processing (e.g. A23K 10/20) to severe structural problems that may affect severely the quality of the final taxonomy (e.g. A01N63/00).

There are in general five types of non-correspondence:

- missing or additional parentheses (e.g. A23K 10/20, A47C1/14, A47B77/08)
- English translation in parentheses (e.g. A61B17/3207)
- “e.g.” and “i.e.” (e.g. B01D24/00)
- missing examples following “e.g.” (e.g. A61Q1/02)
- mismatch in content (e.g. A23G 9/06, A01N63/00)

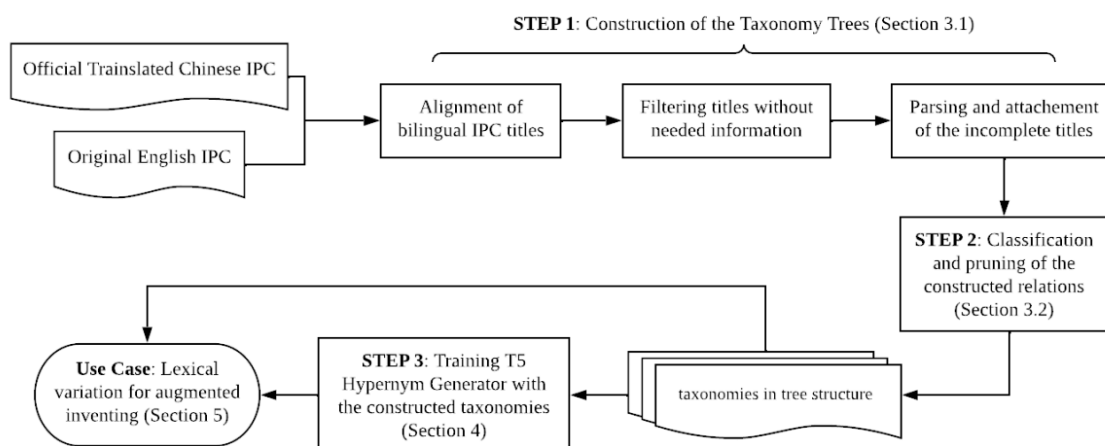
Among which, while the three first can be solved certain automated processes, the mismatch in content and missing examples following “e.g.” are two sources of errors introduced by the transformation into tree structures in Section 5.2.

## 5.2 Mining Hypernyms/Hyponyms in IPC Titles

In this section, we outline the process of building a tree structure for patent-related terms using IPC titles. The entire procedure comprises three steps (refer to Figure 5.2):

1. The creation of taxonomy trees.
2. The classification and refinement of constructed hypernyms/hyponyms relations.
3. The training of a hypernym generator based on the established taxonomies.

The initial part of STEP 1, involving the alignment of bilingual IPC titles, has already been addressed in Section 5.1.2. In this section, we shift our focus to the subsequent steps, which involve filtering inappropriate titles, eliminating irrelevant information from the titles, and supplementing incomplete titles.



**Figure 5.2 - The Process for Lexical Variation**

### 5.2.1 The Construction of the Tree Structure

The creation of taxonomy trees itself consists of three parts: (1) Filtering inappropriate titles or irrelevant information in the titles (Section 5.2.1.1); (2) Extraction of technical expressions (Section 5.2.1.2); (3) Nominalization of incomplete technical expressions (Section 5.2.1.3).

After experimenting with different depths of level, we limited the depth of the tree to eight levels (ipc8, marked as “level 1”). The ipc8 version contains 29,626 titles in total.

## 5.2.1.1 Filtering Inappropriate Titles or Irrelevant Information in the Titles

The original IPC primarily serves the purpose of categorizing patents and utility models based on their specific technological domains. In this capacity, it incorporates categories that establish certain relationships or references to other IPC categories, as demonstrated in Table 5.4. Although these relationships are valuable for classification purposes, they do not provide relevant terminology taxonomy information. To proceed after completing the alignment process outlined in the previous step, our next task involves removing unnecessary information from the aligned IPC titles.

In this section, we provide a detailed analysis of the IPC titles in order to extract the relevant technical terms for our study while excluding irrelevant information. To achieve this, we apply a filtering process based on specific types of frequently encountered unwanted expressions. These expressions include IPC reference codes enclosed in parentheses, abbreviations and translations within brackets or parentheses, as well as certain determiners, pronouns that are not appropriate for inclusion in terms, and so on.

We present the structure of each type along with examples and provide a list of keywords that can be used to identify these unwanted expressions. By implementing this filtering approach, we aim to streamline the extraction of the desired technical terms and minimize computational efforts.

The following Tables 5.4 and 5.5 give examples of the two main types of expressions involving IPC reference codes in the original IPC titles.

A01B	0	SOIL WORKING IN AGRICULTURE OR FORESTRY; PARTS, DETAILS, OR ACCESSORIES OF AGRICULTURAL MACHINES OR IMPLEMENTS, IN GENERAL (making or covering furrows or holes for sowing, planting or manuring A01C_05/00; machines for harvesting root crops A01D; mowers convertible to soil working apparatus or capable of soil working A01D_42/04; mowers combined with soil working implements A01D_43/12; soil working for engineering purposes E01, E02, E21)	农业或林业的整地; 一般农业机械或农具的部件、零件或附件(用于播种、种植或施厩肥的开挖沟穴或覆盖沟穴入A01C_05/00; 收获根作物的机械入A01D; 可变换成整地设备或能够整地的割草机入A01D_42/04; 与整地机具联合的割草机入A01D_43/12; 工程目的的整地入E01, E02, E21)
A01B 01/00	1	Hand tools (edge trimmers for lawns A01G_03/06)	手动工具(草坪修整机入A01G_03/06)
A23K 10/20	1	from material of animal origin	从动物来源的材料(由微生物或生物化学获得的动物饲料入A23K_10/10)
A61K 31/48	5	Ergoline derivatives, e.g. lysergic acid, ergotamine	麦角灵衍生物, 例如麦角酸、麦角胺
A61K 31/495	3	having six-membered rings with two nitrogen atoms as the only ring hetero atoms, e.g. piperazine (A61K_31/48 takes precedence);;	有仅以两个氮原子作为环杂原子的六元环, 例如哌嗪(A61K_31/48优先)
A61L 02/04	2	Heat (A61L_02/08 takes precedence);;	加热(A61L_02/08优先)
A61L 02/08	2	Radiation	辐照

**Table 5.4 (a) - Examples of Irrelevant Information in Parentheses and Brackets from Section A**



As shown in Table 5.4 (a), the first structure is composed of two parts: the principle part describing an object or a process, and the part enclosed in parentheses and brackets, describing in which specific classes or groups certain types of the object or process expressed in the principal part of the title should be included (e.g. A01B, A01B 01/00, A23K 10/20 in Table 5.4 (a)), or which classes or groups “takes precedence” of the current one (e.g. A61K 31/495, A61L 02/04 in Table 5.4 (a)). It is worth noting that in some cases, as in A23K 10/20, the parentheses can only appear in one language (only in the Chinese version here) due to the translation.

A61B5/055	2	involving electronic [EMR] or nuclear [NMR] magnetic resonance, e.g. magnetic resonance imaging	包含电磁共振[EMR]或核磁共振[NMR]的, 例如磁共振成像
A61B17/225	2	for extracorporeal shock wave lithotripsy [ESWL], e.g. by using ultrasonic waves	用于体外震动波碎石法[ESWL]的, 如利用超声波的
A61K38/24	4	Follicle-stimulating hormone (FSH); Chorionic gonadotropins, e.g. HCG; Luteinising hormone (LH); Thyroid-stimulating hormone (TSH)	促卵泡激素 (FSH); 绒毛膜促性腺激素, 例如HCG; 促黄体激素 (LH); 促甲状腺激素 (TSH)
A61K38/25	4	Growth hormone-releasing factor (GH-RF) (Somatoliberin)	生长激素释放因子 (GH—RF) (促生长素释放素)
A61K49/06	1	Nuclear magnetic resonance (NMR) contrast preparations; Magnetic resonance imaging (MRI) contrast preparations	核磁共振 (NMR) 造影剂; 磁共振 (MRI) 成像造影剂
A61H31/02	1	"Iron-lungs", whether or not combined with gas breathing means	带与不带气体呼吸装置的“铁肺”(人工呼吸器)
A61K36/03	2	Phaeophycota or phaeophyta (brown algae), e.g. Fucus	褐藻门(褐藻), 例如墨角藻属
A61B17/3207	3	Atherectomy devices	粥样斑块环切 (Atherectomy) 装置
A47C1/14	1	Beach chairs	沙(海)滩用椅
A47B77/08	2	for incorporating apparatus operated by power, including water power; for incorporating apparatus for cooking, cooling, or laundry purposes	与用动力(包括水力)操作的装置相结合的; 与烹调、冷却或洗涤装置相结合的
A61B16/00	0	Devices specially adapted for vivisection or autopsy (similar devices for medical purposes, see the relevant groups for such devices)	专门适合于活体解剖或尸体解剖用的器械(用于医疗目的类似器械见这类器械的有关组)
A61J3/00	0	Devices or methods specially adapted for bringing pharmaceutical products into particular physical or administering forms (chemical aspects, see the relevant classes)	专用于将药品制成特殊的物理或服用形式的装置或方法(化学方面见有关大类)

**Table 5.4 (b) - Examples of Other Possible Contents in Parentheses and Brackets from Section A**

Some other contents that may appear in the parentheses and brackets are shown in Table 5.4 (b): abbreviations in square brackets (e.g. A61B5/055, A61B17/225 in Table 5.4 (b)) and parathesis (e.g. A61K38/24, A61K38/25, A61K49/06 in Table 5.4 (b)), and in the case of Chinese, additional

explanations as well as corresponding original English terms of translated terminologies marked with parathesis (A61H31/02, A61K36/03, A61B17/3207 in Table 5.4 (b)).

A61K31/46	4	8-Azabicyclo [3.2.1] octane; Derivatives thereof, e.g. atropine, cocaine	8氮杂二环 [3, 2, 1]辛烷;其衍生物, 例如阿托品、可卡因
A61K31/43	6	Compounds containing 4-thia-1-azabicyclo [3.2.0] heptane ring systems, i.e. compounds containing a ring system of the formula , e.g. penicillins, penems	含4硫杂1氮杂双环 [3.2.0]庚烷环系的化合物, 即含下式环系的化合物: 例如青霉素、青霉烯
A61K38/44	3	Oxidoreductases (1)	氧化还原酶(1)
A61K38/45	3	Transferases (2)	转移酶(2)
A61K38/46	3	Hydrolases (3)	水解酶(3)
A61K38/47	4	acting on glycosyl compounds (3.2), e.g. cellulases, lactases	作用于糖基化合物(3, 2), 例如纤维素酶, 乳糖酶
A61K38/48	4	acting on peptide bonds (3.4)	作用于肽键(3, 4)
A61K38/49	5	Urokinase; Tissue plasminogen activator	尿激酶;组织纤溶酶原激活剂
A61K38/50	4	acting on carbon-nitrogen bonds, other than peptide bonds (3.5), e.g. asparaginase	作用于碳—氮键而不是肽键(3, 5), 例如天冬酰胺酶
A61K38/51	3	Lyases (4)	裂解酶(4)
A61K38/52	3	Isomerases (5)	异构酶(5)
A61K38/53	3	Ligases (6)	连接酶(6)
A61K31/15	2	Oximes (; CNO)Hydrazines (; NN)Hydrazones (; NN)	肟CON;肼NN;肼NN
A61K31/155	2	Amidines (), e.g. guanidine (H <sub>2</sub> NC(NH)NH <sub>2</sub> ), isourea (HNC(OH)NH <sub>2</sub> ), isothiourea (HNC(SH)NH <sub>2</sub> )	脒NCN, 例如脒(H <sub>2</sub> N—C(NH)—NH <sub>2</sub> )、异脒(HNC(OH)NH <sub>2</sub> )、异硫脒(NHC(SH)—NH <sub>2</sub> )
A61K31/175	3	having the group , NC(O)NN or , e.g. carbonohydrazides, carbazones, semicarbazides, semicarbazonesThioanalogues thereof	含NC(O)NN、NC(O)NN或NC(O)NN基团的, 例如卡巴肼、卡巴肼、半卡巴肼、半卡巴肼;其硫代类似物
A61K31/56	1	Compounds containing cyclopenta[a]hydrophenanthrene ring systems; Derivatives thereof, e.g. steroids	含环戊[a]氢化菲环系的化合物;其衍生物, 例如甾族化合物

**Table 5.4 (c) - Examples of Contents Involving Chemical Terms in Parentheses and Brackets from Section A**

Nonetheless, a subsequent issue arising from the removal of parentheses and brackets pertains to chemistry terms that also incorporate these symbols, as illustrated in the examples provided in Table

5.4 (c). This challenge persists throughout this study. However, such titles are typically found at deeper levels within the taxonomy, whereas our process is confined to the initial levels of IPC titles.

For this first type of IPC titles including IPC reference codes inside, we keep the principle part and eliminate the enclosed part, although they may also contain potential technical terms, which holds especially true for the first structure in Table 5.4 (a) (e.g. A01B, A01B 01/00, A23K 10/20), such as “machines for harvesting root crops”, “mowers convertible to soil working apparatus or capable of soil working”, “mowers combined with soil working implements” in A01B, and “edge trimmers for lawns” for A01B 01/00, which are identical in the IPC texts. More sophisticated processing will be reserved for future work.

The identification of this first type is by simply using the pattern to match parentheses with IPC codes inside. After the process, we eliminated 2,716 titles.

In the second type shown in Table 5.5, the title as a whole is about the inclusion of certain objects or processes in certain classes or groups (e.g. A01N 57/02, A01P 15/00, A23B 4/14 in Table 5.5).

Like the processing above, at the current stage, we eliminate the whole titles containing IPC codes here, in spite of the observation of potential technical terms, such as “Biocides for specific purposes” in A01P 15/00.

A01N 57/02	1	having alternatively specified atoms bound to the phosphorus atom and not covered by a single one of groups; A01N_57/10, A01N_57/18, A01N_57/26, A01N_57/34	具有与磷原子键合的并且不包含在A01N_57/10, A01N_57/18, A01N_57/26, A01N_57/34
A01P 15/00	0	Biocides for specific purposes not provided for in groups; A01P_01/00-A01P_13/00	在A01P_01/00至A01P_13/00组不包含的用于特殊目的的杀生剂
A23B 4/14	1	Preserving with chemicals not covered by groups; A23B_04/02; or A23B_04/12	用A23B_04/02或A23B_04/12小组不包含的化学品保存

**Table 5.5 - Examples of Unwanted Titles Containing IPC codes**

To match these titles, except the IPC codes, we also used a list of keywords in combination. After the process, we eliminated 1,775 titles.

After the title filter removed titles that were not relevant from a taxonomy perspective during preprocessing, certain strings in the remaining entries are still not part of the technical expression. These include explanations or descriptions marked with “类目 (lèi mù, ‘groups’)”, and “零件 (líng jiàn, ‘details, accessories’)”, etc. Table 5.6 lists different types of unwanted strings along with the eliminated title number of each type.

Chinese Keywords	English Keywords	Number
“类目” “组目”	“subclass”, “groups”, “subgroup”	256
“零件” “细节”	“details”, “accessories”, “tools”	531
“一般”	“in general”	238
“或类似构件” “或类似物” “或其他部分/或其他部位/或其他部件” “在其他类未列入的”	“or like elements”, “or the like”, “or other parts” “not provided for elsewhere”	2,861

**Table 5.6 - Keywords and Number of the Remained Irrelevant Information in the IPC Titles**

#### 5.2.1.2 Extraction of Technical Expressions

Within this subsection, we employ heuristic rules and syntactic analysis to extract pertinent information from IPC titles, which will be instrumental in the creation of taxonomies.

