Some Salient Lexical Features of Spoken Academic Chinese and Their Pedagogical Implications

Hongyin Tao

University of California, Los Angeles (UCLA)

Abstract

This study investigates some of the most salient lexical features of spoken academic Chinese (SAC) based on a 150,000 word video-based corpus of academic lectures at institutions of higher education in China. Keywords are identified in comparison with large-scale corpora of everyday conversation and written academic Chinese, which are then subjected to context-based multimodal discourse analysis. Finally, this study describes the theoretical significance of this approach for a better understanding of linguistic genres and some potential pedagogical applications of the research results in teaching Chinese as a second language.

Keywords: academic Chinese, keywords, pronominal, demonstrative, particle, conjunction

1. Introduction

In the field of Chinese for specific and special purposes, spoken academic Chinese is very much an understudied area (Chen& Tao, 2019; Tao& Chen, 2019). Academic language, broadly referring to language used in academic settings, can be observed in either the spoken or the written mode. However, most related research is conducted on written forms such as textbooks, journal articles, research monographs, student theses, and so forth. In the spoken mode, academic language can be realized as lectures, conference presentations, teaching tutorials, roundtable discussions, oral defenses, office hour talk, student oral presentations, etc. (Simpson-Vlach, 2006, 2013). Historically speaking, research on spoken academic language has concentrated heavily on English language-based studies, especially in the field of English for Special/Specific Purposes (ESP), with recent advancement spurred by corpus linguistics (Coxhead, 2000; Biber, 2006; Simpson-Vlach, 2006, 2013; Gardner& Davies, 2013; Thompson& Nesi, 2001; Alsop& Nesi, 2009).

Academic Chinese, as discussed in Chen and Tao (2019) and Tao and Chen (2019), on the other hand, is almost exclusively associated with language use in the science and engineering fields, hence the term scientific language (科學體 *kēxué tǐ*, W. Chen, 1962, 1997 or 科技漢語 *kējì Hànyǔ*, Li, 1985). While linguistic resources on written academic Chinese have witnessed rapid growth in recent years (Tao, 2013; Chen& Tao, 2019; Liu et al., 2016, 2017), efforts in corpus construction and exploitation involving spoken academic Chinese (SAC) have been conspicuously lacking. So far the only spoken academic Chinese corpus reported in the literature is the small-scale pilot project of Han and Liu (2020) developed at Tongji University in Shanghai. Han and Liu (2020) report that the project, which includes 86,395 Chinese characters[1], 25,902 words, and 5,739 prosodic units, is culled mainly from academic lectures, which appear to be mainly in the field of Chinese linguistics.

In this paper, I first describe the construction of an ongoing spoken academic Chinese

[1] However, in the same paper (Han& Liu, 2020: 73), the authors report the corpus to have 100,000 characters.

corpus called the UCLA Corpus of Spoken Academic Chinese (CSAC) and then move on to discuss some of the most salient lexical features, especially keywords, seen in the corpus as currently constructed and in comparison with that of Han and Liu (2020), where possible. Implications for linguistic genre research and the pedagogy of teaching Chinese as a second language (CSL) will also be discussed.

2. The UCLA Corpus of Spoken Academic Chinese and academic lectures as a genre

2.1 Constructing the CSAC database

Construction of dedicated academic Chinese corpora began with written academic Chinese collections, which are built from a wide variety of written academic texts; for example, the UCLA Corpus of Written Academic Chinese (CWAC, Tao, 2013; Chen& Tao, 2019), a 32-million-word collection, includes sources such as journal articles, book chapters, laboratory manuals, course workbooks, and course notes. Specialized written academic corpora often concentrate on a single type of published written text. The academic Chinese corpus compiled at the National Taiwan Normal University, for example, is comprised entirely of journal articles in the humanities and social science disciplines (Liu et al., 2016, 2017). Spoken academic language, on the other hand, has unique parameters that may not figure prominently in written corpus construction. A list of the 152 events in the Michigan Corpus of Academic Spoken English (MICASE, English Language Institute, the University of Michigan 2003: 4-5) shows a fairly comprehensive range of possibilities:

Table 1. Distribution of MICASE speech events

small and large lectures (62)	lab groups and other meetings (6)
public or departmental colloquia (13)	advising consultations (5)
student presentations (11)	dissertation defenses (4)
discussion sections (9)	one-on-one tutorials (3)
seminars (8)	interviews (3)
undergraduate lab sessions (8)	campus/museum tours (2)
office hours (8)	service encounters (2)
study groups (8)	

Clearly, scale (including number of participant, Lee, 2009), relationship between speakers, the nature of academic events, etc., all play a role in shaping the nature and format of oral interactions in academia (Malavska, 2018).

For the UCLA Corpus of Spoken Academic Chinese (CSAC), we have adopted the following design principles that make the corpus unique in comparison with existing (predominantly written) academic Chinese corpora. First, the target data collection is all video-based. Video-based collections provide a window into the multimodal construction of social interaction in general and in spoken academic practices in particular (Khuwaileh, 1999). Multimodal interaction concerns how lexico-grammar, prosody, and visual/bodily semiotic resources are deployed in concerted ways in meaning making (Goodwin, 2000, 2013; Stivers& Sidnell, 2005; Li& Ono, 2019). This is especially important for understanding cognitively highly demanding genres like the academic lecture (Thompson, 1994, 2003). While corpus linguistics does not typically engage in detailed analysis of specific interactional episodes, as conversation analysis and interactional linguistics do (Sacks et al., 1974; Ochs et al., 1996; Couper-Kuhlen& Selting, 2018), I believe it is time for corpus linguistics of the 21st century to seek ways to integrate such methodologies in analysis. Second, all video data should be open source, so that when the corpus is eventually opened to the wider research community, the source data will have no restrictions for public use. Thus all of the video recordings in CSAC are freely accessible from publicly accessible video sites such as YouTube.com, Bilibili.com, open.163.com, qq.com, and v.youku.com. Third, the data collection includes a wide variety of discipline areas in the spoken context. In comparison with the UCLA

Corpus of Written Academic Chinese (CWAC, Tao, 2013; Chen& Tao, 2019), which follows the New Zealand Academic English Corpus (Coxhead, 2000) and is organized in terms of four different broad disciplinary fields -– the arts, commerce, law, and science – each with a number of subfields, the CSAC corpus will likewise include a broad spectrum of subject areas, including some domains that are oriented toward non-specialists, for example, popular science, training in practical skills (such as public speech), and so forth. Finally, in order to balance data variety, avoid skewing of individual style, and facilitate transcription, each spoken sample is restricted to about 30-50 minutes. It is hoped that more video clips of this size will be added to increase the variety and balance among discipline areas, language features, and individual styles.

Transcription of the data was done initially with speech-to-text software tools and subsequently manually checked by undergraduate and graduate team members. The transcription details are minimal, taking the basic breath unit or intonation unit (Du Bois et al., 1993; Tao, 1996) as the unit of transcription. Tokenization and parts of speech annotation were performed with the Stanford Chinese Parser (Levy& Manning, 2003) through a Windows interface called NLP Wrapper[2].

At the moment, while the project is ongoing, the following samples have been transcribed and annotated.

[2] Available from https://github.com/bobatrance/NLPWrapper. Thanks are due to Edwin Tao for developing the NLP Wrapper for this project.

Table 2. Overview of the UCLA Corpus of Spoken Academic Chinese

File Name	Subject	Speech Event	Duration	Tokens	% of Corpus
F12_Achitecture	Architecture	Lecture	43:15:00	5917	3.85
F11_Microbiology	Biology	Lecture	40:06:00	6470	4.21
F01_Finance	Business	Lecture	30:51:00	4645	3.02
F04_National credit system	Business	Lecture	32:50:00	4844	3.15
S08_Effective management	Business	Lecture	48:19:00	9039	5.88
F08_Bio Chemistry	Chemistry	Lecture	32:09:00	4851	3.15
F10_Computer structure	Computer Science	Lecture	30:21:00	4483	2.91
F22_Computer sci. oral defense	Computer Science	Defense	5:10:00	933	0.61
F03_Monetary systems	Economics	Lecture	34:45:00	5623	3.65
F15_Economics	Economics	Lecture	42:16:00	6022	3.91
S04_Economics principles	Economics	Lecture	34:47:00	6596	4.29
F07_Electric circuits	Electrical Engineering	Lecture	37:36:00	7056	4.59
F05_Speech art	Language	Lecture	31:49:00	5111	3.32
F06_Language acquisition	Language	Presentation	24:50:00	3625	2.36
F21_Applied ling contest	Language	Presentation	35:19:00	6542	4.25
S01_Marketing regulations	Law	Lecture	36:01:00	4822	3.13
F20_Literature teaching contest	Literature	Contest	10:05:00	1139	0.74
S05_Software markets	Marketing	Lecture	52:21:00	9077	5.90
F14_Materials sci.	Material Science	Lecture	38:10:00	4491	2.92
F13_Linear algebra and design	Mathematics	Lecture	39:51:00	6211	4.04
F16_Student mental health	Mental Health	Lecture	33:26:00	4932	3.21
F17_Confucius and the Analects	Philosophy	Lecture	39:07:00	4577	2.98
F18_Wang Yangming	Philosophy	Lecture	35:51:00	4917	3.20
F02_Physics and the arts	Physics	Lecture	35:10:00	4009	2.61
F09_Physics	Physics	Lecture	37:47:00	3582	2.33
F19 Physics teaching contest	Physics	Contest	10:01:00	1132	0.74
S02_Intro to Psychology	Psychology	Lecture	41:16:00	4874	3.17
S07_EQ	Psychology	Lecture	36:15:00	5567	3.62
S06_Intro to Physiology	Physiology	Lecture	36:09:00	5805	3.77
S03_Speech contest intro	Speech Art	Lecture	41:57:00	6954	4.52
Total (17.125 hours) /Average			**34:15:40**	**153,846**	**100**

As Table 2 shows, over 17 hours of video recordings in total are currently in the collection, with an average of 34:15 minutes per recording. There are 30 recordings spanning 19 discipline areas. 26 out of the 30 recordings (87%) are university classroom lectures, with a small number of conference presentations and teaching contest presentations. The overall size of the corpus is 153,846 words, with 10,632 word types, rounding up to a very low type-token ratio of 0.069. In the next section, we will explore in more detail some of the most salient lexical features of spoken academic Chinese based on the corpus.

