

1. ABOUT THE DATASET

Title: The Corpus of Chinese Academic Written and Spoken English (CAWSE)

Creator(s): Yu-Hua Chen^[1], Simon Harrison^[2], David Oakey^[3], Michael Stevens^[4], Shanru Yang^[4], Godwin Ioranim-Uba^[4], Qianqian Zhou^[4], Radovan Bruncak

Organisation(s): 1. University of Leeds, 2. City University of Hong Kong, 3. University of Liverpool, 4. University of Nottingham Ningbo China

Rights-holder(s): Unless otherwise stated, Copyright 2026 University of Leeds

Publication Year: 2026 (earlier version released in 2019)

Description:

The Corpus of Chinese Academic Written and Spoken English (CAWSE) comprises L2 English samples produced by L1 Chinese students enrolled in a preliminary-year programme at an English-medium instruction (EMI) / transnational (TNE) campus in China. Currently available data include coursework (657 scripts; approx. 1,500 tokens per script), interviews (122 sessions, 10 minutes each), and presentations (184 sessions, 10 minutes each), covering the full range of awarded grades. In total, this dataset comprises 657 essays (1,004,523 tokens) and 51 hours of spoken data (352,333 tokens).

In addition, a subcorpus of written exam scripts and a multimodal subcorpus of student group discussions may be available upon request. The corpus was compiled between 2016 and 2019, with transcription completed in the following years. It was first released in 2019, and the current release contains only updates to the documentation, with no changes made to the dataset itself. The data available here have been fully anonymised, with all personally identifiable information removed.

This pre-AI EMI/TNE corpus of over one million tokens offers a valuable resource for analysing L2 English, enabling the investigation of lexical, syntactic, and discourse features across

proficiency levels, task types, genres, and other contexts. Please refer to Chen et al. (2024) for further details about the corpus.

Cite as: Yu-Hua Chen, Simon Harrison, David Oakey, Michael Stevens, Shanru Yang, Godwin Iorati-Uba, Qianqian Zhou, Radovan Bruncak (2026). *The Corpus of Chinese Academic Written and Spoken English (CAWSE)*. University of Leeds / University of Nottingham Ningbo China. [Dataset] <https://doi.org/10.5518/1836>

Related publication:

Chen, Y. H., Harrison, S., Stevens, M. and Zhou, Q. (2024). Developing A Multimodal Corpus of L2 English from an EMI University in China. *Corpora*, 19(1), 1-15.

<https://doi.org/10.3366/cor.2024.0295>

Chen, Y. H. & Bruncak, R. (2020). Transcribear: Introducing a secure online transcription and annotation tool. *Digital Scholarship in the Humanities*, 35(2), 265-275.

<https://doi.org/10.1093/llc/fqz016>

Chen, Y. H., Harrison, S. & Weekly, R. (2019). “I don’t have communicate ability”: Deviations in an L2 Multimodal Corpus of Academic English from an EMI University in China – Errors or ELF? In H. Parviainen, M. Kaunisto and P. Pahta (Eds), *Corpus Approaches into World Englishes and Language Contrasts* (eVarieng).

https://varieng.helsinki.fi/series/volumes/20/chen_harrison_weekly/

Harrison, S. & Chen, Y. H. (2021). From Dancing K-Pop with Chinese and “English in class please”: English language policy negotiations as relational-langaging episodes during classroom interaction. *RELC Journal*, 52(2), 270-286.