This second step also consists of two parts: (1) The separation of conjunction by semicolons; and (2) The extraction of examples within titles as hyponyms.

To separate conjuncts (in Table 5.7), we divide those connected by semicolons in both English and Chinese. Conjunctions linked by simple commas or coordinating conjuncts such as 和 (hé, ‘and’) and 或 (huò, ‘or’) remain unaltered. It's important to note that we only split conjunctions of the second type at the lowest level of the taxonomy trees to ensure clarity when determining the syntactic head for attaching a conjunct.

And from an application perspective, not only simple terms but also expressions containing conjunctions can be found in the real patent corpus, these long expressions are considered useful to a technical taxonomy.

Errors related to the separation of elements within IPC titles can introduce significant challenges and inaccuracies in the process of creating taxonomies. One common issue arises from the absence of the semicolon in English IPC titles, exemplified by entries like “A01N3/00” and “A01M” in Table 5.7. This missing punctuation can disrupt the correct segmentation of terms, making it challenging to identify individual components within the title accurately. Another error occurs when different separation marks are mixed with the semicolon, creating inconsistencies and confusion in the title structure. Furthermore, translation inconsistencies can contribute to errors, especially when corresponding terms in different languages do not align correctly.

A01	-2	AGRICULTURE; FORESTRY; ANIMAL HUSBANDRY; HUNTING; TRAPPING; FISHING	农业;林业;畜牧业;狩猎;诱捕;捕鱼
A01B	-1	SOIL WORKING IN AGRICULTURE OR FORESTRY; PARTS, DETAILS, OR ACCESSORIES OF AGRICULTURAL MACHINES OR IMPLEMENTS, IN GENERAL	农业或林业的整地;一般农业机械或农具的部件、零件或附件
A01B15/02	1	Plough blades; Fixing the blades	犁刀;固定犁刀的
A01B69/00	0	Steering of agricultural machines or implements; Guiding agricultural machines or implements on a desired track	农业机械或农具的转向机构;在所要求的轨道上导引农机具
A01C3/02	1	Storage places for manure, e.g. cisterns for liquid manure; Installations for fermenting manure	厩肥的贮存地点,如贮存液体厩肥的罐车;厩肥的发酵装置
A01B33/08	1	Tools; Details, e.g. adaptations of transmissions or gearings	工作部件;零件,例如传动装置或齿轮装置
A61L15/32	3	Proteins, polypeptides; Degradation products or derivatives thereof, e.g. albumin, collagen, fibrin, gelatin	蛋白质、多肽;它们的降解产物或衍生物,例如白蛋白、胶原蛋白、纤维蛋白、明胶
A01N3/00	0	Preservation of plants or parts thereof, e.g. inhibiting evaporation, improvement of the appearance of leaves Grafting wax	植物或其局部的保存,如抑制蒸发、改进叶子的外观;接蜡
A01M	-1	CATCHING, TRAPPING OR SCARING OF ANIMALS; APPARATUS FOR THE DESTRUCTION OF NOXIOUS ANIMALS OR NOXIOUS PLANTS	动物的捕捉、诱捕或惊吓;消灭有害动物或有害植物用的装置

**Table 5.7 - Examples of Titles as Subjects of Separation**

The other part involves the extraction of example terms as children nodes.

A great number of titles contain examples of the described technical means or instances, normally following the example markers “例如 (lì rú)”, “诸如 (zhū rú)”, “比如 (bǐ rú)” and “如 (rú)” (“e.g.” and “such as” in English) at the end or in the middle of the title (Table 5.8). These examples are viewed as potential subcategories or specific cases related to the means or instances discussed in the main part of the title. We extract these examples as child nodes.

A01C7/16	2	Seeders with other distributing devices, e.g. brushes, discs, screws, slides	带其他撒布装置的播种机,例如用刷子、圆盘、螺旋、滑板的
A61G10/02	1	with artificial climate; with means to maintain a desired pressure, e.g. for germ-free rooms	具有人工气候的治疗室;具有保持所需气压装置的治疗室,如用于无菌室的
B65H9/18	1	Assisting by devices such as reflectors, lenses, transparent sheets, or mechanical indicators	用诸如反射器、透镜、透明薄片或机械指示器等装置协助的

A01C7/16	2	Seeders with other distributing devices, e.g. brushes, discs, screws, slides	带其他撒布装置的播种机, 例如用刷子、圆盘、螺旋、滑板的
A61G5/12	2	Rests specially adapted therefor, e.g. for the head or feet	专用支托, 如用于头或脚的
A47J43/04	1	Machines for domestic use not covered elsewhere, e.g. for grinding, mixing, stirring, kneading, emulsifying, whipping or beating foodstuffs, e.g. power-driven	未列入其他类目的家用机械, 如用于对食品进行研磨、混合、搅拌、揉合、乳化、搅打类的, 如动力驱动的
A61B1/24	1	for the mouth, i.e. stomatoscopes, e.g. with tongue depressors	口腔用的, 即口腔镜, 如带压舌板的口腔镜
A61K31/455	5	Nicotinic acid, i.e. niacin; Derivatives thereof, e.g. esters, amides	烟酸, 即烟碱酸; 其衍生物, 例如酯, 酰胺
A61P25/18	1	Antipsychotics, i.e. neuroleptics; Drugs for mania or schizophrenia	抗精神病药, 例如神经阻滞剂; 用于治疗狂躁或精神分裂症的药物

Table 5.8 - Examples of IPC Titles Containing Examples and Instances

To achieve this, we employ a method where we split the title at the markers (Table 5.9), and any segments that appear after it are considered as child nodes linked to the preceding segments. It's worth noting that in certain titles, the marker “即 (i.e.)” is sometimes mixed with example markers and is followed by examples and instances. In the examples of “A61B1/24” and “A61K31/455” in Table 5.8, “i.e.” is directly followed by instances, and in “A61P25/18”, the English version uses “i.e.” while the Chinese translation is “例如 (e.g.).”

Chinese Keywords	English Keywords	Number
(例如 诸如 比如 如)[^;\n即如]+	e.g. such as	10,755
即[^;\n即如]+	i.e.	1,177

Table 5.9 - The Keywords for Examples and Instances

### 5.2.1.3 Nominalization of Incomplete Technical Expressions

Following the process of splitting lengthy expressions with lists or examples into separate nodes and making necessary adjustments to their relationships based on the enclosed information, the subsequent step revolves around converting incomplete results into nominal expressions. This transformation involves the use of syntactic parsing to pinpoint titles that lack syntactic completeness. We employ SpaCy, a dependency parser that demonstrates superior performance in English compared to Chinese, to detect all English titles characterized by a syntactic head that is not a noun. We then attach them to their parent node title. Additionally, we identify their corresponding Chinese titles that necessitate attachment to ensure the accuracy and completeness of the taxonomy development process.

Titles that are not capitalized typically represent incomplete phrases (refer to Table 5.10 for examples). In the English version of IPC, when an expression starts with a lowercase letter, it signifies that it should be appended to the expressions of the parent node. This decision is straightforwardly implemented based on English expressions since Chinese does not employ capitalization. In English, participial or prepositional phrases are consistently added directly at the end of their parent titles. However, in Chinese, these phrases are added in front of the parent titles and connected using the word “的 (de, meaning ‘of’)”. In most cases, the particle “的 (de, ‘of’)” serves as the connector. Nevertheless, for phrases that commence with specific expressions like “通过 (tōng guò, ‘by means/way of’)”, an additional term “来 (lái, ‘come’)” is used instead of “的 (de, ‘of’)” to establish the connection between the phrase and its parent title. In certain cases, the connector “的” is missing at the end of incomplete titles, as exemplified by “A47H13/01” in Table 5.10. In such instances, it becomes necessary to automatically add the connector “的” to complete the title.

A01B1/04	2	with teeth	带齿的
A47H13/01	1	by clamps; by clamps attached to hooks or rings	用夹子;用附属于钩或环的夹子
A01B1/24	1	for treating meadows or lawns	处理草地或草坪用的
A01B59/042	2	having pulling means arranged on the rear part of the tractor	牵引装置安装在拖拉机尾部的
A01B63/08	2	operated by the movement of the tractor	由拖拉机的运动操作的
A61G17/007	1	characterised by the construction material used, e.g. biodegradable material; Use of several materials	以所使用的结构材料为特征的, 例如生物防腐材料;使用几种材料

Table 5.10 - Examples of Incomplete Titles

Two examples of the constructed taxonomies at the end of the above processing are shown in Figure 5.3 and Figure 5.4.

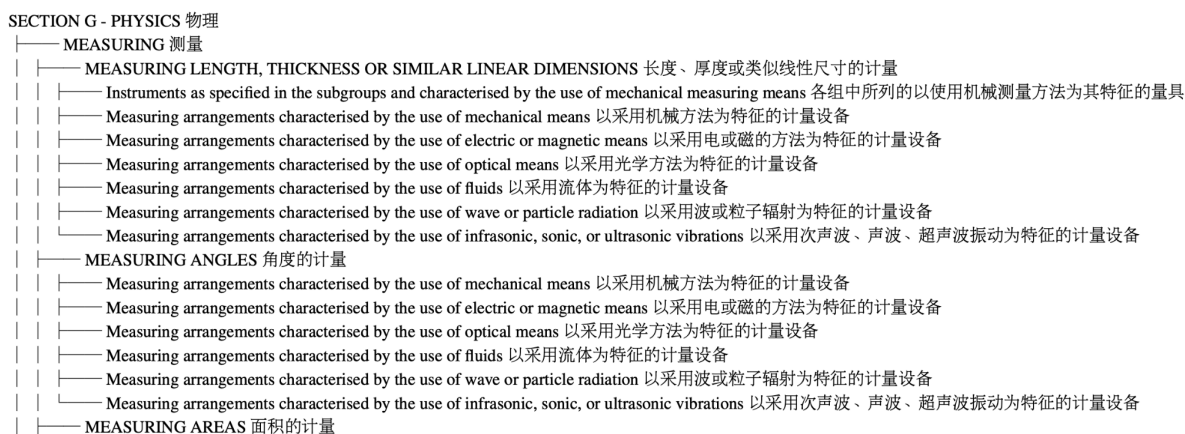


Figure 5.3 - The Beginning of the Constructed Taxonomy of Section G at Level 0 (IPC6)

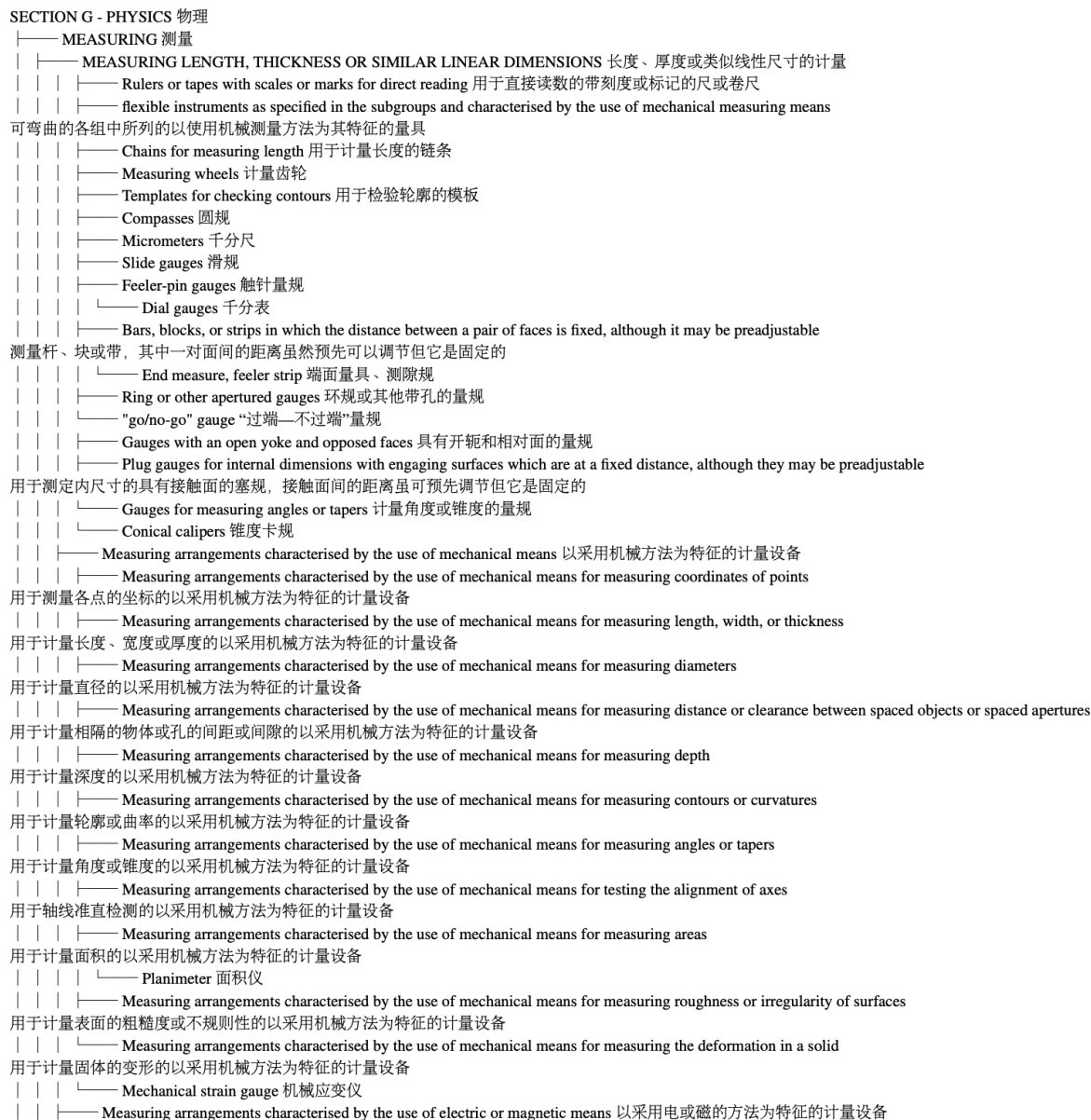


Figure 5.4 - The Beginning of the Constructed Taxonomy of Section G at Level 1 (IPC8)

In the taxonomy trees of both IPC6 and IPC8, it is obvious that there exist two types of the technical expressions extracted: the first one are the typical technical terms (e.g. “Conical calipers 锥度卡规” in Figure 5.4), which are normally noun phrases that are limited in length; the other are the longer expressions (e.g. “Bars, blocks, or strips in which the distance between a pair of faces is fixed, although it may be preadjustable 测量杆、块或带，其中一对面间的距离虽然预先可以调节但它是固定的” and “Measuring arrangements characterised by the use of mechanical means 以采用机械方法为特征的计量设备” in Figure 5.4), which are more like descriptions than technical terms.

It is important to note that not only simple terms but also expressions containing descriptions and conjunctions can be found in the real patent corpus. Although these long expressions may be considered useful for certain types of technical taxonomy, the lexical variation is more interesting in the short nominal technical terms.



Among these two types, the former corresponds to what we define as terms that are of interest for lexical variation. However, the latter contains descriptive contents that are considered as subjects of pruning in Section 5.2.2.