2.2 Academic lectures as a genre

Given that university classroom lectures make up the majority of the current collection, a word about this genre is in order. There has been a great deal of research on genre analysis in general and academic language genres in particular. Malavska (2018) provides a comprehensive overview of research on the genre of academic lectures, where an expert lecturer engages in conveying knowledge to a large number of students in the audience. According to her, the academic lecture can be characterized as a "secondary spoken genre", as it mixes spoken language features with written and multimedia elements, with various degrees of plannedness (Ochs, 1979; Flowerdew& Miller, 1997). From the point of view of a discourse community (Swales, 1990), in the academic lecture setting, the instructor and students forge a social community with a shared goal of delivering "value-laden discourses in which lecturers not only present information to the audience, but also express their attitude and evaluation of the materials" (Lee, 2009: 43, citing Thompson, 1994). While interactivity between the instructor and the audience may not be individualized or one-to-one, especially in large lecture settings (Lee, 2009), involvement (Chafe, 1982), participation, and interactivities can be assumed to function at various levels and at different portions of the discourse process. Malavska (2018: 67) also outlines a few expected features of academic speech: it is logical and consistent, systematic and clear, objective, and intellectually expressive. We will see later that all of these key elements identified in the literature for the genre provide useful perspectives from which to understand the major lexical features found in the CSAC corpus.

3. Initial observations of lexical features of spoken academic Chinese

3.1. Word frequency list

A quick way to understand the lexical features of any corpus is by generating a frequency list of words found in the corpus. The top 50 words in the CSAC corpus are given in Table 3.

Table 3. Top 50 most frequent words in CSAC

Rank	**Token**	**Freq**	17	我	1418	34	和	625
1	的	10871	18	那麼	1356	35	那	590
2	是	6097	19	一	1189	36	可以	579
3	這個	2941	20	他	1162	37	這樣	544
4	我們	2651	21	要	1142	38	去	501
5	就	2597	22	所以	1021	39	上	490
6	了	2145	23	也	915	40	時候	485
7	啊	1980	24	什麼	905	41	好	484
8	一個	1962	25	個	893	42	講	470
9	你	1795	26	都	833	43	把	463
10	有	1749	27	對	827	44	看	463
11	在	1649	28	來	794	45	但是	458
12	它	1632	29	很	726	46	吧	456
13	呢	1593	30	到	708	47	因爲	445
14	這	1521	31	人	652	48	做	444
15	不	1487	32	會	643	49	一些	424
16	說	1430	33	大家	627	50	大	417

Han and Liu (2020) provide a top 100 word list based on their pilot study. Their top 50 words are listed in Table 4.

Table 4. Top 50 high frequency words in Han and Liu (2020).

Rank	**Token**	**Freq**	17	對吧	212	34	可以	120
1	的	1467	18	我	209	35	意思	114
2	啊	885	19	要	197	36	語言	112
3	是	791	20	好	197	37	語境	108
4	呢	584	21	個	193	38	網絡	104
5	了	479	22	在	193	39	沒有	102
6	有	386	23	方言	180	40	語法	101
7	你	369	24	這	167	41	普通話	100
8	我們	356	25	不	160	42	用	100
9	這個	346	26	那	145	43	很	100
10	唉	310	27	都	145	44	理解	92
11	就	302	28	也	140	45	會	89
12	什麼	293	29	是吧	139	46	裏面	86
13	說	285	30	它	131	47	爲什麼	85
14	就是	253	31	講	125	48	問題	84
15	他	226	32	是不是	124	49	下	83
16	那麼	226	33	對不對	123	50	還有	83

Since there are idiosyncratic words likely attributable to the divergent content of the two corpora, we can ignore those for the time being and focus instead on the most frequent 20 words of each corpus for a quick comparison. A glance at the lists shows that 14 of the top 20 words are shared between the two corpora, including:

了　你　啊　我　是　的　這個
他　呢　就　我們　有　說　那麼

Given that common function words such as 的 *de*, 了 *le*, 是 *shì* 'be', 有 *yǒu* 'have, exist', and 就 *jiù* 'just, then' are very common in any Chinese corpus (Tao, 2015) and are thus unlikely to differentiate SAC from other varieties of Chinese, we can again filter those out and focus on the

rest. Among the remaining lexical items, a few lexical categories are noticeable. These include: personal pronouns (我 *wǒ* 'I', 我們 *wǒmen* 'we', 你 *nǐ* 'you', 他 *tā* 'he'), demonstratives (這個 *zhège 'this (one)'*), utterance final particles and/or discourse particles/markers (呢 *ne* and 啊 *a*), conjunctions (那麼 *nàme* 'so, then'), and verbs (說 *shuō* 'say, talk'). Admittedly, these are still very common lexis, yet as we will show later, many of them exhibit important properties in the context of spoken academic Chinese, and academic lectures in particular, that warrant further analysis. We will return to some of those later.

3.2. Keywords

While word lists (especially from a comparative point of view) can yield useful initial insights into some lexical features, given the widely recognized issue that simple frequency counts can be heavily impacted by high frequency lexis, it is important to use keyword lists to reveal critical lexical features of the text (Scott& Trebble, 2006; Xiao& McEnery, 2005; O'Keeffe et al., 2007: 208). Keyword as a corpus analysis tool has been widely used in corpus-based linguistic and literary analysis (Culpeper, 2009), but recent research has highlighted various ways to improve the use of keyword (and key keyword) as a method of analysis (Egbert& Biber, 2019; Gries, 2021), chiefly in regard to the issue of dispersion (roughly the range of distribution of keywords across texts). For this study, however, since I am mainly concerned with core lexical features rather than content features – a main concern of many (key) keyword projects, dispersion induced issues tend not to figure prominently here as the core lexical items we are looking at generally distribute widely across text types (and parts of the corpus), which are often empirically verified.

On the other hand, as a number of researchers have pointed out, for keyword analysis, selection of the benchmark (also known as the reference) corpus is critical (O'Keeffe et al., 2007: 208; Culpeper, 2009). Culpeper (2009), citing Enkvist's extreme example of comparing a poem with a telephone directory (p. 34), argued convincingly that it is important to compare texts that are as compatible as possible. In analyzing the spoken academic English features, O'Keeffe et al. (2007) use a natural conversation corpus as the benchmark. Given that spoken academic language is hybrid (Malavska, 2018), cutting across at least two domains: spoken language and (written and multimedia) academic text (Flowerdew& Miller, 1997), there may be multiple ways of choosing

the appropriate kinds of reference corpora for an optimal set of keywords for deeper understanding of the text. For this reason, I have decided to conduct three sets of keywords analysis, based on 1) ordinary conversation as the benchmark, 2) written academic Chinese as the benchmark, and 3) both ordinary conversation and written academic Chinese combined as the benchmark. The everyday conversation corpus I used has over 500 recordings of talk conducted between family, friends, and acquaintances, and 1.7 million words, while the written academic Chinese corpus (Tao, 2013; Chen& Tao, 2019) has 5.4 million words from multiple disciplines as described earlier. When those two datasets are combined, the total database size is 7.1 million words.

The following are the results of the top 20 items based on these comparisons using the Keyword List tool of AntConc (Anthony, 2020), with relevant settings adjusted as follows: 1) Keyword statistics: Log-Likelihood (4-items); Keyword statistics threshold: $p < 0.01$ (+Bonferroni); 3) Keyword Effect Size Measure: Dice coefficient; and Keyword effect size threshold: top 100.[3]

A quick examination of the three sets of results represented in Table 5 shows that, with ordinary conversation as the benchmark, some of the content words such as 演講 *yǎnjiǎng* 'speech', 貨幣 *huòbì* 'currency', 過程 *guòchéng* 'process', and 功能 *gōngnéng* 'function', as well as a few typical written language markers such as 進行 *jìnxíng* 'get on, be engaged in', are foregrounded. By contrast, when the written academic corpus (CWAC) is used as the benchmark, none of the content words are identified while nearly all of the resultant keywords appear to be typical spoken language elements (with the number one item being the utterance-final particle and/or discourse particle 啊 *a*). The most unique markers in both columns are marked in boldface in Table 5. However, when both ordinary conversation and the written academic data are combined (the 7.1-million-word combo corpus) and used as the benchmark, we see a reduction of extreme types in both directions. For example, for written and content items, we can see just a few items such as 演講 *yǎnjiǎng* 'speech', 講 *jiǎng* 'talk', 當中 *dāngzhōng* 'be in midst of' etc. are now listed; the rest of the items are either common spoken items or all-around high frequency items, with no excessive or exclusive spoken forms represented. Based on this observation, my

[3] Information on some of these statistical measures can be found at http://corpora.lancs.ac.uk/sigtest/#extraHelp.

conclusion is that the third option ("with both" in Table 5) seems to yield the most balanced (or least biased) list of the core lexical items (or key keywords) for the spoken academic language.