<https://doi.org/10.1177/00336882211022123>

Stevens, M., Chen, Y. H. & Harrison, S. (2020). The EMI campus as site and source for a multimodal corpus. In A. Čermáková & M. Malá (Eds), *Variation in Time and Space: Observing the World through Corpora* (series “Diskursmuster - Discourse Patterns”). Berlin: De Gruyter, 377-402. <https://doi.org/10.1515/9783110604719-015>

Contact: Yu-Hua Chen (y.chen8@leeds.ac.uk)

2. TERMS OF USE

This dataset is licensed under a Creative Commons Attribution 4.0 International Licence:

<https://creativecommons.org/licenses/by/4.0/>

3. PROJECT AND FUNDING INFORMATION

Title: The Corpus of Chinese Academic Written and Spoken English (CAWSE)

Dates: 2015-2019

Funding organisation: Ningbo 3315 Plan; University of Nottingham Ningbo China Matching Funding

4. METHODS

In this section, the participants' backgrounds, including their nationalities/L1s and disciplines, as well as the types of assessment tasks and the sampling principles, are described. The file ID system and the transcription conventions are also outlined.

4.1 Participants

4.1.1 Background

Most participants in the CAWSE Assessment (approximately 98%) subcorpus are L1 Chinese students. A small number of participants come from other linguistic and national backgrounds, including Indonesian, Korean, Sudanese, British, Ghanaian, Russian, Tanzanian, Tajikistani, and Thai backgrounds, reflecting the diversity of the preliminary-year student population at this campus (see Table 1).

Table 1 Background of the student participants

Nationality	Number of participants
Chinese	745
Indonesian	3
Korean (S)	3
Sudanese	2
British	1
Ghanaian	1
Russian	1
Tanzanian	1
Tajikistani	1
Thai	1

For Chinese students, the minimum English requirement for undergraduate entry at this campus is 115 out of 150 in the English component of the National Chinese College Entrance Examination (Gaokao). For international students, the equivalent requirement is IELTS 5.5.

4.1.2 Academic disciplines

All assessment data in CAWSE were collected from preliminary-year undergraduate courses focusing on English for Academic Purposes (EAP) or English for Specific Purposes (ESP).

Although the students come from a wide range of majors, the ESP courses are broadly divided into Arts & Social Sciences (A&S) and Science & Engineering (S&E). See Table 2 for a detailed breakdown of the academic disciplines represented in the corpus.

For the assessment data collected, some assessment tasks are discipline-specific, while others use common prompts across disciplines.

Table 2 Academic backgrounds of students

Academic Discipline	Major	Number of participants
Arts & Social Science (A&S)	Business	221
	Economics	64
	English	25
	International Communications	95
	International Studies	49
	English with International Business	11
Subtotal	6	465
Science & Engineering (S&E)	Architecture	15
	Computer Science	80
	Engineering	157
	Environmental Science	7
	Mathematics	35
Subtotal	5	294

4.1.3 Types of assessment data

The assessment component includes both spoken and written assessment data.

Table 3 Composition of the CAWSE written and spoken subcorpus (extracted from Chen et al., 2024, p. 8)

Mode	Written		No. of scripts / tokens	Spoken	No. of sessions, tokens / duration
Individual	Exam	Reading and summary writing (avg. 488 tokens)	365 (178,278 tokens)	10-minute presentation (avg. 1,146 tokens)	184 (210,846 tokens / approx. 30 hrs 40 mins)
		Reading and writing (A & B) (avg. 615 tokens)	612 (376,326 tokens)		
		Listening and writing (avg. 251 tokens)	515 (129,119 tokens)		
	Coursework	Writing assignment (A & B) (avg. 1529 tokens)	657 (1,004,523 tokens)		
Teacher–student				10-minute interview (avg. 1,160 words)	122 (141,487 tokens / approx. 20 hrs 20 mins)
Sub-total			2,149 (1,688,246 tokens)		306 (352,333 tokens / approx. 51 hrs)

For the interviews in spoken assessment, a typical interview structure includes:

- o general questions about university life,
- o a two-minute monologue,
- o and follow-up discussion with the examiner.

In terms of the presentations, students select topics related to their majors. Presentations typically include the following elements:

- o title,
- o background,
- o problem,
- o solution,
- o and evaluation.