### 5.2.2 Evaluation and Pruning

As there exist no benchmarks for specialized taxonomies (Szopinski, 2020; Kaplan, 2022), we have devised our own evaluation criteria to assess the quality and performance of our technical taxonomy.

Our evaluation consists of two key components: precision and recall.

To assess the precision of our technical taxonomy, we conducted a manual evaluation on a randomly chosen sample of term pairs. This evaluation was centred on gauging the accuracy and relevance of the technical terms and the relationships established between them. Through this manual review process, we aimed to measure the precision of our taxonomy by considering both the quality of the terms and the correctness of the relationships assigned to them.

In Table 5.11, we present the outcomes of this evaluation, which encompass assessments of both term quality and the relationships within the taxonomy. After selecting 100 pairs at random from various sections of the taxonomy, we found that both the English and Chinese technical terms exhibited a high precision rate, approximately 90%, with respect to their quality. This suggests that the majority of the terms are indeed accurate and pertinent within their respective domains.

<b>Data</b>	<b>TermEN</b>	<b>TermZH</b>	<b>Relation</b>	<b>Recall EN</b>	<b>Recall ZH</b>
IPC6 (marked as “level 0”)	94.0%	93.0%	56.0%	19.70%	20.52%
IPC8 (marked as “level 1”)	93.0%	86.5%	49.0%	16.05%	14.03%

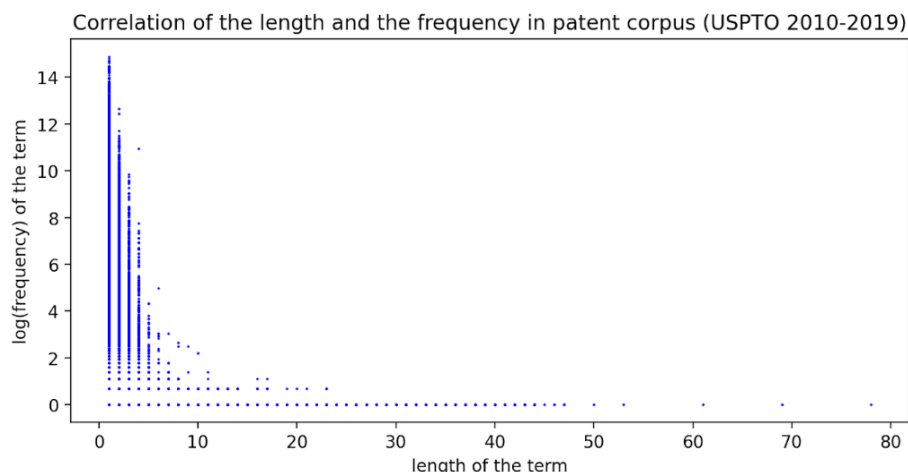
**Table 5.11 - The Evaluation Results on IPC6 and IPC8**

In addition to assessing precision, we also conducted an evaluation of the recall of our taxonomy. Recall measures the capacity of our taxonomy to successfully retrieve pertinent technical terms from actual patent texts. By executing queries on genuine patent documents, we aimed to determine how effectively our taxonomy could identify technical terms in specific domains.

To assess the recall of our technical taxonomy, we performed a retrieval analysis using a dataset of authentic patent applications sourced from the China National Intellectual Property Administration (CNIPA) between 2017 and 2021 and the United States Patent and Trademark Office (USPTO) stored in Solr. This dataset specifically comprised the claims section of patent applications filed between 2010 and 2019.

Unlike the precision, the recall scores<sup>135</sup> for both English and Chinese versions do not appear ideal, with only around 15 to 20 percent of terms being identified in actual patents.

Moreover, our analysis unveiled a correlation between the frequency of occurrence and the length of technical terms. Generally, longer words tend to have lower frequencies. This observation underscores that the quality and effectiveness of our tool diminish as term length increases.



**Figure 5.5 - The Correlation of the Frequency and the Length of English Terms in Constructed Taxonomies**

In the previous step, our process involved the construction of taxonomies by splitting and reorganizing nodes within the IPC titles. However, during this phase, we identified that not all the content contained within child nodes was indicative of hyponymous (is-a) relationships with their respective parent nodes. Following the establishment of the hierarchical tree structure, our subsequent task was to design a relation classifier capable of distinguishing between hypernym/hyponym relationships and other types of relations. These other relations often involve phrases where the heads have different parts of speech (POS) or occur between a concrete term and an abstract term. For instance, these abstract terms could include concepts, activities, events, and similar constructs. Remarkably, we found that more than two-thirds of non-hypernym/hyponym relations fall into these categories, highlighting the importance of a precise relation classification system.

The method involves using a rule-based keyword filter, as outlined in Table 5.12, which specifies the keywords for identifying descriptive titles. The filtering was carried out exclusively using English keywords due to their relatively uniform writing style.

<sup>135</sup> This is defined as the number of found terms divided by the total number of diverse terms.

<b>Keywords (in regular expression)</b>
descriptive_patterns = r'^methods? methods or methods for  methods details means for or methods? or implements? ^machines for machines or instruments for implements for equipments? for specially adapted characterised by ^types? of ^special ^measurement? of therefor thereof therewith thereby  designed for ^treatment  aspects of particular use of general design  them   their  other ^preparation ^Processes for'

**Table 5.12 - List of Keywords to Filter the Descriptive Expressions in the Constructed Taxonomies**

With the filtered taxonomies, we redo the evaluation. The results are shown in Table 5.13 below.

<b>Data</b>	<b>TermEN</b>	<b>TermZH</b>	<b>Relation</b>	<b>Recall EN</b>	<b>Recall ZH</b>
IPC6-light (marked as “level 0”)	98.0%	96.0%	66.0%	54.09%	54.95%
IPC8-light (marked as “level 1”)	90.0%	88.0%	62.0%	51.05%	45.08%

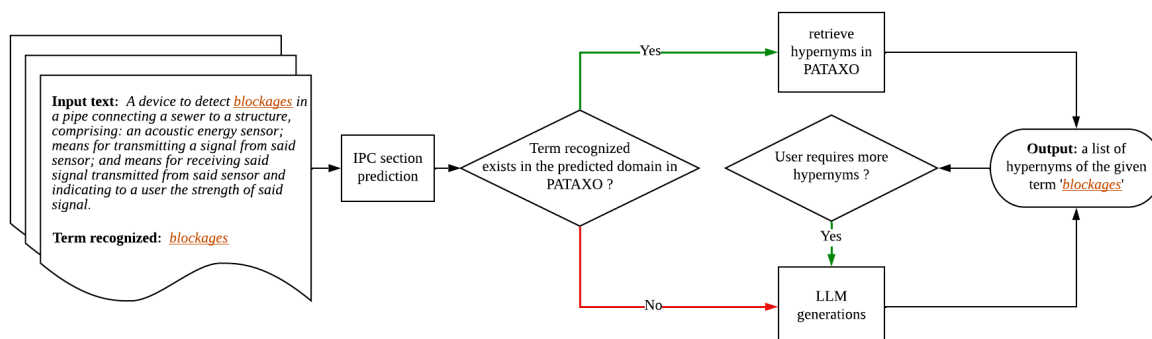
**Table 5.13 - The Evaluation Results on the Light Versions of IPC6 and IPC8**

We can see that the precision and recall have both enhanced following the filtering process, particularly the recall across all four versions.

Our study revealed significant findings, showing that about half of the nodes in both our English and Chinese taxonomies were successfully extracted from at least one patent document. This highlights the efficiency of our taxonomy in encompassing a substantial amount of technical terms present in patent documents.

### 5.3 Oriented Dynamic Lexical Variation

In this section, we will integrate the term recognizer with the IPC-based taxonomy to create another application that aids in patent drafting—specifically, oriented dynamic lexical variation.



**Figure 5.6 - The Procedure of the Oriented Dynamic Lexical Variation**

The complete process of oriented dynamic lexical variation is depicted in Figure 5.6 above. When provided with an input text from a patent claim, we initially identify replaceable terms, such as “blockages” in the example of Figure 5.6. Subsequently, we search for their hypernyms in the IPC-based taxonomy relevant to the corresponding IPC domain. In cases where the hypernym is not found within the constructed taxonomy, we employ a transformer-based Language Model (LLM) trained on the taxonomy to automatically generate a list of potential hypernyms for the recognized term. In the given example of Figure 5.6, the system does not find the term “blockages” in the taxonomy and proceeds to calling the LLM.

In the following subsections, we will outline the training of the transformer-based LLM model as a hypernym generator in Section 5.3.1 and provide examples of term substitution in Section 5.3.2.

#### 5.3.1 Training a Transformer-based Hypernym Generator

Based on the bilingual taxonomies, we trained parallelly an English model and a Chinese model with the task of hypernym generation for each IPC domain. For English, we fine-tuned the FLAN-T5 model (Chung et al., 2022), which is a variant of the original T5 model. And for Chinese, we use the zhuyi-T5-pegasus. In this study, we focus on the Chinese term variation and present in detail only the training process of the Chinese model.

For Chinese data, we used the Chinese model zhuyi-T5-pegasus<sup>136</sup> for fine-tuning as T5’s tokenizer does not support Chinese inference. Compared to the original T5, zhuyi-T5-pegasus uses a Bert tokenizer with Chinese word splitting and has a pre-training task of text summarization and paraphrasing, borrowing ideas from the Pegasus model (Zhang et al., 2019), which is developed by Google AI, designed specifically for abstractive text summarization by pre-training on a large corpus of text with a novel self-supervised objective called “gap sentences generation”.

<sup>136</sup> <https://github.com/ZhuyiTechnology/t5-pegasus>

We used their “t5-pegasus-base”<sup>137</sup> checkpoint<sup>138</sup> with 275 million parameters, and applied the same configurations during fine-tuning as for FLAN-T5 for English.

To simplify the training process of our model, we divided our hierarchical taxonomies into word pairs term-hypernym, allowing us to utilize domain-specific knowledge and the terms as input features. These pairs were then divided into training, validation, and testing subsets in an 80:10:10 ratio. Basic statistics of the datasets for both Chinese and English are presented in Table 5.14 for reference.

	train	valid	test
en	16,316	2,040	2,040
zh	16,316	2,040	2,040

**Table 5.14 - The Size of Training, Validation, and Test Datasets**

We assess the models’ performance using three distinct metrics: (1) Hits at k (Hits@k, where k = 1, 5, 10) and (2) Mean Reciprocal Rank (MRR). These metrics are widely employed in information retrieval as well as for relation prediction tasks. Hits@K measures the proportion of test examples where the correct candidate is ranked within the top K positions. On the other hand, Mean Reciprocal Rank (MRR) provides an absolute ranking score. Below is the formula for calculating the MRR score, where “rank *i*” signifies the rank position of the first relevant document for the *i* th query:

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i}$$

Here, Q represents the total number of queries in the evaluation.

The results for both Chinese and English models are shown in Table 5.15 below.

data	Model	Hits@1	Hits@5	Hits@10	MRR
1) PATA XO_en	FLAN-T5	7.03	17.74	36.04	14.83
2) PATA XO_zh	zhuiyi-T5-pegasus	9.56	22.67	40.58	18.60

**Table 5.15 - The Evaluation Results of the T5-based Hypernym Generator on the Test Set**

In summary, the findings suggest that fine-tuning language models with domain-specific term-hypernym pairs can enhance their ability to generate hypernyms. Nevertheless, there remains room for improvement. Future research endeavours may focus on enhancing model performance through means like integrating supplementary data sources or refining the training process.

Meanwhile, the difficulty of applying the evaluation on the real data is that the generation can be regarded as open-ended questions with no standard answers that can be evaluated automatically when

<sup>137</sup> <https://github.com/renmada/t5-pegasus-pytorch>

<sup>138</sup> In the context of machine learning, a checkpoint refers to a saved state of the model at a specific point in its training process.

it is used for augmented inviting, where the targeted audience does not have exact expectations of the results. In order to better evaluate the usefulness of such a method, we leave it to future work to do a more comprehensive manual extrinsic evaluation by domain experts with real use cases of the hypernym generator in the patent draft.

### 5.3.2 Substitution of Terms

To perform the lexical variation with hypernyms of the recognized terms, we built a test corpus by randomly selecting 1,000 sentences of patent claims in each IPC domain. And following the process in Figure 5.6, we first pass them to the term recognizer and obtain 8 lists of extracted substitutable terms in the 8 IPC domains. For each IPC domain, the total number<sup>139</sup> and ten examples of extracted terms are shown in Table 5.15 below.

IPC domain	Number	Examples
A	4,418	微处理器 (microprocessor), 上隔板 (upper partition), 通水孔 (water passage hole), 轴环 (shaft ring), 液位检测件 (liquid level detection component), 蜂蜜加工工艺 (honey processing technology), 苦白果粉 (bitter white fruit powder), 截留液 (retained liquid), 整体运动姿态 (overall motion attitude), 割草机 (lawn mower)
B	4,807	钻直角槽孔 (drilling right-angled groove holes), 气动组件 (pneumatic components), 钻头安装孔 (drill bit mounting holes), 交流电压 (alternating current voltage), 吸灰管 (vacuum cleaner hose), 病理标本存储转移器 (pathology specimen storage transfer), 陶瓷素胚 (ceramic blank), 电动汽车 (electric car), 多层瓦楞纸板 (multi-layer corrugated cardboard), 车载网络 (onboard network)
C	3,781	热镀锡 (hot-dip tinning), 浓盐水蒸发工艺 (concentrated brine evaporation process), 氧化硅纳米颗粒 (silicon oxide nanoparticles), 温度 (temperature), 烧瓶 (flask), 多室槽 (multi-chamber trough), 生物菌液 (biological bacterial liquid), 煤炭粒料 (coal granules), 碳酸氢钾 (potassium bicarbonate), 阻燃材料 (flame-retardant material)
D	1,217	富氢湿巾机 (hydrogen-rich wet wipes machine), 标准色 (standard color), 墨水 (ink), 打印前图片 (pre-printed image), 皮芯涤棉 (leather core polyester-cotton), 润滑剂 (lubricant), 冰水浴条件 (ice water bath conditions), 真丝织物 (silk fabric), 抗静电整理剂 (anti-static finishing agent), 海藻酸钠溶液 (sodium alginate solution)
E	2,590	装载机 (loader), 机车智能车窗 (locomotive smart car windows), 冲液压支架 (hydraulic flushing bracket), 钢绞线 (steel wire rope), 玻璃胶 (glass adhesive), 门窗洞口 (door and window openings), 木楔 (wooden wedge), 位置线 (position line), 水库调度方式 (reservoir scheduling method), 流域单元水量 (basin unit water volume)
F	5,174	轴承瓦块 (bearing tile), 多功能路灯 (multifunctional street lamp), 地下管线 (underground pipeline), 水泥浆 (cement slurry), 冷凝部 (condenser unit), 双色注胶方法 (two-color injection method), 排料挡板 (discharge baffle), 固态发酵饲料 (solid-state fermented feed), 主动滚轮右侧 (active

<sup>139</sup> The terms can be repetitive in the list.

		right side roller), 沉淀器 (settler)
G	5,475	集团电力企业 (group electric power enterprise), 生产数据 (production data), 供热量数据 (heat supply data), 碳排放基准值 (carbon emissions baseline), 第一预设模型 (first preset model), 热态回转窑 (thermal state rotary kiln), 心线直线度 (coreline straightness), CCD相机 (CCD camera), 砂轮表 (grinding table), 电陶瓷微位平台 (electric ceramic micro-positioning platform)
H	5,050	导电银层 (conductive silver layer), 显影剂 (developer), 智能照明终端 (smart lighting terminal), 显示屏 (display screen), 凹凸卡 (embossed card), 电感线圈 (inductor coil), 公钥验证 (public key verification), 储能模块 (energy storage module), 第一PMOS管 (first PMOS transistor), 网络视频图像压缩方法 network video image compression method)

**Table 5.15 - Total Number and Examples of Extracted Terms from Sampled Patent Claims of Each IPC Domain <sup>140</sup>**

Furthermore, to assess the effectiveness of the substitutions, we gathered five examples featuring indexed terms from Section G - Physics. Again, we look primarily at Section G because it is the easiest to understand for us, as, for example, computer science is a subclass of Section G. For each recognized term, we provide in Table 5.17 the system's top five hypernym suggestions.