Table 5. Top 20 keywords based on three benchmarks

Rank	**w/ Spoken**	**w/CWAC**	**w/Both**
1	的	**啊**	這個
2	這個	這個	我們
3	是	我們	那麼
4	我們	**呢**	呢
5	那麼	你	是
6	一個	那麼	它
7	它	我	一個
8	大家	就	大家
9	呢	是	啊
10	所以	它	所以
11	當中	大家	就
12	**演講**	一個	演講
13	**貨幣**	說	當中
14	和	所以	這
15	地	什麼	說
16	重要	**吧**	要
17	**過程**	他	講
18	**功能**	那	什麼
19	**進行**	**呃**	叫
20	非常	的話	非常

Thus it is important to take a closer look at the keywords identified through the combined benchmark corpus.

4. Keywords in CSAC

Before beginning keyword analysis, it is useful first to note that some of the top words identified purely on the basis of frequency of occurrence fall out of the keyword list. Recall that

in Section 3.1 we have seen pronominals (e.g. *wǒ* 'I', *wǒmen* 'we', *nǐ* 'you', *tā* 'it'), demonstratives (e.g. *zhège 'this (one)'*), utterance-final particles/discourse particles (e.g. *ne* and *a*), conjunctions (e.g. *nàme* 'so, then'), and verbs (e.g. *shuō* 'say, talk') as highly frequent. However, of the remaining top 20 keyword list, only some make the list, including *wǒmen* 'we', *zhège 'this (one)'*, the particle *a*, and *nàme* 'so, then'. In the analysis to be presented next, then, I will concentrate on these four items and the respective lexical categories they represent.

4.1. Pronominals[4]

The prominent role that pronominals such as first and second person pronouns *wǒ* 'I' and *nǐ* 'you' play in spoken language is well documented (Tao, 2015), for good reasons. In ordinary conversational contexts, most of the time we talk about our feelings, views, and opinions; we interact with one another in direct ways (Scheibman, 2002; Kärkkäinen, 2003; Thompson& Hopper, 2001); and we index our epistemic and affective stances with marked agents, even when the language used is the so-called zero-anaphora language (Tao, 1996: chpt. 7), hence the high frequency of first and second person pronouns observed in everyday talk. However, in the case of spoken academic language such as the university lecture, the primary identity is an academic community (Swales, 1990), and the primary communicative goal is to transmit knowledge and forge intersubjectivity, in the sense of approaching issues together and acting together, in hopes of expert and novice learners reaching a common understanding. For these reasons, it is not surprising to see the downplaying of the role of the individual and the elevation of the collective identity, which is most directly represented by the use of the first person plural pronoun. Some representative uses of 我們 can be found in extracts (1)-(3).

[4] The role of the inanimate third person 它 *ta* is also a noticeable feature of CSAC. This is likely because the subject matters discussed in the corpus tend to be inanimate objects, and it may also be used as a tracking device for discourse entities. However, I will leave this topic to future studies and focus instead on personal pronouns.

(1) 所以今天啊，**我們**就一塊來研究一下，波的一種特性，叫作波的干涉。
Suǒyǐ jīntiān a, wǒmen jiù yīkuài lái yánjiū yīxià, bō de yī zhǒng tèxìng, jiào zuò bō de gānshè.
'So today, let's take a look at this together, a special characteristics of waves, called the interference of waves.'

In this example, the instructor explicitly calls for the audience to work together with her in the study of the main topic, wave interference. By contrast, in extract (2), the reference of *women* 'we' is actually the instructor himself (speaker-*we*), as he is the one who is in control of introducing the topics of the lesson.

(2) 還有，有很多看待電路的觀點，**我們**一一會再來給大家做介紹。
Hái yǒu, yǒu hěnduō kàndài diànlù de guāndiǎn, wǒmen yīyī huì zàilái gěi dàjiā zuò jièshào.
'In addition, there are many views on circuits, and we will introduce them (to you) later.'

However, this does not mean that singular pronominals are not used in academic lectures; in fact, they are also quite frequently used (Yeo& Ting, 2014). When they are used, however, the goal is not to highlight individual identity or for identification but rather for forging some shared identity or for seeking intersubjectivity. In (3), we can see clusters of the second person pronoun *nǐ* 'you', and the two first person pronouns *wǒ* 'I' and *wǒmen* 'we':

(3) 從時間上就可以知道**你**是經營大還是管理大。所以**我們**一定要很清楚地知道，**我們**在做這個管理的時候，**我們**雖然定義很瞭解，但是**你**在觀念上，**我**還是希望**你**能夠調整過來。
Cóng shíjiān shàng jiù kěyǐ zhīdào nǐ shì jīngyíng dà háishì guǎnlǐ dà. Suǒyǐ wǒmen yídìng yào hěn qīngchǔ de zhīdào, wǒmen zài zuò zhège guǎnlǐ de shíhòu, wǒmen suīrán dìngyì hěn liǎojiě, dànshì nǐ zài guānniàn shàng, wǒ háishì xīwàng nǐ nénggòu tiáozhěng guòlái.
'You can tell in terms of time whether you are big on deals or big on managing. So we must know very clearly that when we are doing this management, although we know the definition very well, in terms of your concept, I still hope you can adjust it.'

All the instances of *nǐ* 'you' in this example are the so-called generic audience-*you* (Yeo& Ting, 2014), rather than any person in particular (in the audience), as would be the case in most ordinary conversations, and are for collective audience involvement (Chafe, 1982). The cases of *women* can be deemed the inclusive-*we*, referring on the surface to both the audience and the instructor. However, from a semantic and pragmatic point of view, the referential meaning leans heavily toward the audience, as the instructor warns those who only know the definition of management or managing without understanding what management is really about ('we must know very clearly that when we are doing this management, although we know the definition very well') – a proposition that is more identifiable with the novice learner in the audience than with the instructor. In the end, the instructor differentiates the first person from the second person ('in terms of your concept, I still hope you can adjust it'), making the distinction between those in the know and those not more explicit. From this analysis, we can see that although the shifts between first and second person and between the two forms of the first person seem chaotic, there is actually some regularity: *women* 'we' can be used as a way to evoke the notion of instructor and audience togetherness when it comes to negative knowledge states, and once the instructor is also identified as vulnerable to a negative knowledge state, the instructor then transitions to providing explicit lessons on how the novice learner can overcome the hurdle in question, without being perceived as overly imposing or looking down on the students. In short, while the collective identity is most straightforwardly expressed through the use of the plural form *women* 'we', singular pronouns also help achieve intersubjectivity in ways different from their most canonical use.

Moving on to the other prominent second person pronoun on the keyword list, 大家 *dàjiā* '(you) all', we may say that this pronominal form evokes a sense of mass, collective audience identity (audience (all)-*you*). According to Chao (1968: 648), *dàjiā* and similar pronouns denote the meaning of 'all present' or 'all concerned'. Similarly, Zhu (1982: 6.7) characterizes *dàjiā* simply as having a mass designation. In CSAC, however, the referential (or designation) function of this form is downplayed as it is commonly used in directives where the instructor asks the audience explicitly for joint attention and/or to invite the audience's participation in certain cognitive activities. Three examples represented in (4) - (6) can illustrate these uses.

(4) 請**大家**注意，我問的是企業家這些人。
Qǐng dàjiā zhùyì, wǒ wèn de shì qǐyè jiā zhèxiē rén.
'Please note that I am asking about entrepreneurs.'

(5) 那麼什麼樣的人，群體會依隨他呢？品德高尚的人。所以，提拔，**大家**記住，德比才重要。
Nàme shénme yàng de rén, qúntǐ huì yī suí tā ne? Pǐndé gāoshàng de rén. Suǒyǐ, tíbá, dàjiā jì zhù, débǐ cái zhòngyào.
'So what kind of person, the group will follow him? A person of high moral character. Therefore, (for) promotions, everyone remember, morality is of the utmost importance.'

(6) 我們同濟大學，曾經有一位著名的教授叫陳從周**大家**知道嗎？
Wǒmen tóngjì dàxué, céngjīng yǒu yīwèi zhùmíng de jiàoshòu jiào Chén Cóngzhōu dàjiā zhīdào ma?
'At our Tongji University, there used to be a famous professor named Chen Congzhou, do you all know?'

