Abstract of Project entitled Chinese Linguistics Academic Video Analysis.

Submitted by Qin Ying, MC256590 for the degree of Master of Science in Data Science (Computational Linguistics) at the University of Macau in April, 2024.

漢語語言學學術視頻分析

摘要: 針對學術語言的研究歷史悠久，隨著時代的快速發展，學術語言不僅僅以書面語的形式出現，越來越多的學術講座、學術會議使得學術語言逐步出現在口語環境之下，而對于特定專業領域學術視頻分析的研究相對較少。伴隨著我國國際地位的日益提升，國內外對于漢語的相關研究也更加重視，但關于漢語語言學學術口語可使用的語料庫和詞表都相對較少，爲了填補這一空白領域的研究，本文針對漢語語言學學術視頻展開研究。本研究搜集幷最終篩選出 2020 年-至今共計 400 個漢語語言學方向的學術視頻，對視頻轉寫後進行文本過濾清洗，最終得到含 9,535,974 字的漢語語言學學術視頻語料庫。參照公開發表的現代漢語口語詞匯表對分詞過後的語料庫進行口語詞匯抽取，對抽取出的詞匯進行統計分析，篩選出按照詞頻降序排列出現頻率最高的「這個」、「就是」和同時在六個現代漢語口語詞匯表中出現的口語詞匯「無所謂」進行單獨的口語詞匯分析，幷且對比分析各語言學分支下詞匯的异同。將篩選出的口語詞匯參照 AWL（Coxhead, 2002）的構建按照詞頻篩選，幷按照頻段分類，構建出「漢語語言學學術口語詞表」，將「漢語語言學學術口語詞表」和《國際中文教育中文水平等級標準詞匯表》（劉英林等, 2021）進行比較，得出「漢語語言學學術口語詞表」在《國際中文教育中文水平等級標準詞匯表》（劉英林等, 2021）中的分布情況。本研究構建了一個成熟的語料庫和一個詞表，爲漢語語言學類的學術語言研究提供了語料來源和詞匯儲備，利用本研究的研究成果

可以更好的進行國際中文教育的推廣和漢語語言學的相關研究。

關鍵字：語料庫；學術詞表；學術口語；詞匯分析

Abstract of Project entitled Chinese Linguistics Academic Video Analysis.

Submitted by Qin Ying, MC256590 for the degree of Master of Science in Data Science (Computational Linguistics) at the University of Macau in April, 2024.

Chinese Linguistics Academic Video Analysis

Abstract: Academic language has been studied for a long time. With the development of the times, academic language has not only appeared in the form of written language, but also in the oral environment with the emergence of more and more academic lectures and academic conferences, while relatively little research has been conducted on the analysis of academic videos in specific professional fields. As China's international status continues to improve, attention to Chinese language-related research at home and abroad is also increasing. However, there are relatively few corpora and lexicons for the linguistics of spoken academic Chinese. To fill this gap, the study examines academic Chinese linguistics videos. In the study, a total of 400 academic videos in the direction of Chinese linguistics from 2020 to the present were collected. After the videos were translated, text filtering and cleaning were performed, and a corpus of Chinese linguistics academic videos containing 9,535,974 words was finally obtained. According to the published modern Chinese spoken vocabulary lists, spoken words were extracted from the corpus after word segmentation, and the extracted words were statistically analyzed. In descending order of word frequency, "這個" and "就是" with the highest frequency of occurrence, as well as "無所謂", which appears in six modern Chinese spoken vocabulary lists at the same time, were selected for separate spoken word analysis, and compared and analyzed the similarities and differences of words in different branches of

linguistics. Based on the construction of AWL(Coxhead, 2002), the selected spoken vocabulary was screened according to word frequency and categorized by frequency bands to construct the Chinese Linguistics Academic Spoken Word List. The Chinese Linguistics Academic Spoken Word List was compared with the International Chinese Language Education Chinese Proficiency Level Standard Vocabulary List. The distribution of Chinese Linguistics Academic Spoken Word List in the International Chinese Language Education Chinese Proficiency Level Standard Vocabulary List was derived. In the study, we have constructed a mature corpus and a word list for academic research in Chinese linguistics. The results of the study can be used to better promote related researches on Chinese international education and Chinese linguistics.

Keywords: Corpus; Academic Word List; Spoken Academic Language; Lexical Analysis