The outcomes are presented in the table below as Table 5.17. In this table, the indexed terms are visually distinguished through the use of underlined bold formatting, while the recognized terms are highlighted in yellow (for indexed correct terms), red (for indexed incorrect terms, which have been partly recognized), purple (for unrecognized indexed terms, which have not been recognized at all) or green (for non-indexed terms but have been recognized) shade. It's worth noting that since terms may occur repeatedly within a single patent claim, we have opted to provide hypernyms exclusively for the initial occurrence of each term. This decision is based on the understanding that the model consistently suggests the same hypernym for a given term.

1. 一种硬盘驱动器 (Hard disk drive) [指令发信装置; 核算装置; 电数字数据处理; 盘、板、台; 电学 (instruction sending device; accounting device; electro-digital data processing; disks, boards, tables; electrical)] (1)的挠性件 (Flexible component) [机械工程; 测量磁变量; 润滑组合物; 测试; 作业 (mechanical engineering; measuring magnetic variables; lubricating compositions; testing; operations)] (15), 其特征在于, 包括:  
金属基底 (Metal base) [金属的轧制; 润滑组合物; 金属冲压; 核算装置; 电铸 (metal rolling; lubricating compositions; metal stamping; accounting devices; electrocasting)] (18), 所述金属基底由金属板制成。  
布线构件 (Wiring components) [电数字数据处理; 核算装置; 电学; 天线; 物理 (electronic digital data processing; accounting devices; electrical engineering; antennas; physics)] (20), 所述布线构件(20)形成于所述金属基底(18)上;  
 所述布线构件(20)配备有形成于所述金属基底(18)上的绝缘层 (Insulation layer) [导电连接; 导体; 涂层或层; 金属的轧制; 电铸 (conductive connections; conductors; coatings or layers; metal rolling; electroplating)](30)和形成于所述绝缘层(30)上的导体层 (Conductor layer) [导体; 电学; 导电连接; 电铸; 半导体 (conductor; electrical; conductive connections; electroplating; semiconductor)](40);  
 所述绝缘层(30)具有与所述金属基底(18)平行的平坦部 (Flat portion) [核算装置; 测时学; 测量磁变量; 平整或平整; 固定建筑物 (accounting device; horology; measuring magnetic variables; smooth or level; fixed structures)] (23)和从所述平坦部(23)凸起的凸起部 (Raised portion) [核算装置; 润滑组合物; 结构构件; 金

<sup>140</sup> The English translation has been done on <https://www.deepl.com/>.

属冲压; 金属的轧制 (accounting device; lubricating composition; structural component; metal stamping; rolling of metals)] (24);

所述凸起部(24)具有与所述金属基底(18)平行的上表面(Upper surface) [核算装置; 压力机; 电记录术; 表面、结构; 表面或纹理 (accounting device; press machine; electrical recording technique; surface, structure; surface or texture)] (24A)和连接所述上表面(24A)和所述平坦部(23)的侧表面(Side surface) [侧边缘的调节; 侧表面、清洁; 侧边缘; 侧边缘或表面; 侧边缘或支柱 (side edge adjustment; side surface, cleaning; side edge; side edge or surface; side edge or pillar)] (24B);

所述导体层(40)具有沿所述侧表面(24B)形成的连接端子(Terminal) [电数字数据处理; 测量磁变量; 导电连接; 导体; 信号装置 (electronic digital data processing; measurement of magnetic variables; electrical connection; conductor; signal device)] (22)。

所述平坦部(23)中未堆叠有所述第1绝缘覆盖层(Cover layer) [覆盖或衬里; 核算装置; 压力机; 润滑组合物; 覆盖或涂层 (covering or lining; accounting device; press; lubricating composition; covering or coating)] (32)。

所述凸起部(24)中堆叠有所述第1覆盖绝缘层(Insulating layer) [导电连接; 导体; 涂层或层; 金属的轧制; 电铸 (conductive connection; conductor; coating or layer; metal rolling; electrocasting)] (32)。

3. 根据权利要求1所述的挠性性(15), 其特征在于, 包括作为所述凸起部(24)的第1凸起部(241)、第2凸起部(242)及第3凸起部(243);

所述布线构件(20)的延伸方向(Direction) [方向的控制; 测时学; 方位或方位; 方向、逆时针; 方位或方向 (control of direction; horology; orientation or bearing; direction, counterclockwise; orientation or direction)] (X)包括第1方向(X1)和与第1方向(X1)相反侧的第2方向(X2);

所述第1凸起部(241)和所述第3凸起部(243)沿与所述延长方向(Extension direction) [测时学; 电记录术; 电数字数据处理; 测试; 测量磁变量 (horology; electrical recording technique; electrical digital data processing; testing; measurement of magnetic variables)] (X)垂直(Vertical) [测时学; 运动; 作业; 测量磁变量; 电记录术 (horology; motion; operation; measurement of magnetic variables; electrical recording technique)] 的宽度(Width) [宽度、持续时间、宽度; 宽度、持续时间; 宽度、速度、加速度; 宽度、持续时间或宽度; 宽度、伸缩性 (width, duration, width; width, duration; width, speed, acceleration; width, duration or width; width, elasticity)] 方向并列;

所述第1凸起部(241)和所述第3凸起部(243)的所述连接端子(22)分别形成于所述第2方向(X2)侧的所述第1凸起部(241)和所述第3凸起部(243)的所述侧表面(24B)上;

所述第2凸起部(242)的所述连接端子(22)形成于所述第1方向(X1)侧的所述第2凸起部(242)的所述侧表面(24B)。

1. 一种用于汽车(Car) [核算装置; 模型铁路; 作业; 铁路; 机车 (accounting device; model railway; operation; railway; locomotive)] 打滑测试的轮滑锁(Slide lock) [锁; 电数字数据处理; 电学; 标记、锁; 调节 (lock; electronic data processing; electrical; marking, lock; adjustment)] 止装置, 其特征在于, 所述一种用于汽车打滑测试的轮滑锁止装置包括:

4. 根据权利要求1所述的一种用于汽车打滑测试的轮滑锁止装置, 其特征在于:

所述的框架(12)两侧设置有凹型件(Concave component) [压力机; 模印机; 核算装置; 光学部件; 电铸 (press machine; stamping machine; accounting device; optical component; electrocasting)] (4)和连接件(Connector) [电数字数据处理; 锁; 继电器; 钥匙; 电阻器 (electronic data processing; lock; relay; key; resistor)] (5)用于连接其他模块。

2. 根据权利要求1所述的一种用于汽车打滑测试的轮滑锁止装置, 其特征在于:

所述的滑轮(Pulley) [电机; 齿轮机构; 润滑; 机械工程; 运动 (electric motor; gear mechanism; lubrication; mechanical engineering; motion)] (1)的一端设置有防护轮(Protective wheel) [机械工程; 润滑组合物; 电机; 一般车辆; 齿轮机构 (mechanical engineering; lubrication combination; electric motor; general vehicle; gear mechanism)] (3)。

1. 一种加油口(Fuel inlet) [润滑组合物; 核算装置; 润滑; 机械工程; 齿轮机构 (lubrication combination; accounting device; lubrication; mechanical engineering; gear mechanism)] 盖开启角度(Lid opening angle) [测时学; 加热或冷却; 测量磁变量; 一般热交换; 开启角度或持续时间 (horology; heating or cooling; measurement of magnetic variables; general heat exchange; opening angle or duration)] 的检测装置, 包括底板(Bottom plate) [核算装置; 电学; 电记录术; 测时学; 导电连接 (accounting device; electrical; electric recording; horology; electrical connection)] (1), 其特征在于:

所述角度(Angle) [角度、速度、加速度; 角度或角度; 角度、加速度或加速度; 角度、速度或加速度; 角度、角度 (angle, speed, acceleration; angle or angle; angle, acceleration, or acceleration; angle, speed, or acceleration; angle, angle)] 检测机构包括第一支撑块(First support block) [金属冲压; 机械工程; 作业; 基础; 电学 (metal stamping; mechanical engineering; operation; foundation; electrical engineering)] (8), 第一支撑块(8)固定设置于底板(1)上, 第一支撑块(8)上部固定连接有角度检测块(Angle detection block) [角度测量或校准块; 角度扫描或测量块; 角度测量或监测块; 角度测量或测量块; 角度或方位检测块



(angle measurement or calibration block; angle scanning or measurement block; angle measurement or monitoring block; angle measurement or measurement block; angle or orientation detection block)] (7), 角度检测块 (7) 上设置有检测面。

6. 根据权利要求6所述一种加油口盖开启角度的检测装置, 其特征在于:

所述底板 (1) 下表面上固定设置有支撑座 (Support base) [底座或支柱; 牵引; 稳定面; 支座或支柱; 机械工程 (base or pillar; traction; stable surface; support base or pillar; mechanical engineering)] (9), 底板 (1) 上表面上固定设置有基准块 (Reference block) [核算装置; 测量磁变量; 测时学; 测试; 模型或型芯 (accounting device; measurement of magnetic variables; timing studies; testing; model or core)] (10)。

1. 一种基于FPGA的SOC嵌入式质量视觉检测设备, 其特征在于:

该设备包括SOC板卡 (SOC board) [电数字数据处理; 电学; 核算装置; 磁体; 声学 (electric digital data processing; electrical; accounting device; magnet; acoustics)] (1), PC (2)、电源 (Power supply) [电学; 电数字数据处理; 测量磁变量; 物理; 核算装置 (electricity; electrical digital data processing; measurement of magnetic variables; physics; accounting device)]<sup>141</sup> 系统 (3) 和相机 (Camera) [电记录术; 核算装置; 印鉴; 光学部件; 摄影机 (recording; accounting device; seal; optical component; camera)] (5);

所述电源系统 (3) 分别与SOC板卡 (1) 和PC (2) 连接, 用于给SOC板卡 (1) 和PC (2) 供电;

所述相机 (5) 与SOC板卡 (1) 连接, 用于拍摄视频图像;

所述SOC板卡 (1) 还与PC (2) 连接, 用于采集和处理相机 (5) 拍摄的视频图像, 并将处理后的视频图像信息 (Video image information) [电数字数据处理; 电记录术; 图像或视频图像; 数字记录术; 图像、语音识别 (electric digital data processing; electric recording technique; image or video image; digital recording technique; image, voice recognition)] 发送给PC (2);

所述PC (2) 与本地网 (Local network) [测时学; 天象仪; 电数字数据处理; 测量磁变量; 测试 (metrology; celestial globe; electrical digital data processing; measurement of magnetic variables; testing)] 连接, 用于接收并处理所述视频图像信息, 还用于显示视频图像信息和控制SOC板卡 (1)。

2. 根据权利要求1所述的基于FPGA的SOC嵌入 (Embedding) [电数字数据处理; 核算装置; 调制载波系统; 脉冲技术; 数字数据处理 (electrical digital data processing; calculation device; modulated carrier systems; pulse technique; digital data processing)] 式质量 (Mass) [测试; 测量磁变量; 测时学; 电数字数据处理; 测量电变量 (testing; measurement of magnetic variables; chronometry; electrical digital data processing; measurement of electric variables)] 视觉检测设备, 其特征在于:

所述电源转换电路 (Power conversion circuit) [电数字数据处理; 电学; 测量磁变量; 调节; 物理 (electrical digital data processing; electrical engineering; measurement of magnetic variables; regulation; physics)] (141) 与电源端子 (Power terminal) [电学; 导电连接; 继电器; 导体; 电开关 (electrical engineering; electrical connection; relay; conductor; electrical switch)] (140) 连接, 所述电源端子 (140) 与电源系统 (3) 连接, 所述电源转换电路 (141) 用于给SOC板卡 (1) 供电;

所述SOC芯片 (SOC chip) [电学; 电数字数据处理; 核算装置; 磁体; 电通信技术 (electrical engineering; digital data processing; accounting device; magnetism; telecommunication technology)] (110) 分别与DDR3 内存 (Memory) [核算装置; 电数字数据处理; 物理; 磁盘、卡片; 推算 (accounting device; digital data processing; physics; disk, card; estimation)] (111)、CAMERALINK 电路 (CAMERALINK circuit) [电数字数据处理; 光学部件; x光机; 光学; 电记录术 (digital data processing; optical parts; X-ray machine; optics; electrical recording)] (120)、千兆网络 (Network) [核算装置; 物理; 网络数据管理; 电学; 测试 (accounting device; physics; network data management; electrical; testing)] 电路 (Circuit) [电数字数据处理; 电学; 测量磁变量; 核算装置; 物理 (electrical and electronic data processing; electrical; measuring magnetic variables; accounting device; physics)] (130)、JTAG 配置插件 (Configuration plugin) [电数字数据处理; 核算装置; 物理; 调节; 电通信技术 (electrical and electronic data processing; accounting device; physics; regulation; telecommunication technology)] (112) 和RS485 芯片 (Chip) [电数字数据处理; 电学; 核算装置; 测时学; 测量磁变量 (electrical and electronic data processing; electrical engineering; accounting device; time measurement; magnetic field measurement)] (116) 连接;

所述CAMERALINK 电路 (120) 用于与相机 (5) 连接;

所述RS485 芯片 (116) 还与IO 端子 (IO terminal) [电数字数据处理; 导电连接; 继电器; 电学; 测量磁变量 (electrical and electronic data processing; conductive connection; relay; electrical engineering; magnetic field measurement)] (150) 连接;

所述千兆网络电路 (130) 用于与PC (2) 连接。

4. 根据权利要求2所述的基于FPGA的SOC嵌入式质量视觉检测设备, 其特征在于:

所述SOC板卡 (1) 还包括与SOC芯片 (110) 连接的调试IO口 (IO port) [核算装置; 声学; 继电器; 电学; 语音识别 (accounting device; acoustics; relay; electrical engineering; speech recognition)] (114)。

<sup>141</sup> The term 电源 (power supply) is treated as incomplete due to its original form that should be combined with the term 系统 (system) behind it.