In (4), the instructor uses a combination of *dàjiā* and *zhùyì*, 'attention', to call the audience's attention to the subject that he is checking with them about. In (5) the instructor asks everyone to remember (*jìzhù*) an important principle in the promotion of people to important positions. Finally, example (6) shows an indirect way to get the audience's involvement (Chafe, 1982), as the turn is designed in a question form with *zhīdào* 'know' (Tao, 2003).

In sum, most of the personal pronouns in the spoken academic lecture context are group or academic community oriented. This gives rise to the all-around prominence of the first person plural pronominal form *women* 'we', which has a variety of inclusive uses: expressing a strong sense of instructor and audience togetherness in terms of sharing (negative) knowledge states, reaching common ground together, performing actions together, and planning on courses of actions together. Individual personal pronouns, such as *wǒ* 'I' and *nǐ* 'you' do get used, but they are mostly deployed in the service of forging shared identities in various senses rather than for individual identities or identification. Finally, *dàjiā*, an exclusive audience (all)-*you* form, is used to explicitly draw audience attention and/or to issue an invitation for participating in joint cognitive activities along with the instructor.

4.2. Demonstratives

The two most prominent demonstratives on the top 20 keyword list of the spoken academic language corpus are both proximal tokens: 這個 *zhège* (which may also be pronounced as *zhèige* in Beijing Mandarin; however, *zhège* will be used here throughout) and 這 *zhè* (or *zhèi)*, with *zhège* being the very top keyword for CSAC. At first this may look unremarkable, given that proximal demonstratives are consistently shown to be frequent across the board, and they are overwhelmingly more frequent than distal ones (Xu, 1988; Tao, 1999a); however, the fact that these tokens sit at the very top of the list still demands our attention. Existing literature on *zhège* has identified a number of common uses in everyday language. Most importantly, demonstratives are shown to have developed pragmatic uses other than their spatial denotations, which include textual and social meanings (e.g. proximal demonstratives encoding more empathy than their distal counterparts, Tao, 1999a), definiteness (Huang, 1999; Fang, 2002; P. Chen, 2004), and discourse marker use (Liu, 2009). Liu (2009) examined the cohesive use of *zhège* and noted its discourse forward linking (cataphoric) function (as opposed to backward linking by the distal *nà(i)ge*), topic marking function, as well as its cause-introducing function. Furthermore, Liu noted the social correlations of these demonstratives: the proximal forms are said to be more likely to be used by a speaker with higher social status (e.g. senior person, teacher, etc.) than someone of a lower social status. These properties seem to be in congruence with the overall high frequency of *zhège* and *zhè* in the CSAC data, since the speakers in our collection are mostly teachers, who, presumably, have a higher status than those in the audience at the moment of lecturing. However, careful analysis of the data is still needed in order to better understand how proximal demonstratives actually work in the spoken academic Chinese context.

Using the AntConc Concordance tool, a randomly selected set of 148 cases of *zhège* shows that 104 of them (70%) are used as a modifier, i.e. in the attributive slot before a nominal, while 44 (30%) are used as an independent token (i.e. without a head noun). These two types of use are exemplified in (7) and (8) respectively.

(7) 實際上是影響了整個歐洲的**這個**貨幣，啊影響歐元穩定。
Shíjì shang shì yǐngxiǎngle zhěnggè ōuzhōu de zhège huòbì, a yǐngxiǎng ōuyuán wěndìng.
'In reality, it affects the entire European currency, ah, affects the stability of the euro.'

(8) 是以牟利爲目的的，那麼**這個**沒有利潤他能維持經營嗎？
Shì yǐ móulì wéi mùdì de, nàme zhège méiyǒu lìrùn tā néng wéichí jīngyíng ma?
'It is for the purpose of making a profit, so can he maintain the business if there is no profit?'

In (7) *zhège* modifies 貨幣 *huòbì* 'currency', whereas in (8) *zhège* is used alone, referring back to the antecedent 牟利爲目的 *móulì wéi mùdì* 'purpose of making a profit'.

The fact that 70% of the proximal demonstratives are used in an attributive role suggests that their referential function is important in the spoken academic genre. An examination of the contexts in which they are used shows that they play what can be called a double role: tracking a previously introduced referent and marking definiteness. Below are some examples illustrating these patterns.

(9) 我們把它叫做排毒作用，抗腫瘤抗衰老，那麼**這個**作用呢實質上…
Wǒmen bǎ tā jiàozuò páidú zuòyòng, kàng zhǒngliú kàng shuāilǎo, nàme zhège zuòyòng ne shízhì shàng
'We call it detoxification, anti-tumor and anti-aging, then this effect is essentially…'

In (9), after the detoxification (排毒 *páidú*) function is introduced, it is immediately referred back to as *zhège* 'this function'. 作品 *Zuòpǐn* 'work' in (10) is marked by *zhège* in a similar manner, only with one more clause in between the two mentions.

(10) 最後給大家展現一幅抽象主義畫派，畫家克利的**作品**。克利被稱之爲教授型的畫家。你看**這個作品**，非常神秘。
Zuìhòu gěi dàjiā zhǎnxiàn yì fú chōuxiàng zhǔyì huà pài, huàjiā kèlì de zuòpǐn. Kèlì bèi chēng zhī wéi jiàoshòu xíng de huàjiā. Nǐ kàn zhège zuòpǐn, fēicháng shénmì.
'Finally, I will show you an abstract painting school, the work of the painter Klee. Klee is called a scholar painter. Take a look at this work; it's rather mysterious.'

In (11), what is referred to as 這麼個事 *zhème gè shì* "this incident" has been mentioned over ten

clauses prior: the discovery of a big cache of Confucius antique documents from the resident of the Lord of Lugong during the Western Han period. By using the term 事 *shì* 'thing, incident', the speaker is able to keep track of the event throughout the subsequent discussion.

(11) 那麼因爲有**這麼個事**，所以好事之徒那些爲了這個。使自己出名的。水平也還蠻不錯的，這麼一撥人，然後就開始借著**這個事**呢製造僞書。
Nàme yīnwèi yǒu zhème gè shì, suǒyǐ hǎoshì zhī tú nàxiē wèile zhège. Shǐ zìjǐ chūmíng de. Shuǐpíng yě hái mán búcuò de, zhème yī bō rén, ránhòu jiù kāishǐ jièzhe zhège shì ne zhìzào wěishū
'So, because of such an incident, those who are ambitious do it for this, to make themselves famous. Their standards is pretty high. With such a group of people, they began to make fake books through (by taking advantage of) this matter.'

Interestingly, the hand gestures depicting the two references to 'the incident' in extract (11) are constructed with very similar shapes, and the speaker's postures while producing them are also remarkably close, as shown in Figures 1 and 2. In other words, both the lexical expressions and their affiliated gestures work together in keeping track of an introduced entity in subsequent discussions.

Fig. 1. 因爲有這麼個事 (12:34)
Yīnwèi yǒu zhème gè shì 'Because of this incident'

Fig. 2. 借著這個事 (12:44)
Jièzhe zhège shì ‘through this incident’

What these examples show is that while existing views on the proximal demonstrative *zhège* offer a great deal of useful insight into their uses in discourse, the hybrid nature of the uses seen here has hitherto not been well documented. However, I believe that such a double-role use is well motivated: in classroom lectures, there are a great deal of references to be tracked, and lecturers need to present the content in systematic, cohesive, and clear ways, a major challenge for both the lecturer and the audience (Thompson, 1994, 2003; Malavska, 2018). With the lack of grammaticalized definiteness markers such as the definite article *the* in English, the need for *zhè(me)ge* to fill this role is great (P. Chen, 2004). Thus it is not surprising to see both the dominance of the attributive use (as opposed to the independent use) and the statistically identified high level keyness of these tokens in the CSAC corpus.

The simplex demonstrative form *zhè* serves a similar role to the composite form *zhège* (comprising *zhè* plus the classifier *ge*), with the difference being that *zhè* can be combined with other numeral classifier expressions (e.g. 三種 *sān zhǒng* ‘three kinds’, 幾個 *jǐ gè* ‘several’, and so forth) besides the individuation form of *ge*, as shown in (12) and (13).

(12) 切的結果有三種，注意**這三種**結果將來你們一旦做試驗可能都會涉及到。
Qiè de jiéguǒ yǒusān zhǒng, zhùyì zhè sān zhǒng jiéguǒ jiānglái nǐmen yīdàn zuò shìyàn kěnéng doūhuì shèjí dào.
‘There are three results of cutting. Note that these three results may be relevant once you start doing experiments in the future.’

(13) 最後是通脹和失業。**這幾個**概念都是宏觀經濟學的概念，
Zuìhòu shì tōngzhàng hé shīyè. Zhè jǐ gè gàiniàn dōu shì hóngguān jīngjì xué de gàiniàn,
'Finally, inflation and unemployment. These concepts are all concepts of macroeconomics,'

An interesting case of *zhège* can be found in (14), where *zhège* seems to have taken the place of *zhè*:

(14) 那 I triple E 呢，嚴格地給出來了需求的五個基本性質。也就是說，具有了**這個**五個基本性質的描述，它才可以稱爲需求。
Nà I triple E ne, yángé de gěi chūláile xūqiú de wǔ gè jīběn xìngzhì. Yě jiùshì shuō, jùyǒule zhège wǔ gè jīběn xìngzhì de miáoshù, tā cái kěyǐ chēng wéi xūqiú.
'So IEEE provides a strict definition of the five basic properties of demand. In other words, only with the description of these five basic properties can it be called a demand.'