As can be seen in Table 3, for the written assessment, the coursework and the reading and writing exam are divided into the two broad disciplinary groups, A (Arts & Social Sciences) and

B (Science & Engineering). The other two exams — listening and summary writing, and reading and summary writing — use a single set of exam papers for all students.

4.1.4 Sampling

To ensure representativeness, assessment samples were selected across score bands ranging from 0–39 to 90–100 wherever possible.

For written assessment: a maximum of 100 scripts was collected per score band. See Table 4 for a breakdown of the written data.

Table 4 Scoring distribution of written assessment samples

Assessment Score band	Essays		Exam scripts			
	Reading & writing (A&S)	Reading & writing (S&E)	Listening & writing	Reading & writing (General)	Reading & writing (A&S)	Reading & writing (S&E)
0-39	24	19	28	11	8	5
40-49	59	58	72	100	79	42
50-59	100	96	100	100	100	100
60-69	100	86	100	100	100	94
70-79	71	28	100	50	52	23
80-89	13	3	80	4	4	4
90-100	0	0	35	0	1	0
Total	367	290	515	365	344	268

For spoken assessment, due to transcription resource constraints, a maximum of 30 recordings was collected per score band. See Table 5 for a breakdown of the spoken data.

Table 5 Scoring distribution of spoken assessment samples

Assessment Score band	Interview	Presentation
0-39	1	4
40-49	7	30
50-59	16	30
60-69	30	30
70-79	30	30
80-89	30	30
90-100	8	30
Total	122	184

4.2 File ID system

Each file ID contains:

- a register marker,
- a task abbreviation,
- and a four-digit participant code.

See some examples of the file ID system below.

Table 6 The File ID system for different tasks

Register	Task	File ID Example
Spoken - s	Individual Interview - ii	sii0001
	Academic Oral Presentation - aop	saop0001
Written - w	Independent Writing (A&S) - iwas	wiwas0001
	Independent Writing (S&E) - iwse	wiwse0001
	Listening and Writing - lw	wlw0001
	Reading and Writing (General) - rwg	wrwg0001
	Reading and Writing (A&S) - rwas	wrwas0001
	Reading and Writing (S&E) - rwse	wrwse0001

4.3 Transcription conventions

The spoken data were orthographically transcribed and anonymised. Verbal features (e.g., code-switching) and non-vocal features (e.g., laughter) were annotated, along with various linguistic features (e.g., lexical deviation) and non-linguistic features (e.g., time stamps) and conventions. The written data also underwent similar anonymisation and annotation procedures.

For detailed transcription and annotation conventions, please refer to two separate documents:

- 1) transcription-conventions-exam-scripts, and
- 2) transcription-conventions-spoken-data.

5. SAMPLE EXTRACTS

This section includes sample extracts from spoken assessment (interviews and presentations) and written assessment (essays, listening and writing exam and reading and writing exam scripts).

These samples preserve second language features, which are annotated for, including but not limited to:

- grammatical deviations
- lexical errors
- hesitations
- repetitions
- discourse features

For detailed transcription and annotation conventions, please refer to the two transcription convention documents also available in this repository:

1. transcription-conventions-exam-scripts
2. transcription-conventions-spoken-data

5.1 Spoken assessment

5.1.1 Interview sample script

<Greetings & recording of the date, location, interviewer & student names, ID number, etc.>

(00:00.000)-(00:31.241)

t0005: I'll now ask you some general questions (1.4) which parts of your first semester here have you enjoyed the most

s0094: erm maybe the: (1.0) the life in: the life I hh come I take part in the roo- roommate er:

because I have three very good roommates en: we often go sh- shopping and eat some delicious food

t0005: very nice

s0094: yeah

t0005: great (1.0) when you study do you usually study alone or with friends

s0094: en: sometimes er I I often study with my roommates we go er library and self-study er room