1. 一种像素结构 (Pixel structure)(10), 其特征在于, 包括:  
 两条平行于第一方向的数据线 (Data cable) [电数字数据处理; 核算装置; 测量磁变量; 测时学; 电学 (digital data processing; accounting device; measurement of magnetic variables; chronometry; electrical engineering)] (100)和两条平行于第二方向 (Second direction) [方向的控制; 顺时针或逆时针; 测时学; 电数字数据处理; 测试 (direction control; clockwise or counterclockwise; chronometry; digital data processing; testing)] 的两条栅极线 (Stripe gate electrode) [条带、栅极线; 电记录术; 电数字数据处理; 细丝状材料的导向器; 条带、带子、弹簧 (strip, gate line; electrical recording; digital data processing; guide for filamentary material; strip, belt, spring)] (101), 所述第一方向与所述第二方向垂直, 两条所述数据线(100)与两条所述栅极 (Gate electrode) [电学; 电记录术; 测量磁变量; 核算装置; 光学部件 (electricity; electrical recording; measurement of magnetic variables; accounting device; optical component)] 线(101)围成像素区域 (Pixel region) [计数; 电数字数据处理; 电记录术; 比色法; 放大的控制 (counting; electronic digital data processing; electric recording; colorimetry; control of amplification)] (102);  
 两条所述数据线(100)之间设有沿所述第一方向延伸的第一公共电极 (Electrode) [电学; 电记录术; 测量磁变量; 电铸; 电阻器 (electrical engineering; electric recording; measurement of magnetic variables; electroplating; resistor)] (103), 所述第一公共电极(103)与所述数据线(100)同层 (Same layer) [电学; 核算装置; 电数字数据处理; 电记录术; 导电连接 (electrical engineering; accounting device; electrical digital data processing; electric recording; electrical connection)] 设置, 用于与外围公共电极 (Encircle common electrode) [电学; 电阻器; 导电连接; 测量磁变量; 导体 (electrical engineering; resistor; electrical connection; magnetic variable measurement; conductor)] 连接;  
 两条所述栅极线(101)之间设有沿所述第二方向延伸的第二公共电极(104), 所述第二公共电极(104)与所述栅极线(101)同层设置, 用于与所述外围公共电极连接;  
 所述第一公共电极(103)从靠近所述栅极线(101)的位置沿第二方向分别延伸至所述第一公共电极(103)两侧的所述薄膜 (Thin film) [电记录术; 核算装置; 分离; 物理; 薄膜、玻璃 (electrical recording; accounting device; separation; physics; thin film, glass)] 晶体 (Crystal) [电学; 电记录术; 核算装置; 电铸; 物理 (electrical; recording; accounting device; electrocasting; physics)] 管(106)处。

**Table 5.17 - Five Example Claim Texts from Section-G Demonstrating the Results of Lexical Variation<sup>142</sup>**

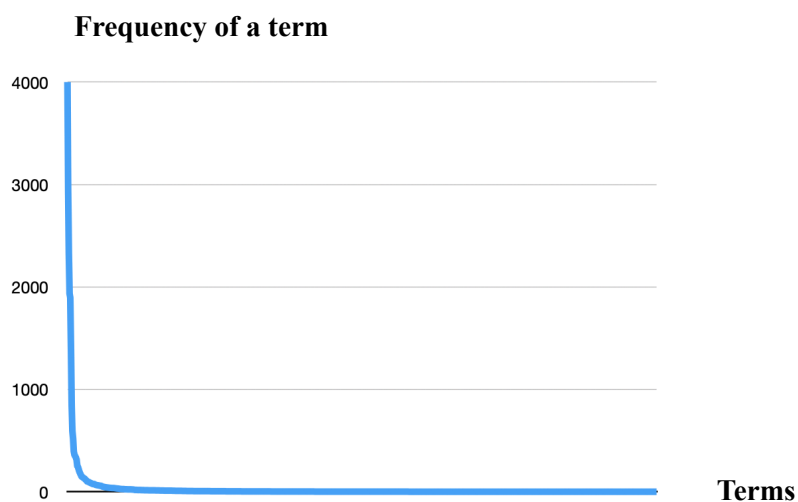
The results presented in Table 5.17 reveal several issues related to the term substitution process. One of the key challenges is that the terms available in the taxonomies often prove to be overly general (e.g. for the term 数据线 (Data cable) the proposed top five hypernyms are 电数字数据处理 (digital data processing), 核算装置 (accounting device), 测量磁变量 (measurement of magnetic variables), 测时学 (chronometry) and 电学 (electrical engineering) and lack the specificity required for accurate substitutions. Additionally, the taxonomies tend to focus primarily on titles at the top levels that are more frequent in the training set, which can limit the availability of suitable hypernyms for more specialized terms. Figure 5.7 demonstrates the distribution of frequency of the predicted hypernyms: among all 840 identical predicted terms, while the most frequently predicted term “核算装置 (hé suàn zhuāng zhì, ‘Accounting device’), which is a very frequent term in the training set, has appeared 4,000 times in the prediction file, more than 600 terms appear less than 10 times.

Table 5.18 gives the top 20 most frequent predicted terms in the results with their level code<sup>143</sup> in IPC titles. It proves that the most frequent predictions are from the top-level titles (not only from Section G, but also from top-level titles of other sections, such as “电铸 (Electroforming)” is from Section C).

This issue becomes particularly evident when dealing with terms like “第1凸起部 (dì 1 tū qǐ bù, ‘Bulge 1’)” and “第一支撑块 (dì yī zhī chēng kuài, ‘First support block’)”, where finding appropriate hypernyms becomes a challenge due to their unique nature.

<sup>142</sup> The English translations have been produced using <https://www.deepl.com/>.

<sup>143</sup> As the level code starts from 0 from the fourth level, the first three levels do not have a level code. We call them here from the top “level -3”, “level -2” and “level -1”.



**Figure 5.7 - The Frequency of Predicted Hypernyms from Section G**

Predicted Hypernyms	IPC Level	Frequency
核算装置: Accounting device	-2	4,000
电数字数据处理: Digital data processing	-2	2,934
测量磁变量: Magnetic variable measurement	-2	2,314
测时学: Timing studies	-2	1,934
物理: Physics	-3	1,896
测试: Testing	-2	1,399
电记录术: Electric recording	-2	839
调节: Regulaing	-2	594
印鉴: Seal	-2	521
声学: Acoustics	-2	394
电铸: Electroforming	_(C -2)	361
测量电变量: Electrical variable measurement	-2	358
计数: Counting	-2	334
数据识别: Data recognition	-2	313
香料磨: Spice grinding	_(A -2)	253
作业: Performing operations	_(B -3)	242
润滑组合物: Lubricating composition	_(C -1)	237
信号装置: Signal device	-2	196
模型或型芯: Model or core	_(B 0)	194
指令发信装置: Instruction sending device	-1	166

**Table 5.18 - The Top 20 Most Frequent Predicted Hypernyms for Terms Extracted from Sentences in Section G**

A potential solution to overcome the limitations of the taxonomies is to incorporate information from deeper levels of the IPC titles, thereby enriching the taxonomies with more specific technical terms. Currently, the mining process only utilizes less than half of the available information in the IPC titles by stopping at “level 1” due to the memory constraints.

Additionally, to address the issue of frequency imbalance in predictions, the problem mainly arises from the abundance of pairs involving top-level IPC terms in the training dataset. To mitigate this, one approach is to restrict the frequency of highly frequent top-level IPC terms by limiting the hypernyms to levels lower than “level 1” during the training phase. The exploration of this hypothesis will be deferred to future research.

Furthermore, these results shed light on potential problems with the term extraction process itself, indicating the need for further refinement and enhancement in this aspect of the workflow. Out of the 56 terms found in all five example claims in Table 5.17, 24 of them are recognized incorrectly (2 were not recognized, 13 were partially recognized, and 9 were recognized incorrectly).

# Conclusion and Outlook

This thesis focused on the analysis of syntax and terminology in patent texts, and its main contributions are twofold:

1. The creation of a character-level treebank, representing the first-ever treebank for Chinese patent claims, with precise annotation guidelines based on widely accepted Chinese morphology theories;
2. The development of a novel method for handling lexical variation was achieved through the construction of a patent-related taxonomy and the term recognizer.

This thesis has made significant contributions to the field of computational linguistics in the domain of the analysis of Chinese patent claims. The creation of a character-level dependency analysis schema marks an advancement, which enables a deeper understanding of the complex structure of technical terms in patent language. This tool not only aids in the breakdown of technical terms but also paves the way for a more nuanced analysis of their syntactic and semantic properties.

We also showed how the annotation at the character level allows to train a parser that learns to distinguish morphological relations from standard syntactic relations and thus can function as a context-aware multi-word term recognizer that shows state-of-the-art performance on patents. This is particularly interesting for these highly technical texts where any vocabulary-list-based approach to term recognition will require a very high technical domain granularity and frequent updates to be able to predict termhood in patents.

On the other side, the development of a technical taxonomy based on the International Patent Classification titles serves as a foundation for generating lexical variants of patent terms, offering a more dynamic approach to understanding and processing patent language. The taxonomy's focus on hypernyms and hyponyms enriches the study of patent texts, offering insights into their hierarchical structure and terminology.

The methodology employed in this research showcases an innovative blend of linguistic theory and technical domain knowledge, effectively addressing the unique challenges posed by Chinese patent texts. This research epitomizes an interdisciplinary convergence, amalgamating computational linguistics with patent analysis to innovatively automate the generation of lexical variants in Chinese patent claims. The methodology, characterized by its computational rigor and linguistic sensitivity, offers a nuanced understanding of patent language idiosyncrasies, thereby augmenting the corpus of knowledge in both computational linguistics and patent documentation analysis. This scholarly endeavor paves the way for future research, potentially catalyzing advancements in the processing and interpretation of specialized technical texts across various linguistic and domain-specific landscapes.

In the conclusion section of this thesis on syntax and terminology analysis in patent texts, we acknowledge that our work has opened new avenues for exploration in computational linguistics and patent analysis. However, several challenges and opportunities for advancement remain.

This study has made significant strides in the field of computational linguistics and patent analysis, yet there remain several areas that warrant further attention and development. Firstly, the limited quantity of sentences in our treebank has led to suboptimal parsing results, indicating a need for the

## Conclusion and Outlook

bootstrapping of the treebank. Additionally the exclusive use of section “G” limits the breadth of our analysis. Future research should also delve deeper into term recognition and lexical variation, incorporating expert manual evaluation to enhance accuracy and relevance.

The further experiments on the new version of the International Patent Classification (IPC) and the potential inclusion of the Chinese Cooperative Patent Classification (CPC) can represent significant advancements in our field. These developments not only reflect the evolving nature of patent classifications but also offer new opportunities for linguistic and analytical exploration.

Looking forward, a systematic comparison of our approach with more classical algorithms such as the application of TF-IDF for candidate term identification, exploring different selection methods such as variations of C-value<sup>144</sup> and Pointwise Mutual Information (PMI), and the implementation of a character-level parser on taxonomy titles are essential next steps. These techniques will enable a more nuanced and precise analysis of patent documents.

Additionally, the fusion of taxonomies with other resources to create a dynamic, more comprehensive databases is a critical avenue for future research. This approach, coupled with a more detailed treatment of specific domains (such as chemical substances in section “C” in IPC) and the handling of conjunctions in taxonomies, will greatly enhance the depth and utility of our data.

Moreover, the exploration of advanced machine learning techniques, such as deep learning and transformer-based models, could significantly improve the performance of our parser and term recognizer. These models, known for their ability to capture complex patterns in large datasets, could offer more nuanced insights into the linguistic intricacies of patent texts.

Another promising direction is the exploration of interlingual analysis, examining how patent terminology and syntax vary across languages. This could involve creating parallel treebanks for patents in different languages, offering insights into cross-linguistic variations and similarities in patent discourse.

Furthermore, the application of our methodologies to other technical documents, such as scientific papers or technical manuals, could validate the generalizability of our findings and techniques. This would not only broaden the scope of our research but also contribute to the wider field of technical document analysis.

In conclusion, while this study has laid a solid foundation and introduced new perspectives, the path ahead is ripe with opportunities for further exploration and refinement in the complex interplay of computational linguistics, patent analysis, and data science.

---

<sup>144</sup> The C-value method, already mentioned in Se, proposed by Frantzi et al. in 2000, is a statistical approach for automatic term recognition in text, which quantifies the importance of a phrase based on its frequency in the text, its length, and its degree of nesting within other terms.

## References

- Baker, M. (1985). The mirror principle and morphosyntactic explanation. *Linguistic inquiry*, 16(3), 373-415.
- Bian, C. (1998). Analysis of the structure of three-syllable compound words in modern Chinese. *Chinese Language Learning*, (4).[卞成林. 现代汉语三音节复合词结构分析[J]. 汉语学习,1998. (4).]
- Bloomfield, L. (1933). *Language*. New York: Holt, Rinehart & Winston.
- Bourigault, D., & Slodzian, M. (1999). Pour une terminologie textuelle. *Terminologies nouvelles*, 19, 29-32.
- Budin, G. (2001). A CRITICAL EVALUATION OF THE STATE-OF-THE ART OF TERMINOLOGY THEORY. *Terminology Science & Research/Terminologie: Science et Recherche*, 12(1-2), 7-23.
- Burga, A., Codina, J., Ferraro, G., Saggion, H., & Wanner, L. (2013, August). The challenge of syntactic dependency parsing adaptation for the patent domain. In *ESSLLI-13 workshop on extrinsic parse improvement*.
- Castellví, M. T. C. (2003). Theories of terminology: Their description, prescription and explanation. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 9(2), 163-199.
- Cao, W. (2004). *The Study of Modern Chinese Lexicon*, Beijing: Peking University Press. [曹炜. 现代汉语词汇. 北京:北京大学出版社. 2004.]
- Chao, Y. (1948). *Mandarin primer: An intensive course in spoken Chinese*. Harvard University Press.
- Chen, C. Y., & Duanmu, S. (2016). Categorization and statistical analysis of morphemes in disyllabic compound words. *Chinese Language Learning*, 1. [陈昌勇 & 端木三. "双音节复合词内部语素的词类标注和统计分析." 汉语学习 1 (2016)]
- Chen, K. J., Luo, C. C., Chang, M. C., Chen, F. Y., Chen, C. J., Huang, C. R., & Gao, Z. M. (2003). Sinica treebank: Design criteria, representational issues and implementation. *Treebanks: Building and using parsed corpora*, 231-248.
- Chen, Y., Huang, Y., & Fang, J. (2006). *Patent Information Collection and Analysis*. Beijing: Tsinghua University Press. [陈燕, 黄迎燕, 方建国. 专利信息采集与分析[M]. 北京: 清华大学出版社. 2006.]
- Cheng, X. (2008). *A Study of Polysyllabic Words of Special Books in Chinese History (Revised Edition)*, Beijing: The Commercial Press. [程湘清. 汉语史专书复音词研究 (增订本). 北京:商务印书馆. 2008.]
- Chi, C & Lin, Z. (2019). Reconsidering the Parallelism of Chinese Compound Words Structure and Syntactic Structure. [池昌海 & 林志永. "汉语复合词的结构与句法结构平行" 说新议." 浙江大学学报 (人文社会科学版) 5. 2019.]