In this example, 五個基本性質 *wǔ gè jīběn xìngzhì* 'five basic properties' could have been grammatically referred to with the simplex form *zhè 'this'*; instead, here the composite form *zhège* is used, attesting to the highly grammaticalized status of the *zhè* and *ge* combination.

In summary, the proximal demonstratives in the spoken academic context play the dual role of tracking an introduced referent and marking definiteness. This finding reveals new properties that contrast with previous studies in multiple ways. First, while earlier studies have shown that some demonstratives are moving toward becoming a definiteness or discourse marker, they attribute this use mostly to either the distal form (Huang, 1999) or the independent proximal forms (Fang, 2002; Liu, 2009). On the other hand, while other studies have highlighted the pragmatic aspects of the referential use, they often associate this with a forward linking function (Liu, 2009; Yin, 2009), which is not entirely true for the CSAC data, where backward tracking is quite common. Spoken academic language thus offers a window into an alternative context in which (proximal) demonstratives work.

4.3. Particles

Two particles on the CSAC keyword list stand out: 啊 *a* and 呢 *ne*. As a highly frequent lexical class in spoken language, particles have received extensive treatment in the literature, with most attention paid to utterance-final particles in spoken Chinese (Chao, 1968; Lee-Wong, 1998, 2001; Chu, 1998, 2002, 2015). Some of the tokens, such as *a*, can also appear in non-final positions, in which case they are often called interjections in standard reference grammars (Chao, 1968: 815), discourse particles (Tseng, 2001) or even discourse markers (Lin, 2003) in more recent studies. In reference grammars, interjections are often said to express a wide range of meanings, depending on form and context. *A*, for example, is said to "express mild feelings" as its English counterpart *ah* and for "repeated requests" (Chao, 1968:817). However, in the context of spoken academic Chinese, its uses are found to have a much wider scope than those described in reference grammars.

First, in terms of the status of *a* as a particle of any kind, more detailed prosodic analysis will be needed. Earlier studies on the utterance-final *a* have generally divided the token into two types: a so-called strong version and weak version. However, the identification of these versions is not free of controversy (Chu, 2002; Lin, 2003). In addition, previous prosodic analyses have not taken into account the fuzzy nature of the positions of the *a* token. An examination of the prosodic shapes of *a* in academic lectures shows that there are potentially at least three distinct forms that can be identified: utterance final, utterance initial, and free standing, and sometimes it is difficult to tell where the boundaries are due to the weak prosodic forms it sometimes takes in spontaneous speech. This is illustrated in extract (15), where the numbers in the parentheses indicate, through measurements in Praat (Boersma& Weenink, 2021), the estimated duration (in tenths of a second) of the vocalic syllable. Due to the functions of these varied tokens, the rough transliterations also vary, from *ah* to *uh* or *uhm*.

(15) …( ) **啊(.23)**，那麼什麼是電路呢？**啊(.13)**先要給一個解釋。**啊(.2)**電路英文叫 electric circuits，那麼也有說法叫 electrical, **啊(.19)**，我們采用了這個更常用的 electric circuits，由元件，**啊(.14)**若干的這個 elements, 相互連接構成的電流的通路,**啊(.13)**。

A(.23), Nàme shénme shì diànlù ne? A(.13) Xiān yào gěi yígè jiěshì. A(.2) Diànlù yīngwén jiào electric circuits, nàme yěyǒu shuōfǎ jiào electrical, a(.19), Wǒmen cǎiyòngle zhège gèng chángyòng de electric circuits, yóu yuánjiàn, a(.14) Ruògān de zhège elements, xiānghù liánjiē gòuchéng de diànliú de tōnglù, a(.13).

'Ah (.23), so what is a circuit? Uh (.13) I need to give an explanation first. uh (.2) The circuit is called *electric circuits* in English, so it is also called electrical, uhm (.19), we use this more commonly used (term) electric circuits, which are composed of components, uh (.14) several of these elements, connected to each other, forming the path of current, uh (.13).'

In this extract, there are six tokens of *a* observed. Some are clearly standing alone, as in the cases of the first and the last (sixth), others may not be so clear cut. If they appear at the end of an utterance which has a falling or terminal intonation (Du Bois et al., 1993), one is more likely to treat the *a* token as an utterance initial element. However, if the previous intonation unit is continuing or non-terminal, and the *a* token is produced in short form (e.g. less than 0.2 second), it could go either way: the end of the previous utterance (as a final particle) or the beginning of the second unit (as a unit initial particle), as in the cases of the fourth and fifth tokens in the extract. In other words, the status of the *a* tokens is at best a continuum. The Praat produced graph in Figure 3 displays just the first three tokens of *a*, which exhibit three different shapes: a large token in the clear in 1); 2) a small token closely attached to the following utterance; and an intermediate token in 3), likely attaching to the unit that follows.

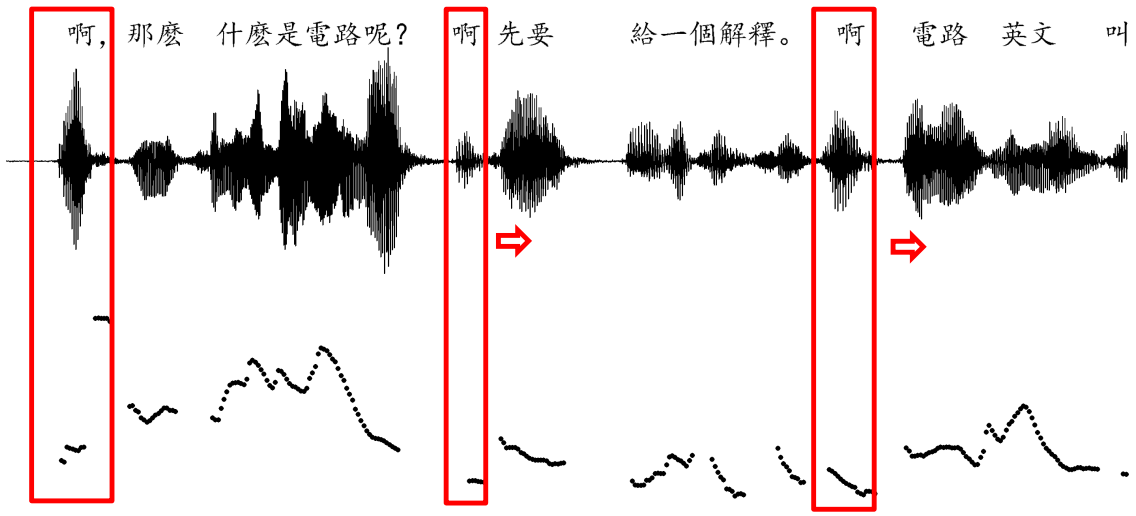

Fig. 3. Diverse prosodic shapes of the first three tokens of *a* in extract (15).

Second, in terms of functions, some earlier studies based on natural conversation have made a distinction between speaker-oriented (i.e. expressing more subjective attachment to the associated message) and addressee-oriented (agreeing or appealing to the addressee, Chu, 2002) uses. Such a distinction may not be completely relevant in the academic lecture setting. Most of the tokens in examples such as (14) can generally be called discourse particles (DP, Tseng, 2001) or discourse markers (DM, Lin, 2003), in the sense of indicating discourse boundaries (as signposts) for discourse organization as well as for drawing the attention of the audience. While there is no consensus as to which term to use or how to define them (Schiffrin, 1987; Fraser, 1990; Tseng, 2001), some evidence can be gathered to argue for a finer distinction between the two, similar to the differences between a global level *discourse marker* and a local level *conjunction* as argued by Georgakopoulou and Goutsos (1998). For example, a discourse marker can be said to be used at major discourse boundaries, whereas a discourse particle may not be, and both can be used to draw the attention of the audience. In this sense, we can treat the DM as a stronger version of the DP. If we follow such a distinction, then, particle number one is clearly a DM, as it starts a new segment (about how to define electric circuit) after a long pause, and the rest are used in the middle of their respective discourse segments, marking minor boundaries. Their phonological shapes also vary iconically: major boundary markers exhibit more prominent shapes in terms of longer duration and longer gaps before and after the token, whereas minor ones are less likely to be expressed in this

way (cf. Thompson, 1994, 2003 on similar properties in academic English).

The particle *ne* is exclusively used at the utterance-final position. Chao (1968: 802) analyzes *ne* as functioning to "question with a specific point", to mark a "deliberate pause", etc., while Li and Thompson (1984) treat it as a response token, one that is not used to initiate discourse (p. 306). However, in the CSAC data, *ne* is found to be used mainly for drawing the attention of the audience (Lee-Wong, 2001) or appeal to the listener's active participation (Alleton, 1981, cited in Chu, 1998: 160) through interrogative, semi-interrogative (borderline), or non-interrogative forms. Explicit interrogative or semi-interrogative forms can be found in (16), while non-interrogative forms are presented in (17).