t0005: mhm (1.0) OK (1.0) we'll now move on to part two in this part I would like you to give a short speech on the topic I will give you you have one minute to think about what you are going to say and up to two minutes to speak here is the you can make notes here is (1.4) excuse me a piece of paper (2.1) I'd like you to try to speak for the full two minutes if you can OK

s0094: OK

t0005: and I would like you to talk about (1.2) artificial intelligence

<prep time 1 min>

(01:50.702)-(02:57.048)

t0005: please begin speaking now

s0094: mm: er about artificial intelligence er artificial intelligence is a engineering and science er about making a machines which is intelligent especially the computer intelligence program and er I will tal- er the intellige- artificial intelligence can be used in two <dv>x{aspects}</dv> the first is expert system er: this means the knowledge of expert and second is (1.0)

<dv>x{aspect}</dv> is ro- robotic er that means er (1.1) s- scientis- scientist use artificial intelligence to make a robot er the robot can: en do some cleaning or: erm or: ma- many other things (1.0) hh er they can: do the explo- they can explore the space er and do some things the human can't do hh mm (1.4) er I er my family usually have a cleaning robot er which can help my mother to do some housework I think this er I think this reduce the: housework pla- <dv>x{pressure}</dv> of my mother er my feeling about the artificial intelligence is I think artificial intelligence is very useful for er our life because er it can do many things people can't do and many er: dangerous work er and boring work mm but I think it also have the disadvantage er because if many work are did are did by the robot er many people will lose their j- job hh er...

5.1.2 Presentation sample script

(00:58.216)

s0032: hello everyone my name is <an>xxx</an> I'm from Finance Accounting and Managements Management <laughing> and Management today I'm glad to be here and en for the next ten minutes or so er: I will looking at the China in globalisation mm: in other words (1.8) mm we will talk about facing the challenge of economic globalisation in China I've divided I have divide my presentation to three parts er: I will describing China's economic overview and definition of globalisation firstly and then I will (1.4) introduce the challenge China face er finally in my final part of presentation I will give some solutions of globalisation China face (1.2) hh <laughing> well I will star- start it I will start it by looking at China's economic

overview hh I like to draw your a- attention to the graphs that I shows in Power Point (1.0) hh en: the interest rate and repo rate have <ut>x</ut> decrease from two thousand and eight to: two thousand and nine (1.0) hh and and in this picture res- reserve cu- currency have show the same decrease (1.1) hh en it may shows that China has more and more become a part of: globalisation en and: (1.0) as we all know China China's economic overview have present a picture of general <dv>x{prosperity}</dv> and in two thousand and eight China overcome the German become the third biggest economy in the world (1.0) hh so so that we should say erm China needs globalisation and needs the cooperation with west countries and world hh en (2.8) er: now I like to move on to my second part I will focus on the definition of globalisation and give some instruction of disadvantages and advantages of globalisation globalisation bring us a lot of convenience we eat: <laughing> Kentucky Fried Chicken er we use L V Louis Vuitton' bags and we wear Nike erm also we drive B M W cars (1.0) en without globalisation we can't have such convenient and comfortable life but: however en: globalisation's not always be advantages (1.4) it it also have some disadvantages and shortcomings...

5.2 Written assessment

5.2.1 Essay Sample (A&S)

Critically evaluate Ducker's MBO and assess its usefulness in the public sector

The concept of the MBO system was firstly created by Peter Drucker in the mid 1950s. It was widely used and had a long and deep influence in the managing field of all professions and trades. Based on Peter Drucker's concept, the MBO system mainly brings various advantages to the management of the public sector. It can not only allow the employees take part in the goal setting process, but can also motivate them to improve the efficiency when participating in challenging tasks. In addition, the MBO system can also create tacit agreement and better cooperation which can benefit a long-term development. It has to be conceded that although there are some disadvantages, for example, the system should be systematically done, and it may also be resented by subordinates causes they may be stressful when setting goals with the manager. However, the benefits outweigh the disadvantages. It will argued here that the MBO system is widely applied in the public sector and truly benefits it. The essay will evaluate the applicability of MBO system in the public sector, it will critically evaluate the disadvantages of MBO at first, then it will show the advantages in government's goal accomplishment, protection of the power of democracy and laws, improvement of national happiness index and setting up evaluation system.