## References

- Chinese Academy of Social Sciences, Institute of Language Studies, Dictionary Editorial Office. (2005). *The Modern Chinese Dictionary (5th ed.)*. Commercial Press.
- Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on information theory*, 2(3), 113-124.
- Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., ... & Wei, J. (2022). Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.
- Dai, X. L. (1992). *Chinese Morphology and its Interface with the Syntax*. The Ohio State University.
- Ding, J., Lü, X., & Liu, K. (2015). A method for extracting terminology from patent literature based on boundary marker set. *Computer Engineering and Science*, 37(8), 1591-1598. [丁杰, 吕学强, 刘克会. 基于边界标记集的专利文献术语抽取方法[J]. 计算机工程与科学, 2015, 37(8):1591-1598.]
- Ding, S. (1961). *Lectures on Modern Chinese Grammar*. Commercial Press, Chinese Language Magazine Society. [丁声树. 现代汉语语法讲话. 商务印书馆, 中國語文杂志社. 1961.]
- Dong, X. (2011). *Lexicalization: the origin and evolution of Chinese disyllabic words*. [董秀芳. 词汇化: 汉语双音词的衍生和发展. 2011.]
- Dong, X. (2016). *The Lexicon and Morphology of Chinese*. Peking University Press. (p. 6). [董秀芳. 汉语的词库与词法. 北京大学出版社 2016 年版, 第 6 页.]
- Dong, Z., & Dong, Q. (2003, October). HowNet-a hybrid language and knowledge resource. In *International conference on natural language processing and knowledge engineering, 2003. Proceedings. 2003* (pp. 820-824). IEEE.
- Dozat, T., & Manning, C. D. (2016). Deep biaffine attention for neural dependency parsing. *arXiv preprint arXiv:1611.01734*.
- Du, J., Alexantris, C., & Yu, P. (2021). Towards Chinese Terminology Application of TERMONLINE. In *Advances in Artificial Intelligence, Software and Systems Engineering: Proceedings of the AHFE 2021 Virtual Conferences on Human Factors in Software and Systems Engineering, Artificial Intelligence and Social Computing, and Energy, July 25-29, 2021, USA* (pp. 190-198). Springer International Publishing.
- Duanmu, S. (1998). Wordhood in Chinese. *New approaches to Chinese word formation: Morphology, phonology and the lexicon in modern and ancient Chinese*: 135-196.
- Federico M. (1993). *The formation of the modern Chinese lexicon and its evolution toward a national language: the period from 1840 to 1898*. [马西尼. 现代汉语词汇的形成: 十九世纪汉语外来词研究. 1993.]
- Feng, Z. (1988). FEL formula - The economy law of term formation. *Information Science*, 5, 8-15. [冯志伟. "FEL 公式—术语形成的经济律." 情报科学 5 (1988): 8-15.]
- Feng, Z. (1989a). Preliminary analysis of the structure of Chinese word-type terms. *China Terminology*, 02, 17. [冯志伟. "汉语单词型术语的结构初析." 中国科技术语 02 (1989): 17.]



- Feng, Z. (1989b). Structural description and potential ambiguities in Chinese scientific and technical terms. *Journal of Chinese Information Processing*, 2, 1-16. [冯志伟. 中文科技术语的结构描述及潜在歧义. 中文信息学报 2 (1989): 1-16.]
- Feng, Z. (1989c). Ambiguous structures in Chinese scientific and technical terms and their identification methods. *Journal of Chinese Information Processing*, 3(3), 12-27. [冯志伟. 中文科技术语中的歧义结构及其判定方法. 中文信息学报 3, no. 3 (1989): 12-27.]
- Feng, Z. (1994). Some non-grammatical factors in determining word segmentation units. *Journal of Chinese Information Processing* [冯志伟. 确定切词单位的某些非语法因素. 中文信息学报. (1994).]
- Feng, Z. (1995). On the potential of ambiguous structures. *Journal of Chinese Information Processing*, 9(4), 14-24. [冯志伟. 论歧义结构的潜在性. 中文信息学报 9, no. 4 (1995): 14-24.]
- Feng, Z. (1996). Methods of ambiguity resolution in natural language processing. *Applied Linguistics*, 1, 55-60. [冯志伟. 自然语言处理中的歧义消解方法. 语言文字应用 1 (1996): 55-60.]
- Feng, Z. (2001). Some grammatical factors in determining word segmentation units. *Terminology Standardization & Information Technology*, 2, 24-30. [冯志伟. 确定切词单位的某些语法因素. 术语标准化与信息技术 2 (2001): 24-30.]
- Feng, Z. (2001). The main schools of modern terminography. *China Terminology*, 3(01), 33. [冯志伟. 现代术语学的主要流派. 中国科技术语 3, no. 01 (2001): 33.]
- Feng, Z. (2004a). The structure of Chinese word-type terms. *China Terminology*, 6(01), 14. [冯志伟. 汉语单词型术语的结构. 中国科技术语 6, no. 01 (2004): 14.]
- Feng, Z. (2004b). The structure of Chinese phrase-type terms. *China Terminology*, 6(02), 35. [冯志伟. 汉语词组型术语的结构. 中国科技术语 6, no. 02 (2004): 35.]
- Feng, Z. (2006). Potential ambiguities in Chinese scientific and technical terms (continued). *China Terminology*, 8(02), 14. [冯志伟. 汉语科技术语中的潜在歧义 (续). 中国科技术语 8, no. 02 (2006): 14.]
- Feng, Z. (2010). *Formal Models of Natural Language Processing*. University of Science and Technology of China Press.
- Frantzi, K., Ananiadou, S., & Mima, H. (2000). Automatic recognition of multi-word terms: the c-value/nc-value method. *International journal on digital libraries*, 3, 115-130.
- Gao, J., Li, M., Huang, C. N., & Wu, A. (2005). Chinese word segmentation and named entity recognition: A pragmatic approach. *Computational Linguistics*, 31(4), 531-574.
- Ge, B. (Ed.). (2003). *Chinese Lexicology*. Shandong University Press. [葛本仪主编. 汉语词汇学. 济南: 山东大学出版社. 2003.]
- Gerdes, K., Guillaume, B., Kahane, S., & Perrier, G. (2018, November). SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In *Universal dependencies workshop 2018*.

## References

- Gerdes, Kim, Bruno Guillaume, Sylvain Kahane, and Guy Perrier (2019). "Improving Surface-syntactic Universal Dependencies (SUD): surface-syntactic relations and deep syntactic features." In TLT 2019-18th International Workshop on Treebanks and Linguistic Theories.
- Gibson, E. (2000). The Dependency Locality Theory: A Distance-Based Theory of Linguistic Complexity. In Marantz, A. P., Miyashita, Y., & O'Neil, W. (Eds.), *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium* (pp. 95-126). MIT Press, Massachusetts, US.
- Gildea, D., & Temperley, D. (2010). Do grammars minimize dependency length? *Cognitive Science*, 34(2), 286-310.
- Givón, T. (1971). On the verbal origin of the Bantu verb suffixes. *Studies in African linguistics*, 2(2), 145.
- Grodner, D., & Gibson, E. (2005). Consequences of the Serial Nature of Linguistic Input for Sentential Complexity. *Cognitive Science*, 29(2), 261-290.
- Guiller, K. (2020). Analyse syntaxique automatique du pidgin-créole du Nigeria à l'aide d'un transformer (BERT): Méthodes et Résultats. *Mémoire de Master, Sorbonne Nouvelle*.
- Haspelmath, M. (2016). The serial verb construction: Comparative concept and cross-linguistic generalizations. *Language and Linguistics*, 17(3), 291-319.
- Hatori, J., Matsuzaki, T., Miyao, Y., & Tsujii, J. I. (2012, July). Incremental joint approach to word segmentation, POS tagging, and dependency parsing in Chinese. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1045-1053).
- Hu, A., Zhang, J., & Liu, J. (2013). Chinese terminology extraction based on an improved C-value method. *Modern Library and Information Technology*, 29(2), 24-29. [胡阿沛, 张静, 刘俊丽. 基于改进 C-value方法的中文术语抽取[J]. 现代图书情报技术, 2013, 29(2):24-29.]
- Hu, Y. (1995). *Modern Chinese*. Shanghai Educational Publishing House. [胡裕树. 现代汉语. 上海: 上海教育出版社, 1995]
- Huang, B., & Liao, X. (2012). *Modern Chinese*. Peking University Press. [黄伯荣, 廖序东. 现代汉语. 2012.]
- Huang, C. (2006). Focus on Bakeoff. In Zhang, P., & Lin, S. (Eds.), *Research and Application of Digital Chinese Language Teaching* (pp. 20-27). Hong Kong City University. [黄昌宁. 聚焦 Bakeoff[A]. 张普, 蔺荪, 等编. 数字化汉语教学的研究与应用[c]. 香港城市大学. 2006.]
- Huang, C. R., Chen, F. Y., Chen, K. J., Gao, Z. M., & Chen, K. Y. (2000, October). Sinica treebank: Design criteria, annotation guidelines, and on-line interface. In *Second Chinese Language Processing Workshop* (pp. 29-37).
- Huang, C. T. J. (1984). Phrase structure, lexical integrity, and Chinese compounds. *Journal of the Chinese Language Teachers Association*, 19(2), 53-78.
- Huang, Y. (1995). A Study on Compound Words. *Linguistics Abroad*, (2), 19. [黄月圆. 复合词研究, 国外语言学. 1995.]

- Humbley, J. (1997). Is terminology specialized lexicography? The experience of French-speaking countries. *HERMES-Journal of Language and Communication in Business*, (18), 13-31.
- Jackendoff, R. S. (1972). *Semantic interpretation in generative grammar*. Cambridge, Mass.: MIT Press.
- Jin, G., & Chen, X. (2008). The fourth international chinese language processing bakeoff: Chinese word segmentation, named entity recognition and chinese pos tagging. In *Proceedings of the sixth SIGHAN workshop on Chinese language processing*.
- Kageura, K., & Umino, B. (1996). Methods of automatic term recognition: A review. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 3(2), 259-289.
- Kahane, S., Vanhove, M., Ziane, R., & Guillaume, B. (2021, March). A morpheme-based treebank for Beja. In *TLT 2021-20th International Workshop on Treebanks and Linguistic Theories*.
- Kilicoglu, E., & Kaplan, A. (2022). Predicting the Mathematical Abstraction Processes Using the Revised Bloom's Taxonomy: Secondary School 7th Graders. *Athens Journal of Education*, 9(2), 237-256.
- Kondratyuk, D., & Straka, M. (2019). 75 languages, 1 model: Parsing universal dependencies universally. *arXiv preprint arXiv:1904.02099*.
- Kratochvíl, P. A. U. L. (1966). Modern standard chinese. *Lingua*, 17(1-2), 129-152.
- Lee, J. S., & Hsiang, J. (2019). Measuring patent claim generation by span relevancy. *arXiv preprint arXiv:1908.09591*.
- Lee, J. S., & Hsiang, J. (2020). Patent claim generation by fine-tuning OpenAI GPT-2. *World Patent Information*, 62, 101983.
- Leung, H., Poiret, R., Wong, T. S., Chen, X., Gerdes, K., & Lee, J. S. (2016, December). Developing universal dependencies for Mandarin Chinese. In *Proceedings of the 12th workshop on Asian Language Resources (ALR12)* (pp. 20-29).
- Li, H., Zhang, Z., Ju, Y., & Zhao, H. (2018, April). Neural character-level dependency parsing for Chinese. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 32, No. 1).
- Li, J. (1962). *New Chinese Grammar*. Commercial Press. [黎錦熙. 新著國語文法. 商務印書館, 1962.]
- Li, X., Meng, Y., Sun, X., Han, Q., Yuan, A., & Li, J. (2019). Is word segmentation necessary for deep learning of Chinese representations?. *arXiv preprint arXiv:1905.05526*.
- Li, X. (1982). A problem in the study of Chinese word formation: On the structure of words like "nursing illness," "fighting fire," and "standing up against injustice". *Language Research*, (2).[李行健 汉语构词法研究中的一个问题——关于“养病”“救火”“打抱不平”等词语的结构. 语文研究. 1982.]
- Li, Y & Gerdes, K. (2019). Chinese Word Segmentation with External Lexicons on Patent Claims. *TOTh 2019*.

## References

- Li, Y., Dong, C., & Gerdes, K. (2019, August). Character-level annotation for Chinese surface-syntactic universal dependencies. In *Depling 2019-International Conference on Dependency Linguistics*.
- Li, Y. (2023). Character-level Dependency Annotation of Chinese. In *Proceedings of the Seventh International Conference on Dependency Linguistics (Depling, GURT/SyntaxFest 2023)* (pp. 42-53).
- Li, Z. (2011). Parsing the internal structure of words: a new paradigm for Chinese word segmentation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 1405-1414).
- Li, Z., & Wei, Q. Introduction to the Cooperative Patent Classification (CPC) System. Patent Examination Cooperation Center, Beijing. [李真 魏巧莲. 联合专利分类 CPC 系统介绍. 专利审查协作北京中心]
- Li Z., Zhou G. (2012). Unified dependency parsing of Chinese morphological and syntactic structures. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 1445–1454.
- Li, Z., Zhou, J., Zhao, H., Zhang, Z., Li, H., & Ju, Y. (2022). Neural Character-Level Syntactic Parsing for Chinese. *Journal of Artificial Intelligence Research*, 73, 461-509.
- Liang, J. (1991). A Preliminary Investigation of the Number of Characters in Chinese Scientific and Technical Terms and Related Issues. *China Terminology*. [梁际翔: "汉语科技术语构词字数及有关问题初步研究." 中国科技术语 02 (1991): 74.]
- Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2), 159-191.
- Liu, H. (2007). Probability Distribution of Dependency Distance. *Glottometrics*, 15, 1-12.
- Liu, H. (2008). Dependency Distance as a Metric of Language Comprehension Difficulty. *Journal of Cognitive Science*, 9(2), 159-191.
- Liu, H., Hudson, R., & Feng, Z. (2009). Using a Chinese treebank to measure dependency distance.
- Liu, H., Xu, C., & Liang, J. (2015). Dependency length minimization: Puzzles and Promises. arXiv preprint arXiv:1509.04393.
- Liu, H., Xu, C., & Liang, J. (2017). Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of life reviews*, 21, 171-193.
- Liu, J., Tang, H. F., & Liu, W. Y. (2014). A Chinese terminology extraction method based on statistical techniques. *China Terminology*, 16(5), 10-14. [刘剑, 唐慧丰, 刘伍颖. 一种基于统计技术的中文术语抽取方法[J]. 中国科技术语, 2014, 16(5):10-14.]
- Liu, Q., & Huang, Z. H. (2003). Some characteristics that scientific and technical terminologies should have. *China Terminology*, 5(01), 22. [刘青 & 黄昭厚. "科技术语应具有的若干特性." 中国科技术语 5, no. 01 (2003): 22.]
- Liu, S. (1990). *Chinese Descriptive Lexicology*. The Commercial Press. (p. 46). [刘叔新. 汉语描写词汇学. 北京:商务印书馆. 1990.]