(16) (Semi)interrogative *ne*

a) 哪樣的模型電路會、會有這樣的結果呢，你猜一猜。
Nǎyàng de móxíng diànlù huì, huì yǒu zhèyàng de jiéguǒ ne, nǐ cāi yī cāi.
'Which model circuit will, will have such a result? Can you guess?'

b) 這個媒介是什麼呢？
Zhège méijiè shì shénme ne?
'What is this medium?'

c) 有什麼作用呢？
Yǒu shénme zuòyòng ne?
'What effects does it have?'

d) 意思是什麼呢？
Yìsi shì shénme ne?
'What does it mean?'

(17) Non-interrogative *ne*

a) 那這裏面呢，也是作爲一個什麼，設計的約束。
Nà zhè lǐmiàn ne, yěshì zuòwéi yígè shénme, shèjì de yuēshù.
'What about it, it is also a constraint of design.'

b) 那麼日本呢，也是數量上來看呢，略有下降，呃人體的觀察呢，
Nàme rìběn ne, yěshì shùliàng shàng lái kàn ne, lüè yǒu xiàjiàng, è réntǐ de guānchá ne,
'But, in Japan, from a quantitative point of view, there has been a slight decrease, er, the observation of the human body,'

c) 一個呢，叫南方模式動物中心，那麼內毒素呢？
Yígè ne, jiào nánfāng móshì dòngwù zhōngxīn, nàme nèidúsù ne?
'One is called the Southern Model Animal Center. What about endotoxins?'
d) 那麼我們在她的領導下，我們一起呢，就研製出來了重組的鏈激酶，
Nàme wǒmen zài tā de lǐngdǎo xià, wǒmen yìqǐ ne, jiù yánzhì chūláile chóngzǔ de liànjīméi,
'So we, under her leadership, together we developed a recombinant streptokinase,'
e) 所以呢，手很巧手很巧，動手能力很强啊，
Suǒyǐ ne, shǒu hěn qiǎo shǒu hěn qiǎo, dòngshǒu nénglì hěn qiáng a,
'So, the hands are very skillful, the hands are very skillful, with strong manual skills,'
f) 然後呢，在細胞層面，在動物層面，都可以去研究。
Ránhòu ne, zài xìbāo céngmiàn, zài dòngwù céngmiàn, dōu kěyǐ qù yánjiū.
'Then, at the cellular level and at the animal level, you can study both.'

In both types of use, a pause plus the particle combined serve to draw the attention of the audience to some key elements of the lecture or to a particular point in a sequence of events, and the interrogative form, which plays a fundamental role in learning (Camiciottoli, 2008), can be especially helpful for getting the audience involved, whether or not they actually answer any questions posed by the instructor.

The variety of materials that *ne* can be attached to varies greatly, often indicating the formulaic character of the utterance. Many of the formulas involve a conjunction of some kind, including 但是 *dànshì* 'but, however', 首先 *shǒuxiān* 'firstly', 然後 *ránhòu* 'then', 接下來 *jiē xiàlái* 'then', 下面 *xiàmiàn* 'next', 第 N 個 *dì* N *gè* 'No. N', 另外 *lìngwài* 'besides', 同時 *tóngshí* 'at the same time', 因此 *yīncǐ* 'thus', 所以 *suǒyǐ* 'so', 而且 *érqiě* 'furthermore', etc. We will touch on the issue of conjunctions in the next section.

4.4. Conjunctions

Our last prominent lexical category on the keywords list is conjunction, where 那麼 *nàme* 'so, then' and 所以 *suǒyǐ* 'so' both make the top list. *Nàme* is found to be overwhelmingly used when a previous segment has come to a closure, as exemplified in (18).

(18) 有的國家長期保持逆差，有的國家長期保持順差，都很難調整。**那麼**我們在後面的這個均衡這部分，我們會給大家呢，再進一步討論，爲什麼順差，長期順差也不好。

Yǒu de guójiā chángqī bǎochí nìchā, yǒu de guójiā chángqī bǎochí shùnchā, dōu hěn nán tiáozhěng. Nàme wǒmen zài hòumiàn de zhège jūnhéng zhè bùfèn, wǒmen huì gěi dàjiā ne, zài jìnyībù tǎolùn, wèishéme shùnchā, chángqī shùnchā yě bù hǎo.

'Some countries have a long-term deficit, and other countries have a long-term surplus, and both are difficult to change. Then we will discuss in the section on balancing, and we will further discuss why neither the surplus nor the long-term surplus are good.'

In this example, the first three clauses lay out the two extreme cases of deficit and surplus, after which *nàme* is used to mark the beginning of a new segment and proffers what will be brought up next.

The use of *suǒyǐ* is also noticeable in a number of ways. First, there is no causal marking (e.g. with *yīnwèi*, Song& Tao, 2009) to go with *suǒyǐ* as prescribed in the typical apodosis and protasis formation commonly described in reference grammars. Nearly 90 percent of the time, the *suǒyǐ*-prefixed expressions in CSAC are used after a terminal intonation in the prior clause, indicating the independent status of the *suǒyǐ*-prefaced expression from the preceding clause or clause nexus. In dialogic discourse, *suǒyǐ*-prefaced utterances have been argued to manage suspension and help steer the talk to a pre-prior course of action (Wang, 2020). In the lecture context, however, the patterns are often different: most of the *suǒyǐ*-prefaced utterances can be seen as moving the discourse forward without returning to a pre-prior course of action. This is illustrated in (19).

(19) 那麼我的幻燈一般重要的概念我給大家是用英文，**所以**大家呢，可以願意看英文也可以，

Nàme wǒ de huàndēng yībān zhòngyào de gàiniàn wǒ gěi dàjiā shì yòng yīngwén, suǒyǐ dàjiā ne, kěyǐ yuànyì kàn yīngwén yě kěyǐ,

'So I will show you all the important concepts using English in my slides, so you can read English if you like.'

In this context, the function of *suǒyǐ* is more frequently to indicate a sense of inference, with which the audience is instructed to follow a certain course of action suggested by the instructor in the utterances immediately following the conjunction (‘reading the materials in English’ in this case). Such functions are often described as procedural, in the sense that the speaker points recipients “to particular – more or less schematic – frames of interpretation for the utterances hosting such expressions” (Hansen, 2012: 595). In ordinary conversations, *suǒyǐ* and the English counterpart *so* as well as other similar discourse particles are often argued to serve to facilitate participation transition, for example, for transition of turns at talk (Schiffrin, 1987: 217). However, in monologic discourse such as the academic lecture, participation transition is not critically relevant, and boundary marking and invited inferencing (Traugott, 2018) can be said to play a more important role in the use of such tokens.

5. Summary and discussion

In the preceding section, I have discussed four types of selected top keywords and their associated lexical categories – pronominals, proximal demonstratives, discourse particles, and conjunctions – that have been identified as being statistically significant for spoken academic Chinese, noting especially the unique features they display in academic lectures that may differ from their uses in other contexts such as everyday conversations. These key lexical features can be seen as integral parts of a cluster of related properties, which together define spoken academic Chinese, especially university lectures, as a genre. These common features can be summarized as follows: 1) academic community driven identity and distributed cognition; 2) reference and entity tracking; 3) instructor directed joint attention; and 4) boundary marking and invited inferencing. All these features mesh together and are fitted in for a genre that shares some elements with ordinary conversation on the one hand and some with written academic text on the other hand, yet distinguishes itself from both with its own set of characteristics based on and driven by its unique communicative goals.

To illustrate these features in a more holistic way, let us examine our final example (20) where all of the discussed major features manifest.

(20) 當有致病菌進來的時候某些致病菌還會被**這個**抗體作用**啊**，阻止它致病，**所以這個呢**叫免疫作用。**那麼**第三個**呢**是營養作用,這個**大家**都很清楚了對吧,**我們**在**這個**新陳代謝裏面也講過，有些細菌它能够合成一些營養物質最常見的大腸杆菌，合成維生素 k 哦, **那麼**最後**呢**,還有些其他的作用,包括**我們**把它叫做排毒作用，抗腫瘤抗衰老, 那麼**這個**作用呢實質上**我們**現在看到能够抗腫瘤抗衰老能够排毒，實質上**這個**作用都是間接的，

Dāng yǒu zhì bìngjùn jìnlái de shíhòu mǒu xiē zhì bìng jūn hái huì bèi zhège kàngtǐ zuòyòng a, zǔzhǐ tā zhì bìng, suǒyǐ zhège ne jiào miǎnyì zuòyòng. Nàme dì sān gè ne shì yíngyǎng zuòyòng, zhège dàjiā dōu hěn qīngchǔle duì ba, wǒmen zài zhège xīnchéndàixiè lǐmiàn yě jiǎngguò, yǒuxiē xìjūn tā nénggòu héchéng yīxiē yíngyǎng wùzhì zuì chángjiàn de dàcháng gǎn jūn, héchéng wéishēngsù k ó, nàme zuìhòu ne, hái yǒuxiē qítā de zuòyòng, bāokuò wǒmen bǎ tā jiàozuò páidú zuòyòng, kàng zhǒngliú kàng shuāilǎo, nàme zhège zuòyòng ne shízhì shàng wǒmen xiànzài kàn dào nénggòu kàng zhǒngliú kàng shuāilǎo nénggòu páidú,
shízhì shàng zhège zuòyòng dōu shì jiànjiē de,
'When pathogenic bacteria come in, certain pathogenic bacteria will be affected by this antibody to prevent it from causing disease, so this is called immune function. And the third one is nutrition. Everyone knows this well, right? We also said in this metabolism that some bacteria can synthesize some nutrients, the most common one is Escherichia coli, which synthesizes Vitamin K, so in the end, there are also other effects, including what we call detoxification, anti-tumor and anti-aging, so these effects are actually seen now that it can fight tumors, fight aging, and detoxify. In essence, all these effects are indirect.'