Some opponents claim that the MBO system cannot be applied in the public sector, for four main reasons. First, they say that based on Drucker's theory, the system must be systematically applied,

or it will lead to the missing links, which may damage the working process in some situations. Second, it is claimed that MBO does not consider the available sources and presentation, in addition, it may have bad influence on the cooperators. A suitable example of this can be seen in the argument made by Brunson (2002), during the original use period of the MBO system, the opposite effect appeared as nearly none of the Swedish government officials were willing to take charge of the task, because they knew the difficulty of it. The existence of the MBO system was controversial and most political leaders refused to tolerate it. (Ibid.) Third, the goals set by the MBO system in public sector may be difficult to achieve. Murphy (1983) argues that the circumstances of the society are changing constantly, therefore, the goals change as well. Since the employees cannot reach their initial goals, they cannot obtain the promised reward smoothly, which adds burden to organization. Thus, it would be hard to motivate the employees to improve the efficiency. Finally, the managers begin to force the employees to achieve the best standard of developed by MBO. As a result, they ignore the demand of the workers, which will encourage their discontent. It can cause dissatisfaction and expand the sense of distance between managers and their subordinates...

5.2.2 Listening and writing exam sample script

This lecture talks about the impact of design on branding, and discusses the success of Apple's branding design. Furthermore, the speaker also talks about the changes of Starbucks logo. First of all, Apple company was created by Steve Jobs, and is successful now. Apple's products include many areas. For instance, ipod is for music industry, and iphone refers to phone industry. As for computer industry imac absorbs plenty of <dv>cosumers{consumers}</dv>. Secondly, the color of Apple is sample, usually is black and white. Whereas ipod has other color like gold, and the colorful products absorbs many young people. Another success is the design of Apple stores. The Apple in New York is sharp and fashion. Apple store even has children's table. Thirdly, the Apple's logo is versus a green apple. So Steve Jobs don't have to worry about the copyright. Finally, as for starbucks, the logo is a star <dv>godness{goddess}</dv>. The <dv>godness{goddess}</dv> changed from naked body to be covered.

5.2.3 Reading and writing exam sample script

(Q1: define; Q2: compare & contrast)

Q1:

Organic farming refers to a rational solution for the global demand for the food production compared to conventional farming. It is a traditional type of agriculture and its existence lasted for

a long time before the Industrial Revolution. Organic farming is more natural and less harmful to the environment, thus it should be implemented instead of conventional farming. There are three main reasons for why it's necessary to promote organic farming. First of all, organic farmers use seeds from previous crops rather than genetically modified (GM) seeds. At the same time, by growing various kinds of seeds to guarantee biodiversity, it lowers the risk of product reduction when disease attacks. Secondly, avoiding using chemical fertilisers, organic farming allows more natural ones. For instance, 'green manure' such as a plant, seed or animal waste plays a significant role in crop production and soil condition. Finally, organic farming promotes insect use and protects plants from diseases by crop rotation, which keeps a balance between plants and solid nutrients. Although there are various advantages of organic farming, people still have a doubt if this way can solve global feeding problems in the future.