- Liu, Y., Lin, Z., & Kang, S. (2018). The extraction of morpheme concepts in Chinese and semantic word formation analysis. *Journal of Chinese Information Processing*, 32(2), 12-21. [刘扬, 林子, 康司辰. "汉语的语素概念提取与语义构词分析." 中文信息学报 32, no. 2 (2018): 12-21.]
- Liu, Y., Li, Y., & Huang, Y. (2016). Research on semantic and syntactic analysis of patent literature. *ICIC Express Letters*, 10(2), 471-477.
- Lu, J. (2003). *A Course in Modern Chinese Grammar Research*. Peking University Press. [陆俭明. 现代汉语语法研究教程[M]. 北京: 北京大学出版社. 2003.]
- Lu, Z. (1957). *The Morphology of Chinese*. Science Press. [陆志韦. 汉语的构词法. 科学出版社. 1957.]
- Lu, Q., Xu, C., & Liu, H. (2015). The influence of chunking on dependency crossing and distance. *arXiv preprint arXiv:1509.01310*.
- Lu, Q., Xu, C., & Liu, H. (2016). Can chunking reduce syntactic complexity of natural languages?. *Complexity*, 21(S2), 33-41.
- Lü, S., & Zhu, D. (1979). *Lectures on Grammar and Rhetoric*. China Youth Publishing House. [吕叔湘, 朱德熙. 语法修辞讲话[M]. 北京中国青年出版社. 1979.]
- Lü, S. (2001). Problems in Chinese Grammar Analysis. In Huang, G. Y. (Ed.), *Selected Works of Lü Shuxiang*. Northeast Normal University Press. (Originally published 1979). [吕叔湘. 汉语语法分析问题[A]. 商务印书馆. 1979. 黄国营编. 吕叔湘选集[C]. 长春: 东北师范大学出版社. 2001.]
- L'homme, M. C. (2004). *La terminologie: principes et techniques*. Pum.
- L'Homme, M. C. (2004). A Lexico-semantic Approach to the Structuring of Terminology. In *Proceedings of CompuTerm 2004: 3rd International Workshop on Computational Terminology* (pp. 7-14).
- Magistry, P., & Sagot, B. (2012, July). Unsupervised word segmentation: the case for mandarin chinese. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 383-387).
- Magistry, P. (2013). *Unsupervised Word Segmentation and Wordhood Assessment* (Doctoral dissertation, Paris Diderot; Inria).
- Mel'čuk, I. (1998). Collocations and lexical functions. *Phraseology. Theory, analysis, and applications*, 23-53.
- Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT press.
- Muller, B., Sagot, B., & Seddah, D. (2020). Can multilingual language models transfer to an unseen dialect? A case study on north african arabizi. *arXiv preprint arXiv:2005.00318*.
- Nakagawa, H. (2001). Experimental evaluation of ranking and selection methods in term extraction. *Bourigault D, L'Homme M.-C., Jacquemin C.(éd.), Recent advances in computational terminology, John Benjamins Publishing Company*, 303-26.

## References

- Nguyen, É. V. T. (2006). *Unité lexicale et morphologie en chinois mandarin: vers l'élaboration d'un dictionnaire explicatif et combinatoire du chinois*. PhD thesis, Université de Montréal (Canada).
- Oh, T. H., Han, J. Y., Choe, H., Park, S., He, H., Choi, J. D., ... & Kim, H. (2020). Analysis of the Penn Korean Universal Dependency treebank (PKT-UD): Manual revision to build robust parsing model in Korean. *arXiv preprint arXiv:2005.12898*.
- Oya, M. (2011). Syntactic dependency distance as sentence complexity measure. In *Proceedings of the 16th International Conference of Pan-Pacific Association of Applied Linguistics (Vol. 1)*.
- Oya, M. (2021). Three types of average dependency distances of sentences in a multilingual parallel corpus. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation* (pp. 652-661).
- Packard, J. L. (2000). *The morphology of Chinese: A linguistic and cognitive approach*. Cambridge University Press.
- Patent Law of the People's Republic of China* (2020). Agent, C. P.
- Patent Examination Guidelines* (2021) State Intellectual Property Office of the People's Republic of China. [专利审查指南. 中华人民共和国知识产权局. 2021.]
- Peng, Z., Gerdes, K., & Guiller, K. (2022, November). Pull your treebank up by its own bootstraps. In *Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain (LIFT) et Traitement Automatique des Langues (TAL)* (pp. 139-153). CNRS.
- Porter, A. L. (2009). Tech mining for future-oriented technology analyses. *Futures research methodology*, 1-20.
- Qiu, H., & Shi, L. (2015). Construction of multi-domain chinese dependency treebanks and analysis of influencing factors on dependency parsing. *Journal of Chinese Information Processing*, 29(5), 69.
- Qiu, X., Pei, H., Yan, H., & Huang, X. (2019). A concise model for multi-criteria Chinese word segmentation with transformer encoder. *arXiv preprint arXiv:1906.12035*.
- Royle, P. (2001). Cabré, M. Teresa (1998): Terminology: Theory, methods and applications, Philadelphia PA, John Benjamins, 248 p.[transl. of La Terminologia. La teoria, els mètodes, les aplicacions, Barcelona, Emúries, 1992]. *Meta*, 46, 3.
- Ruzsics, T., Sozinova, O., Gutierrez-Vasques, X., & Samardzic, T. (2021, April). Interpretability for morphological inflection: from character-level predictions to subword-level rules. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 3189-3201).
- Sager, J. C. (1990). *Practical course in terminology processing*. John Benjamins Publishing.
- Savary, A. (2008). Computational inflection of multi-word units: A contrastive study of lexical approaches. *Linguistic Issues in Language Technology*, 1.
- Selkirk, E. (1982). The Syntax of Words. *Linguistic Inquiry Monographs*, 7. Cambridge, MA: MIT Press.
- Selkirk, E. (1984). On the Major Class Features and Syllable Theory. In *Language Sound Structure*.

- Shao, J. (Ed.). (2016). *An Introduction to Modern Chinese*. Shanghai Educational Publishing House. (p. 87). [邵敬敏主编. 现代汉语通论. 上海:上海教育出版社. 2016]
- Sproat, R. (1996). Multilingual text analysis for text-to-speech synthesis. *Natural Language Engineering*, 2(4), 369-380.
- Sproat, R., & Shih, C. (2002). Corpus-based methods in Chinese morphology and phonology. In *Proceedings of the 19th International Conference on Computational Linguistics*.
- Sproat, R., & Emerson, T. (2003). The first international Chinese word segmentation bakeoff. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing* (pp. 133-143).
- Su, X. (2001). Reflections on the quantitative study of vocabulary in the Modern Chinese Dictionary. *Chinese Teaching in the World*, 2001(4), (Issue 58). [苏新春. 关于《现代汉语词典》词汇计量研究的思考. 世界汉语教学. 2001年第4期(总第58期)]
- Szopinski, D., Schoormann, T., & Kundisch, D. (2020). Criteria as a Prelude for Guiding Taxonomy Evaluation.
- Tang, T. (1991). The "Incorporation Phenomenon" in Chinese Grammar. *Tsinghua Journal (Taiwan)*, New Series, 21(1). [汤廷池. (1991). 汉语语法的“并入现象”. 清华学报》(台湾)新, 21(1).]
- Tang, T. (1992). The Structure, Function, and Origin of Complement-Compound Verbs in Chinese. In *Chinese Morphology and Syntax*, 95-154. [湯廷池. (1992). 漢語述補式複合動詞的結構, 功能與起源. 漢語詞法句法, 95-154.]
- Tang, T. (1994). Contrastive Analysis and Grammatical Theory: [X-bar Theory] and [Case Theory]. *Journal of Japanese Language Literature in Taiwan*, (6), 1-40. [湯廷池. (1994). 對比分析與語法理論:[X 標槓理論] 與 [格位理論]. 台灣日本語文學報, (6), 1-40.]
- Tesnière, L. (1959). *Éléments de syntaxe structurale*. Klincksieck Paris.
- Tran, H. T. H., Martinc, M., Caporusso, J., Doucet, A., & Pollak, S. (2023). The Recent Advances in Automatic Term Extraction: A Survey. *arXiv preprint arXiv:2301.06767*.
- Trojar, M. (2017). Wüster's View of Terminology. *Slovenski jezik/Slovene Linguistic Studies*, 11.
- Tsou, B. K., Chow, K. P., Lee, J. S., Yip, K. F., Ji, Y., & Wu, K. (2020, October). Bilingual Multi-word Expressions, Multiple-correspondence, and their cultivation from parallel patents: The Chinese-English case. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation* (pp. 589-602).
- Zhou, W., Ge, T., Xu, K., Wei, F., & Zhou, M. (2019). BERT-based Lexical Substitution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3368-3373). Florence, Italy: Association for Computational Linguistics.
- Wang, H. (1994). Observing words and phrases from characters and character groups: A discussion on the criteria for word division in Chinese. *Chinese Language*, 1994(2), 102-112. [王洪君. 从字和字组看词和短语——也谈汉语中词的划分标准[J]. 中国语文, 1994(2): 102-112.]

## References

- Wang, H. (1998). The internal structure of "boxing" and "nurturing injuries" from the analogy of free phrases. *Language Research*, (4). [王洪君. 从与自由短语的类比看“打拳”、“养伤”的内部结构. 语文学研究第4期. 1998.]
- Wang, L. (1980). *Collected Works of Longchongbingdiao Zhai*. Zhonghua Book Company. [王力. 龙虫并雕斋文集. 北京:中华书局. 1980.]
- Wang, Y., & Liu, H. (2017). The effects of genre on dependency distance and dependency direction. *Language Sciences*, 59, 135-147.
- Wang, S. (2020). *Chinese multiword expressions*. Singapore: Springer Singapore.
- Wu, A. (2003). Customizable Segmentation of Morphologically Derived Words in Chinese. *Computational Linguistics and Chinese Language Processing*, 8(1), 1-28.
- Wüster, E. (1931). *Internationale Sprachnormung in der Technik, besonders in der Elektrotechnik*. VDI-Verlag.
- Xia, F. (2000a). Segmentation Guidelines for the Penn Chinese Treebank (3.0). University of Pennsylvania Institute for Research in Cognitive Science Technical Report No. IRCS-00-06. Retrieved from [http://repository.upenn.edu/ircs\\_reports/37/](http://repository.upenn.edu/ircs_reports/37/)
- Xia, F. (2000b). The Part-of-Speech Tagging Guidelines for the Penn Chinese Treebank (3.0). University of Pennsylvania Institute for Research in Cognitive Science Technical Report No. IRCS-00-07. Retrieved from [http://repository.upenn.edu/ircs\\_reports/38/](http://repository.upenn.edu/ircs_reports/38/)
- Xu, C., Shi, S. C., Fang, X., et al. (2013). Chinese patent literature terminology extraction. *Computer Engineering and Design*, 34(6), 2175-2179. [徐川, 施水才, 房祥, 等. 中文专利文献术语抽取[J]. 计算机工程与设计, 2013, 34(6):2175-2179.]
- Yan, H., Qiu, X., & Huang, X. (2020). A graph-based model for joint Chinese word segmentation and dependency parsing. *Transactions of the Association for Computational Linguistics*, 8, 78-92.
- Yang, S. L., Lü, X. Q., Li, Z., et al. (2016). Research on automatic identification of terminology in Chinese patent literature. *Journal of Chinese Information Processing*, 30(3), 11-17. [杨双龙, 吕学强, 李卓, 等. 中文专利文献术语自动识别研究[J]. 中文信息学报, 2016, 30(3):11-17.]
- Yu, Y., Chen, L., Jiang, J., & Zhao, N. (2019). Research on Chinese Patent Candidate Term Selection Based on Dependency Syntax Analysis. *Library and Information Service*, 63(18), 109. [俞琰, 陈磊, 姜金德, and 赵乃瑄. "基于依存句法分析的中文专利候选术语选取研究." 图书情报工作 63, no. 18 (2019): 109.]
- Yuan, C. F., & Huang, C. N. (1998). Research on Chinese morphemes and word formation based on a morpheme database. *Applied Linguistics*, 3, 86-91. [苑春法 & 黄昌宁. 基于语素数据库的汉语语素及构词研究. 语言文字应用 3 (1998): 86-91.]
- Yue, J. Y., Xu, J. A., & Zhang, Y. J. (2013). Research on Chinese word segmentation technology for patent literature. *Journal of Peking University: Natural Science Edition*, 49(1), 159-164. [岳金媛, 徐金安, 张玉洁. "面向专利文献的汉语分词技术研究." 北京大学学报: 自然科学版 49, no. 1 (2013): 159-164.]



- Zeman, D., Hajic, J., Popel, M., Potthast, M., Straka, M., Ginter, F., Nivre, J., & Petrov, S. (2018). CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies* (pp. 1-21).
- Zeng, Z., Lü, X., & Li, Z. (2016). A method for extracting domain terminology from patent abstracts. *Computer Applications and Software*, 33(3), 48-51. [曾镇, 吕学强, and 李卓. "一种面向专利摘要的领域术语抽取方法." 计算机应用与软件 33, no. 3 (2016): 48r51.]
- Zhang, G., Liu, D., Yin, B., Xu, L., & Miao, X. (2010). Chinese word segmentation techniques for patent documents. *Journal of Chinese Information Processing*, 24(3), 112-116. [张桂平, 刘东生, 尹宝生, 徐立军, and 苗雪雷. "面向专利文献的中文分词技术的研究." 中文信息学报 24, no. 3 (2010): 112-116.]
- Zhang, J., Zhang, H., & Zhai, D. (2014). Research on segmentation methods for Chinese patent claims. *Modern Library and Information Technology*, 9, 91-98. [张杰, 张海超, 翟东升. 面向中文专利权利要求书的分词方法研究. 现代图书情报技术 9 (2014): 91-98.]
- Zhang, J., Zhao, Y., Saleh, M., & Liu, P. (2020). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning* (pp. 11328-11339). PMLR.
- Zhang, M., Zhang, Y., Che, W., & Liu, T. (2013). Chinese Parsing Exploiting Characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*.
- Zhang, M., Zhang, Y., Che, W., & Liu, T. (2014). Character-Level Chinese Dependency Parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*.
- Zhang, M. (2020). A survey of syntactic-semantic parsing based on constituent and dependency structures. *Science China Technological Sciences*, 63(10), 1898-1920.
- Zhang, S. (1907). *Intermediate Chinese Grammar*. [章士钊. 中等国文典. 商务印书馆. 1911.]
- Zhang, S. (1956). *The Grammatical Analysis of Texts*. [張世祿. 文章的語法分析. 1956.]
- Zhang, W. (2004). The development and structure of the Chinese Thesaurus for subject indexing. *The International Information & Library Review*, 36(1), 47-54.
- Zhai, D., & Ma, W. (2011). Research on Chinese patent claim segmentation algorithm. *Journal of Intelligence*, 30(11), 152-155. [翟东升, and 马文姗. "中文专利权利要求书分词算法研究." 情报杂志 30, no. 11. 2011.]
- Zhao, H. (2009). Character-level dependencies in Chinese: Usefulness and learning. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL)*, 879-887.
- Zheng, H., Liu, Y., Yin, Y. Q., Wang, Y., & Dai, D. M. (2022). Research on recognition of Chinese word formation structures based on word information embedding. *Journal of Chinese Information Processing*, 36(5), 31-40. [郑娅, 刘扬, 殷雅琦, 王悦, and 代达劼. "基于词信息嵌入的汉语构词结构识别研究." 中文信息学报 36, no. 5. 2022.]

## References

- Zheng, S. (2004). Research methods in terminology studies. *Terminology Standardization & Information Technology*, 2, 5-7. [郑述谱. 术语学的研究方法. 术语标准化与信息技术 2. 2004.]
- Zhou, J. (2014). *Chinese Vocabulary Structure Theory*. People's Education Press. (pp. 147-149). [周荐. 汉语词汇结构论. 北京:人民教育出版社. 2014.]
- Zhou, J. (2016). *On Lexicology* (1st ed.). Commercial Press. [周荐. 词汇论. 商务印书馆, 第一版 2016.]
- Zhou, Z. (1959). *Lectures on Chinese Vocabulary*. People's Education Press. [周祖谟. 汉语词汇讲话. 人民教育出版社. 1959.]
- Zhu, D. (1982). *On grammar*. Beijing: Commercial Press. [朱德熙. 语法讲义. 商务印书馆. 1982.]
- Zhu, D. (1985). PhD dissertation on the grammaticalization of verbs and nominal verbs in modern written Chinese. [朱德熙. 现代书面汉语里的虚化动词和名动词 为第一届国际汉语教学讨论会而作. 1985.]
- Zhu, H., & Lei, L. (2021). Jingyang Jiang & Haitao Liu (eds.), Quantitative analysis of dependency structures (*Quantitative Linguistics* 72). Berlin: De Gruyter Mouton, 2018. Pp. xii+ 368. *Journal of Linguistics*, 57(2), 449-453.
- Zou, S. (Ed.). (1988). *Modern Terminology and Dictionary Compilation*. [邹树明等编译. 现代术语学与辞书编纂. 科学出版社. 1988.]
- You, Z., Li, Y., Parias García, A., & Gerdes, K. (2022). Technological Taxonomies for Hypernym and Hyponym Retrieval in Patent Texts. *TOTh* 2022.