In the excerpt, we can see that the instructor introduces different key terms and references (the functions of antibodies and bacteria) and keeps track of them with demonstrative-derived definiteness markers (*zhège*) over time; the instructor also uses conjunctions (*nàme*) to lead the audience to make inferences about the nature and types of the antibody functions; the instructor often breaks the flow of thought and text down into chunks with conjunctions and discourse particles (*suǒyǐ ... ne*); finally the instructor uses the first person plural pronoun (*wǒmen*) multiple times as well as the second person audience-*you* form (*dàjiā*) to help create a blended collective identity, which simultaneously draws the attention of the audience and invites the students to participate in joint attention and in the thinking and reasoning processes. Almost every one of the

utterances cited in the example contains at least one key lexical element discussed in the preceding sections. Although there is no doubt about individual stylistic variations in academic lecturing (Malavska, 2018; Flowerdew& Miller, 1997), these are almost certainly some of the key structural and discourse ingredients of a typical university lecture genre.

6. Conclusions and pedagogical implications

As a video-based collection of spoken academic Chinese, the UCLA Corpus of Spoken Academic Chinese (CSAC) can be exploited as a valuable resource for insights into a ubiquitous yet unique discourse genre. This paper reports some of the initial findings concerning four types of top keywords, identified on statistical grounds, and their associated lexical categories – pronominals, proximal demonstratives, discourse particles, and conjunctions. In each instance, we show the unique features they display in the predominantly academic lecture genre and how they often differ from their uses in other contexts such as ordinary conversations. It is clear even from this cursory study that expanding the scope of genre research can lead to useful findings in discourse and grammar (Tao, 1999b).

Finally, on the topic of pedagogical applications, one of the major benefits of using corpora for linguistic insights is the potential to inform and improve language learning and teaching in the area of academic language (O'Keeffe et al., 2007; Donley& Reppen, 2001; Swales, 2002; Tao, 2013; Tao& Chen, 2019). For example, video recordings of academic lectures can be carefully curated for courses in language for special and specific purposes, which has the advantage of enabling the examining and learning of this highly challenging type of language genre in a multimodal environment. Corpora with diverse disciplinary content and style differences can be helpful in guiding the selection of teaching materials (Tao& Chen, 2019; Lin, 2003) and exposing the learner to a wide variety of input. Corpus findings can be used directly in material development and related curricular activities (Donley& Reppen, 2001). As discussed earlier, high frequency vocabulary items, keywords, and the associated lexical bundles in spoken academic language as elsewhere are all important types of information that can be brought to bear on the efficient integration of lexical and grammatical learning (Conrad, 2000) and can be incorporated into curriculum design. As the construction of the CSAC database evolves, we hope to be able to

engage in more comprehensive research with increased data accumulation in both quantity and variety and eventually translate research insights into the practice of academic language teaching and learning in the future.

Acknowledgements

This research is partially funded through a UCLA Senate Faculty Research Grant (2020-21) on patterns of interaction in Mandarin Chinese. I wish to thank the following UCLA students for their research assistance with collection and transcription of the UCLA Corpus of Spoken Academic Chinese data: Yuchen Chen, Ye Jiang, Chenyang Lin, Adele Ng, Yi Ren, Sandra Tandiono, Liting Wang, Anyi Wu, and Jiaxin Zheng, and Liz Carter for her valuable editorial assistance and insightful comments. Standard disclaimers apply.

References

Alleton, V. (1981). Final particles and expression of modality in modern Chinese. *Journal of Chinese Linguistics, 9*(1), 91-115.

Alsop, S., & Nesi, H. (2009). Issues in the development of the British Academic Written English (BAWE) corpus. *Corpora, 4*(1), 71-83.

Anthony, L. (2020). AntConc 3.5.9 [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software (accessed 1 March, 2021).

Biber, D. (2006). *University Language: A Corpus-Based Study of Spoken and Written Registers*. Amsterdam: John Benjamins.

Boersma, Paul & David Weenink (2021). Praat: Doing Phonetics by Computer [Computer program]. Version 6.1.40. Retrieved from http://www.praat.org/ (accessed 27 February 2021).

Camiciottoli, B. C. (2008). Interaction in academic lectures vs. written text materials: The case of questions. *Journal of Pragmatics, 40*(7), 1216–1231.

Chafe, W. (1982). Integration and involvement in speaking, writing, and oral literature. In D. Tannen (Ed.), *Spoken and Written Language: Exploring Orality and Literacy*, 35-53. Norwood, NJ: Ablex.

Chao, Y. R. (1968). *A Grammar of Spoken Chinese*. Berkeley: University of California Press.

Chen, H. H. J., & Tao H. (2019). Academic Chinese: From Corpora to Language Teaching. In Xiaofei Lu and Berlin Chen (Eds.), *Computational and Corpus Linguistic Approaches to Chinese Language Teaching and Learning* (pp.57-79). Berlin & Singapore: Springer.

Chen, P. (2004). Identifiability and definiteness in Chinese. Linguistics, 42(6), 1129-1184.

Chen, W. (陳望道) (1962/1997). *Introduction to Rhetoric* (修辭學發凡). Shanghai: Shanghai Education Press (上海教育出版社).

Chu, C. C. (1998). *A Discourse Grammar of Mandarin Chinese*. New York: Peter Lang.

Chu, C. C. (2002). Relevance theory, discourse markers and the Mandarin utterance-final particle *a/ya*, *Journal of the Chinese Language Teachers Association, 37*(1), 1-42.

Chu, C. C. (2015). Utterance final particles. In Rint Sybesma (Ed.), *Encyclopedia of Chinese Language and Linguistics*. Consulted online on 01 March 2021. First published online: 2015.

Conrad, S. (2000). Will Corpus linguistics revolutionize grammar teaching in the 21st century? *TESOL Quarterly, 34*(3), 548–560.

Couper-Kuhlen, E., & Selting, M. (2018). *Interactional Linguistics*. Cambridge: Cambridge University Press.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213–238.

Culpeper, J. (2009). Keyness: Words, parts-of-speech and semantic categories in the character-talk

of Shakespeare's Romeo and Juliet. *International Journal of Corpus Linguistics, 14*(1), 29–59.

Donley, K. K., & Reppen, R. (2001). Using corpus tools to highlight academic vocabulary in SCLT. *TESOL Journal, 10*(2-3), 7–12.

Du Bois, J. W., Schuetze-Coburn, S., Cumming, S., & Paolino, D. (1993). Outline of discourse transcription. In Jane Edwards and Martin Lampert (Eds.), *Talking Data* (pp.45-89). Hillsdale, NJ: Lawrence Erbaum Associates.

Egbert, J. & Biber, D. (2019). Incorporating text dispersion into keyword analyses. *Corpora, 14*(1), 77–104.

English Language Institute (2003). *MICASE Manual: The Michigan Corpus of Academic Spoken English*. Ann Arbor: The University of Michigan.

Fang, Mei (方梅) (2002). Grammaticalization of the demonstratives zhe and na in Beijing Mandarin (指示詞"這"和"那"在北京話中的語法化). *Zhongguo Yuwen* (**中國語文**), *4*, 343-356.

Flowerdew, J., & Miller, L. (1997). The teaching of academic listening comprehension and the question of authenticity. *English for Specific Purposes, 16*(1), 27-46.

Fraser, B. (1990). An approach to discourse markers. *Journal of Pragmatics, 14*(3), 383-398.

Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics, 35*(3), 305-327.

Georgakopou Lou, A., & Goutsos, D. (1998). Conjunctions versus discourse markers in Greek: The interaction of frequency, position, and functions in context. *Linguistics, 36*(5), 887-917.

Goodwin, C. (2000). Action and embodiment within situated human interaction. *Journal of Pragmatics*, *32*(10), 1489-1522.

Goodwin, C. (2013). The co-operative, transformative organization of human action and knowledge. *Journal of Pragmatics*, 46(1), 8-23.

Gries, S. T. (2021). A new approach to (key) keywords analysis: Using frequency, and now also dispersion. *Research in Corpus Linguistics, 9*(2), 1-33.

Han, Yi & Yuntong Liu (韓毅、劉運同) (2020). Construction and applications of a spoken academic Chinese corpus (漢語學術口語語料庫的創建與應用研究). *Corpus Linguistics* (**語料庫語言學**), *7*(2), 70-82.

Hansen, M. B. M. (2012). The semantics of pragmatic expressions. In Hans-Joerg Schmid (Ed.), *Cognitive pragmatics: Handbook of Pragmatics* (pp. 589-613). Berlin: Mouton de Gruyter.