Q2:

The graph demonstrates the percentage of total agricultural land used for organic farming in different countries. As shown in the graph, all the countries begin with the low percentage of organic farming in 2006, but Austria, Liechtenstein and China promote the organic farming and leads to its continually increasing percentage from 2006 to 2015. Austria has the highest percentage of organic farming and it has risen because people pay more attention to healthy food. Another reason for this is that Austria has wide natural land and advanced technology and experts agricultural production around the world, thus it's rational to promote organic farming to satisfied people's demand for healthy crops. Liechtenstein's organic farming percentage has doubled from the 6.5% in 2006 to 13% in 2015. It may result from some government agricultural policies to encourage farmers to grow organic crops. As for China, although there's an increasing tendency in organic farming, from 2.5% to 4%, the percentage is still low compared to other countries. China has developed quickly in the past few years. With the development of economy and technology, people attach great importance to diet, which contributes to the increasing demand for organic food. But because of its large population, the widespread of organic food supply may be impossible. Despite three countries' organic farming percentage increase, organic farming percentage in several countries has no differences and even dropped. Organic farming in USA and Argentina remains 7.5% and 8.5% from 2006 to 2015. USA owns advanced technology and wide land. Farmers grow and starve crops with the help of technology and machine, so it lowers the possibility for them to implement organic farming because of its lower efficiency. Argentina doesn't own wide lands to grow organic crops thus the percentage of land used for organic farming is still 8.5%.

Nevertheless, India is the only country among all, of which organic farming has dropped from 5.5% to 5%. The continually increasing population in India results in more demand for food. Farmers should increase the output with the help of chemical fertilisers and even GM seeds. Government focuses more on hunger problem solving rather than the quality of food. Therefore, the percentage of organic farming in India has decreased.

5.2.4 Reading and writing exam sample script (S&E)

Question 1

Contemporary art museum refers to a place where has <dv>exhibitions{exhibitions}</dv> of modern art, including paintings, sculptures, movies and so forth. When architects are designing this building in suburban area, there are three factors, the choice of materials, the size of the museum, and the context of the site, that need to be considered. To begin with, the choice of materials is of paramount importance regarding the success of an art museum. Specifically, it is argued that concrete walls which have a smooth surface are to the detriments of art display because the <dv>maintainence{maintenance}</dv> of these walls is demanding. In this case, drawings may not be hang on the wall but hang form the top. To avoid the inconvenience other materials such as masonry could be chosen as the material of walls. Another contributing factor is the size the museum. <dv>Mordern{Modern}</dv> art displays often require large space to be installed, in which case, the space of contemporary art museum should be adequate for the transportation and installation of art work. In addition to material and size factor, the context of the location should also be taken into consideration. Situated in the suburban area, this museum would be less attractive due to the transportation factors. To alleviate this problem, the entrance of the museum should be designed facing the main avenue or the exit of subway. This practice enables people to travel easier. To conclude, the design of a contemporary art museum situated in the suburban area should be based on the materials, the size and the context factors.

Questions 3

Micro-apartments refer to small size homes designed for small household families living in the city. This type of buildings receives varied responses regarding the question that whether it is a paradigm to address the cities' housing crisis. This essay will argue that although these apartments are criticized for affordability problems, they are preferable for tackling housing shortages to a larger extent. Three factors, including the affordability of housing, the function of these buildings and the size of these apartments, would be analyzed in the following essay.

The first and foremost factor is the affordability of micro-apartments. It is claimed by architecture

firms that this type of buildings is designed for low or average paid workers who work in the city. However, complains can be found that these initial targets are incapable of affording to live in these apartments. Although this is the reality that these working force couldn't afford \$2750 to live in a 28 square metre room, the situation can be altered. The current dilemma arises because the modular units is a new technology to develop and micro-apartments are perceived as high-risk projects which need more interest to receive loans from lenders. Once this category of housing goes viral worldwide, the price of these units would be remarkably decreased.

Another trigger factor would be the function of these unites. Multifunctional furniture, kitchens and bathrooms are installed in these apartments. Nevertheless, critics argue that these functions are insufficient and even the architects agree it because they designed gym, storage room, lounge room and community room outside of these apartments to compensate for the loss of function. Despite of the fact that certain activities are unable to occur within the apartments, it can be achieved within inches of dwellers' homes. In addition, people strengthen the sense of community by sharing these rooms with their neighbourhoods.