# Appendix

## Appendix A - Complete annotated term lists for each deprel

conj\_m: 195

{'程度', '符号', '发送', '更动', '设定', '行驶', '验证', '添加', '模型', '方法', '弯曲', '管理', '扭转', '积分', '系统', '进而', '业务', '创建', '约束', '地址', '方式', '按压', '安装', '感测', '作为', '区域', '请求', '距离', '查询', '同一', '分布', '绘制', '搜索', '继续', '参考', '信息', '网络', '携带', '拟合', '变化', '触发', '阶段', '统计', '获得', '需要', '进行', '设备', '边界', '配置', '出现', '具有', '分析', '汇聚', '关联', '围绕', '铸造', '以及', '选择', '程序', '要求', '重新', '标识', '权利', '判定', '运算', '监视', '容纳', '行为', '情况', '识别', '存放', '划分', '判断', '解析', '调用', '形成', '使用', '软硬', '形式', '匹配', '安全', '触摸', '发生', '类型', '接收', '传输', '查找', '依据', '产生', '侦测', '所述', '寄存', '指示', '步骤', '测试', '间隔', '均匀', '选取', '饱和', '架构', '筛选', '存储', '利用', '兴趣', '状态', '提取', '施加', '关键', '收集', '执行', '控制', '操作', '规则', '应用', '响应', '目的', '指令', '包含', '停止', '尺寸', '校验', '作用', '确定', '服务', '调整', '位置', '消息', '处理', '展示', '重复', '设计', '构建', '实现', '设置', '计算', '包括', '转换', '通知', '显示', '检测', '修改', '据根', '按照', '封闭', '关系', '存在', '推导', '结构', '完成', '电磁', '煤电', '初始', '分别', '生产', '振荡', '适应', '基准', '启动', '连结', '是否', '索引', '插接', '监测', '基础', '历史', '移动', '部分', '运行', '装置', '即为', '开关', '规格', '交换', '传送', '测量', '采集', '对应', '方案', '条件', '卡片', '扫描', '学习', '认知', '建立', '机器', '保存', '获取', '时刻', '保障', '规范', '结合', '删除', '累积', '图形', '连接'}

mod\_m: 248

{'符号', '版本', '变电', '电压', '自度', '稳态', '遍历', '信号', '热值', '内核', '面积', '参数', '模型', '快照', '资料', '配表', '壳体', '自身', '智能', '相关', '客车', '来源', '自容', '内存', '常用', '最终', '电源', '符表', '电极', '间段', '地址', '信表', '封式', '接卡', '节点', '含率', '平台', '终端', '预定', '间戳', '用户', '压感', '绘制', '集极', '油层', '回传', '全部', '复用', '两个', '桶状', '控笔', '数据', '共振', '要素', '电子', '处器', '以口', '非显', '四则', '事表', '软件', '时长', '扣卡', '元据', '软测', '质心', '相应', '实时', '内部', '存区', '点图', '瞬态', '存器', '均值', '地图', '经验', '运算', '虚拟', '水平', '寄存器', '算法', '马氏', '每个', '预处', '地质', '结体', '页面', '梯度', '图层', '预设', '物理', '主板', '方钢', '头像', '内容', '基座', '基站', '平动', '力量', '指针', '全局', '不同', '分式', '频率', '实质', '功率', '选器', '状部', '强度', '根据', '遥测', '网格', '主机', '平井', '特征', '向量', '孔隙', '兴点', '哪个', '硅油', '键帽', '宿主', '工区', '异常', '渗率', '接点', '桌面', '制剂', '轮胎', '上方', '腔体', '连线', '似度', '同时', '感区', '目的', '键盘', '相似', '曲线', '液冷', '栅控', '反映', '镜像', '互信', '表单', '确定', '函数', '相渗', '走线', '整车', '上述', '多条', '传器', '空间', '对比', '图标', '户端', '后端', '可化', '运器', '压力', '工况', '前卡', '其中', '一条', '缓存', '输端', '金属', '相对', '线图', '嵌槽', '变站', '车身', '卡点', '相同', '插卡', '据库', '单元', '按键', '面板', '未读', '滑动', '中板', '接度', '学生', '日志', '偏置', '隙度', '模块', '基准', '控件', '电路', '过程', '阈值', '积比', '时间', '容式', '界面', '卫星', '等值', '流量', '接着', '数量', '导体', '目录', '客户', '优化', '厚度', '直接', '地线', '拟机', '正常', '累值', '散点', '触控', '源据', '系数', '指数', '一种', '对象', '同步', '服端', '对应', '饱和度', '整体', '芯片', '密封', '区块', '多个', '合性', '接口', '电容', '电流', '数学', '煤比', '环境', '差值', '变量', '模态', '元件', '之间', '发机', '服器', '复制', '号线', '显区', '图形', '目标'}

subj\_m: 3

{'水淹', '针对', '自适'}

comp\_m: 103

{'事件', '取证', '使得', '升级', '小于', '绑定', '返回', '积分', '最大', '可配', '传感', '通话', '做差', '接地', '发电', '作为', '合法', '带入', '方向', '复用', '当前', '以下', '载荷', '可视', '接近', '从中', '以及', '通过', '输入', '选中'}

## Appendix

','转向','器件','生成','然后','运算','行为','超过','待执','划分','清零','评价','在线','联网','文件','所述','','开机','指向','等于','待验','结果','定义','大于','调谐','燃煤','通信','来自','如下','应用','在轨','含油','','面向','驱动','表征','第二','作用','服务','插入','所有','一个','依序','渗透','得到','修正','含水','签名','','以上','至少','输出','加载','接入','第三','制冷','部分','制动','随意','可信','第一','建模','按下','同步','','取值','卡片','发热','导电','交互','自由','时刻','任意','浸没','能够','变电','累积','是否'}

flat\_m: 7

{'其中','拓扑','以太','部分','或者','以及','思维'}

## Appendix B - Sentences with the WSS score lower than 0.8

### Without the pretrained model:

1. 所述第二导电部分相对于所述第二感测区的面积比与所述第一导电部分相对于所述第一感测区的面积比不同。
2. 将选取到的至少部分第一类参考信息对应的网络地址的位置信息作为异常的网络地址的位置信息。
3. 解析所述验证配置文件以生成与所述待验证数据对象对应的规则对象；
4. 基于与所述待验证数据对象对应的规则对象对所述待验证数据对象进行验证，并输出验证结果信息；
5. 2.2在步骤2.1建立的应用电路中采集多组测试数据，所述测试数据包括所述解析软测量建模筛选出的所述栅控器件的关键参数和对应的栅控器件的集电极电流；
6. 2.3根据步骤2.2采集的多组测试数据建立统计数据模型，即利用机器学习算法建立所述关键参数与对应的栅控器件的集电极电流之间的关系；
7. 所述测量阶段为：
8. 在栅控器件应用电路中，实时采集所述关键参数，并将采集的关键参数带入经验数据软测量建模建立的统计数据模型进行计算，得到所述栅控器件的集电极电流。
9. 一桶状部；
10. 1.一种水平井水淹程度评价方法，其特征在于，该水平井水淹程度评价方法包括：

**With the pretrained model:**

1. 基于所述至少一条第一类参考信息中的所述用户数量, 选取得到异常的网络地址的位置信息, 对所述异常的网络地址的位置信息进行修正以使得修正后的所述网络地址的位置信息与用户常用地址相同。
2. 将选取到的至少部分第一类参考信息对应的网络地址的位置信息作为异常的网络地址的位置信息。
3. 解析所述验证配置文件以生成与所述待验证数据对象对应的规则对象;
4. 基于与所述待验证数据对象对应的规则对象对所述待验证数据对象进行验证, 并输出验证结果信息;
5. 1.一种用于栅控器件的集电极电流软测量方法, 其特征在于, 包括解析软测量建模、经验数据软测量建模和测量阶段,
6. 所述经验数据软测量建模包括如下步骤:
7. 2.2在步骤2.1建立的应用电路中采集多组测试数据, 所述测试数据包括所述解析软测量建模筛选出的所述栅控器件的关键参数和对应的栅控器件的集电极电流;
8. 所述测量阶段为:
9. 数学思维可视化模块, 数据收集模块, 数据统计模块, 以数据通信方式连接。
10. 2.如权利要求1所述的基于卡片点的数学图形认知思维可视化系统, 其特征在于, 所述数学思维可视化模块, 通过学生在终端界面上对卡片点的绘制连线显示其思维过程。
11. 一桶状部;
12. 一接近传感器, 设定为根据接近传感器上的触摸的存在或不存在, 来输出一接近度接近数据;
13. 一处理电路, 具有一通信子电路。
14. 其中, 所述基准事件标识用于反映所述设备上已保存的数据信息, 所述目标事件标识用于反映所述设备上未保存的数据信息;

## Appendix C - Sentences with low LAS score

### Sentence 88

LAS: 0.4318181818181818

基于所述至少一条第一类参考信息中的所述用户数量, 选取得到异常的网络地址的位置信息, 对所述异常的网络地址的位置信息进行修正以使得修正后的所述网络地址的位置信息与用户常用地址相同。

WSS: 0.7386363636363636

基于所述至少一条第一类参考信息中的所述用户数量, 选取得到异常的网络地址的位置信息, 对所述异常的网络地址的位置信息进行修正以使得修正后的所述网络地址的位置信息与用户常用地址相同。

### Sentence 43

LAS: 0.4186046511627907

将选取到的至少部分第一类参考信息对应的网络地址的位置信息作为异常的网络地址的位置信息。

WSS: 0.7441860465116279

将选取到的至少部分第一类参考信息对应的网络地址的位置信息作为异常的网络地址的位置信息。

### Sentence 44

LAS: 0.4772727272727273

基于与所述待验证数据对象对应的规则对象对所述待验证数据对象进行验证, 并输出验证结果信息;

WSS: 0.7727272727272727

基于与所述待验证数据对象对应的规则对象对所述待验证数据对象进行验证, 并输出验证结果信息;

### Sentence 61

LAS: 0.47540983606557374

2.3根据步骤2.2采集的多组测试数据建立统计数据模型, 即利用机器学习算法建立所述关键参数与对应的栅控器件的集电极电流之间的关系;

Sentence 34

LAS: 0.3235294117647059

数学思维可视化模块, 数据收集模块, 数据统计模块, 以数据通信方式连接。

WSS: 0.7058823529411765

数学思维可视化模块, 数据收集模块, 数据统计模块, 以数据通信方式连接。

Sentence 75

LAS: 0.4

2.如权利要求1所述的基于卡片点的数学图形认知思维可视化系统, 其特征在于, 所述数学思维可视化模块, 通过学生在终端界面上对卡片点的绘制连线显示其思维过程。

WSS: 0.7466666666666667

2.如权利要求1所述的基于卡片点的数学图形认知思维可视化系统, 其特征在于, 所述数学思维可视化模块, 通过学生在终端界面上对卡片点的绘制连线显示其思维过程。

Sentence 5

LAS: 0.2

一桶状部;

WSS: 0.6

一桶状部;

Sentence 41

LAS: 0.3902439024390244

一接近传感器, 设定为根据接近传感器上的触摸的存在或不存在, 来输出一接近度接近数据;

WSS: 0.7560975609756098

一接近传感器, 设定为根据接近传感器上的触摸的存在或不存在, 来输出一接近度接近数据;

## Appendix

### Sentence 31

LAS: 0.45161290322580644

步骤6, 对水平井进行水淹程度划分, 并评价水平井整体的水淹情况。

### Sentence 110

LAS: 0.4727272727272727

2.根据权利要求1所述的能任意更动按键位置的键盘, 其特征在于, 每个按键模块的识别电路为调谐电路且具有一共振频率, 当一侦测信号的频率实质上相同于该共振频率时, 该扫描单元所接收到的回传信号的频率将实质上相同于该侦测信号的频率。

### Sentence 97

LAS: 0.4845360824742268

运行在主操作系统的客户端在接收到服务端发送的第一签名结果后, 调用运行在可信执行环境TEE中的对应可信应用TA, 其中, 所述第一签名结果由所述服务端针对所述客户端发送的业务要素进行签名后返回给所述客户端;



# Abstract

## **Title: Tech-mining on Chinese Patents: Syntax and Terminology**

This thesis aims to contribute to the field of research by automating the generation of lexical variations for technical terms found in Chinese patent claims. It achieves this through two primary contributions. Firstly, a character-level dependency parser specifically pre-trained on Chinese patent claims is developed. This parser enables the analysis of the internal structure of the terms and thus avoids the long-existing segmentation problem in Chinese. Secondly, a technical taxonomy is constructed based on the titles of the International Patent Classification (IPC) system, providing promising hypernym/hyponym substitutes for the production of variants of a base claim text.

Chapter 1 serves as an introduction, providing the necessary linguistic and technical background for the research. In Chapter 2, the collection and preprocessing of the corpus used in the study are detailed. Chapters 3 and 4 focus on annotating the Chinese character-level dependency treebank and describe the training process used to bootstrap the parser. Chapter 5 presents the construction and evaluation of the technical taxonomy, which utilizes the IPC system. Finally, in the end of Chapter 5, the methodology for recognising and selecting lexical variations is demonstrated, employing the measurement of semantic distance.

**Keywords: dependency parsing, Chinese morphology, terminology, automatic term extraction, lexical variation, term substitution, taxonomy, patent**

## **Résumé en français**

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

Cette thèse vise à contribuer au domaine de la recherche en automatisant la génération de variations lexicales pour les termes techniques présents dans les demandes de brevet chinoises. Elle réalise cela grâce à deux contributions majeures. Tout d'abord, un analyseur de dépendance au niveau des caractères, spécifiquement pré-entraîné sur les demandes de brevet chinoises, est développé. Cet analyseur permet d'analyser la structure interne des termes et évite ainsi le problème de segmentation qui existe depuis longtemps en chinois. Deuxièmement, une taxonomie technique est construite en se basant sur les titres de la Classification internationale des brevets (IPC), fournissant des substituts prometteurs d'hyperonymes/hyponymes pour la production de variantes d'un texte de demande de brevet de base.

Le chapitre 1 sert d'introduction, en fournissant les connaissances linguistiques et techniques nécessaires à la recherche. Le chapitre 2 détaille la collecte et la préparation du corpus utilisé dans l'étude. Les chapitres 3 et 4 se concentrent sur l'annotation de l'arbre de dépendance au niveau des caractères chinois et décrivent le processus d'entraînement utilisé pour démarrer l'analyseur. Le chapitre 5 présente la construction et l'évaluation de la taxonomie technique, qui utilise le système de la Classification internationale des brevets. Enfin, à la fin de chapitre 5, la méthodologie de reconnaissance et de sélection des variations lexicales est démontrée, en utilisant la mesure de la distance sémantique.

**Mots-clé: analyse de dépendance, morphologie chinoise, terminologie, extraction automatique de termes, variation lexicale, substitution de termes, taxonomie, brevet**

**Université Sorbonne Nouvelle**

Ecole Doctorale 622 - Sciences du langage,

Laboratoire de Phonétique et Phonologie (LPP) UMR 7018

19, rue des Irlandais Paris