Huang, S. (1999). The emergence of a grammatical category definite article in spoken Chinese. *Journal of Pragmatics, 31*(1), 77-94.

Kärkkäinen, E. (2003). *Epistemic Stance in English Conversation: A Description of Its Interactional Functions, with a Focus on 'I think'*. Amsterdam: John Benjamins.

Khuwaileh, A. A. (1999). The role of chunks, phrases and body language in understanding coordinated academic lectures. *System, 27*(2), 249-260.

Lee, J. J. (2009). Size matters: An exploratory comparison of small- and large-class university lecture introductions. *English for Specific Purpose, 28*(1), 42-57.

Lee-Wong, S. M. (1998). Face support - Chinese particles as mitigators: A study of *ba a/ya* and *ne*. *Pragmatics, 8*(3), 387-404. International Pragmatics Association.

Lee-Wong, S. M. (2001). Coherence, focus and structure: The role of discourse particle *ne*. *Pragmatics, 11*(2), 139-153. International Pragmatics Association.

Levy, R., & Manning, C. D. (2003). Is it harder to parse Chinese, or the Chinese Treebank? *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, 439-446.

Li, C. N. & Thompson, S. A. (1981). *Mandarin Chinese: A Functional Reference Grammar*. Berkeley and Los Angeles: University of California Press.

Li, Y. (李裕德) (1985). *Grammar of scientific Chinese* (**科技漢語語法**). Beijing: Metallurgical Industry Press (冶金工業出版社).

Li, X., & Ono, T. (2019). Introduction: A multimodal approach to Chinese Interaction. In Xiaoting Li & Tsuyoshi Ono (Eds.) *Multimodality in Chinese Interaction* (pp.1-9). Berlin: De Gruyter Mouton.

Lin, Chin-hui (林欽惠) (2003). A pedagogical grammar perspective on sentence-final particle *a* in Mandarin Chinese (漢語句末助詞「啊」之教學語法初探). MA thesis, National Taiwan Normal University. Taipei.

Liu, Liyan (劉麗艷) (2009). *Zhege* and *nage* as discourse markers (作爲話語標記的“ 這個” 和“ 那個”). Language Teaching and Research (**語言教學與研究**), 1, 89-96.

Liu, C. (劉貞妤), Chen, H. (陳浩然) & H. Yang (楊惠媚) (2016). Compiling a Chinese academic wordlist based on an academic corpus (藉學術語料庫提出中文學術常用詞表: 以人文社會科學爲例). *Journal of Chinese Language Teaching* (**華語文教學研究**), *13*(2), 43-87.

Liu, C. (劉貞妤), Chen, H. (陳浩然) & H. Yang (楊惠媚) (2017). Study on the lexical bundles in Chinese academic writing (中文人文社會科學論文常用詞串之研究). *Journal of Chinese Language Teaching* (**華語文教學研究**), *14*(1), 119-152.

Malavska, V. (2016). Genre of an academic lecture. *International Journal on Language, Literature and Culture in Education*, *3*(2), 56-84.

Ochs, E. (1979). Planned and unplanned discourse. In T. Givón (Ed.), *Discourse and Syntax*. (Vol 12, pp. 51-80). New York: Academic Press.

Ochs, E., Schegloff, E. A., & Thompson, S. A. (Eds.) (1996). *Interaction and Grammar*. New York: Cambridge University Press.

O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A Simplest systematics for the organization of turn-taking for conversation. *Language, 50*(4/1), 696-735.

Scheibman, J. (2002). *Point of View and Grammar: Structural Patterns of Subjectivity in American English Conversation*. Amsterdam: John Benjamins.

Schiffrin, D. (1987). *Discourse Markers*. Cambridge: Cambridge University Press.

Scott, M., & Tribble, C. (2006). *Textual Patterns: Key Words and Corpus Analysis in Language Education*. Amsterdam: John Benjamins.

Simpson-Vlach, R. (2006.) Academic speech across disciplines: Lexical and phraseological distinctions. In K. Hyland & M. Bondi (Eds.), *Academic Discourse Across Disciplines* (pp.295-316). Bern: Peter Lang.

Simpson-Vlach, R. (2013). Corpus analysis of spoken English for academic purposes. In C. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*. Chichester, UK: Blackwell Publishing Ltd.

Stivers, T., & Sidnell, J. (2005). Introduction: Multimodal interaction. *Semiotica, 156*(1/4), 1-20.

Song, Z., & Tao, H. (2009). A unified account of causal clause sequences in Mandarin Chinese and its implications. *Studies in Language, 33*(1), 69-102.

Swales, J. (1990). *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Swales, J. M. (2002). Integrated and fragmented worlds: EAP materials and corpus linguistics. In J. Flowerdew (Ed.), *Academic discourse* (pp.150-164). London: Longman.

Tao, H. (1996). *Units in Mandarin Conversation: Prosody, Discourse, and Grammar*. [Studies in Discourse and Grammar, 5]. Amsterdam: John Benjamins.

Tao, H. (1999a). The grammar of demonstratives in Mandarin conversational discourse: A case study. *Journal of Chinese Linguistics, 27*(1), 69-103.

Tao, H. (1999b). Discourse taxonomies and their grammatico-theoretical implications (試論語體分類的語法學意義). *Contemporary Linguistics* (**當代語言學**), *1*(3), 15-24.

Tao, H. (2003). Phonological, grammatical, and discourse evidence for the emergence of *zhidao* 'to know' constructions in Mandarin conversation (從語音、語法和話語特徵看"知道"格式在談話中的演化). *Zhongguo Yuwen* (**中國語文**), 4, 291-302.

Tao, H. (2013). *Corpus of Written Academic Chinese*. ACTFL CALPER Brochure. State College, PA: Pennsylvania State University.

Tao, H. (2015). Profiling the Mandarin spoken vocabulary based on corpora. In William Wang & Chaofen Sun (Eds.), *Oxford Handbook of Chinese Linguistics* (pp.336-347). Oxford: Oxford University Press.

Tao, H. & Chen, H. H. J. (2019). Chinese for specific/professional purposes: An introduction. In H. Tao & H. Chen (Eds.), *Chinese for Specific/Professional Purposes: Theory, Pedagogical Applications, and Practices*, vii-xiii. Singapore: Springer Nature.

Thompson, S. A., & Hopper, P. J. (2001). Transitivity, clause structure, and argument structure: Evidence from conversation. In Joan Bybee & Paul J. Hopper, (Eds.), *Frequency and the Emergence of Linguistic Structure* (pp.27-60). Amsterdam: John Benjamins.

Thompson, S. (1994). Frameworks and contexts: A genre-based approach to analysing lecture introductions. *English for Specific Purposes, 13*(2), 171-186.

Thompson, S. E. (2003). Text-structuring metadiscourse, intonation and the signalling of organization in academic lectures. *Journal of English for Academic Purposes, 2*(1), 5-20.

Thompson, P., & Nesi, H. (2001). The British Academic Spoken English (BASE) Corpus project. *Language Teaching Research, 5*(3), 263-264.

Traugott, E. C. (2018). Rethinking the role of invited inferencing in change from the perspective of interactional texts. *Open Linguistics 4*(1), 19-34.

Tseng, S. C. (2001). Highlighting utterances in Chinese spoken discourse. In *Language, Information and Computation. Proceedings of the 15th Pacific Asia Conference on Language, Information and Computation*, 163-174.

Wang, X. (2020). Managing a suspended course of action: A multimodal study of suoyi 'so'-prefaced utterances in Mandarin conversation. *Chinese Language and Discourse, 11*(2), 306-334.

Xiao, Z., & McEnery, A. (2005). Two approaches to genre analysis: Three genres in Modern American English. *Journal of English Linguistics, 33*(1), 62-82.

Xu, Dan (徐丹) (1988). Notes on the asymmetrical properties of *zhe* and *na* (淺談這/那的不對稱). *Zhongguo Yuwen* (**中國語文**), *2*, 128-130.

Yeo, J. Y., & Ting, S. H. (2014). Personal pronouns for student engagement in arts and science lecture introductions. *English for Specific Purposes,* 34, 26–37.

Yin, Shulin (殷樹林) (2009). Grammaticalization of the discourse markers *zhege* and *nage* and relevant factors (話語標記"這個"、"那個"的語法化和使用的影響因素). *Foreign Language Research* (**外語學刊**), *149*, 92-96.

Zhu, D. X. (朱德熙) (1982). *Lectures on Chinese grammar* (**語法講義**). Beijing: Commercial Press (商務印書館).

中文學術口語的若干主要詞彙特徵及其教學啟示

陶紅印

美國加州大學洛杉磯分校

摘要

本文以一個 15 萬詞的高校課堂講課學術口語語料庫為基礎，探討中文學術口語基本詞彙的重要特徵。在描述了基本詞彙的幾個主要特點後，我們繼而確認學術口語中的關鍵詞，主要方法是將學術口語語料庫與其他相關較大型語料庫進行比較。然後我們再對若干重要關鍵詞（含人稱代詞、指示代詞、語氣詞及連接成分）進行更全面的多模態分析。結尾討論了本研究對語體理論的貢獻以及華語二語教學的務實意義。

關鍵詞：學術中文、關鍵字、人稱代詞、指示代詞、語氣詞、連接成分