


A pilot study of data-driven learning approach in teaching Chinese vocabulary

Yu-Ting Tseng^a and Li-Ping Chang^b

^aGraduate Program of Teaching Chinese as a Second Language, National Taiwan University, r07146012@ntu.edu.tw and

^bGraduate Program of Teaching Chinese as a Second Language, National Taiwan University, , lchang@ntu.edu.tw

How to cite: Tseng, Y.; Chang, L. (2023). A pilot study of data-driven learning approach in teaching Chinese vocabulary. In *CALL for all Languages - EUROCALL 2023 Short Papers*. 15-18 August 2023, University of Iceland, Reykjavik. <https://doi.org/10.4995/EuroCALL2023.2023.16972>

Abstract

The Data-Driven Learning (DDL) approach advocates a shift from passive knowledge recipients to active researchers among learners. This is achieved by utilizing abundant and context-rich target language inputs in a bottom-up learning process (Johns, 1990). Despite the scarcity of empirical research on the implementation of DDL in Chinese as a Second Language (CSL) classrooms, this study conducted a teaching experiment focused on confusable words to explore the approach's effects and learners' attitudes. Five advanced-level CSL learners from diverse native language backgrounds participated in the study, being taught ten sets of confusable words over a five-week period. The first five lessons employed an indirect DDL method, while the latter five utilized a direct DDL method with Sketch Engine. To evaluate the effectiveness of the DDL approach, a pre-test was administered prior to the course, followed by a post-test, a questionnaire, and one-on-one interviews after the completion of the course. The results of the post-test revealed an average improvement of 24% compared to the pre-test with statistical significance. Additionally, learners exhibited a positive attitude towards the course, expressing a preference for learning vocabulary through collocations and showing a desire to observe pre-selected concordance lines under the guidance of the instructor.

Keywords: data-driven learning, corpus, DDL instructional design, Chinese teaching, empirical study.

1. Introduction

Johns (1990) proposed Data-Driven Learning (DDL), which advocates a bottom-up learning process driven by analyzing typical, large, and contextualized data of the target language. Numerous studies have validated the benefits of the DDL approach for students (Mizumoto & Chujo, 2015; Boulton & Cobb, 2017; Lee, Warschauer & Lee, 2019). However, its application in teaching Chinese as a Second/Foreign Language (CSL/CFL) has remained limited (Chang, 2022). Only Smith (2011) and Yeh & Zhang (2018) have implemented DDL in CFL classrooms, with the former incorporating five DDL tasks within textbook content, and the latter conducting an empirical study aimed at enhancing learners' usage of the discourse-linking connection *jiù* 'then' in oral storytelling. It is hoped that through this study, more CSL instructors will become aware of and be inclined to this approach in their teaching. Therefore, this study poses two primary research queries:

1. How effective is the application of DDL in enhancing learners' understanding and usage of confusable words in Chinese?
2. How do learners' preferences vary between direct and indirect DDL approaches, and what are their perspectives on instructional arrangements?

2. Method

This study conducted a five-week course consisting of ten sessions for two hours per week, while recruiting five learners with an advanced level of Chinese proficiency to participate. The first five sessions employed an indirect DDL (hands-off) approach, and the subsequent five sessions employed a direct DDL (hands-on) approach (Boulton, 2010).

To assess the effectiveness, learners were requested to undergo both a pre-test and a post-test, each comprising 40 multiple-choice vocabulary items. Upon completing all courses, they are required to fill out a questionnaire mainly adapted from Yoon and Hirvela (2004) and participate in one-on-one interviews.

3. DDL Instructional Design

3.1. Instructional Design for Indirect DDL Sessions: *jiéguǒ* 'result' and *hòuguǒ* 'consequence'

The key to the successful implementation of indirect DDL lies in the creation of paper-based materials for learners to observe. For illustration, we outline the steps involved in preparing materials focused on *jiéguǒ* 'result' and *hòuguǒ* 'consequence'. To commence, the instructor initiated an examination of the Chinese learner corpora, with the objective of scrutinizing and identifying instances of confusion prevalent among learners. The majority of these errors stemmed from disparities in the semantic prosody of the two words. 'Semantic prosody,' as elucidated by Louw (2000), refers to "a form of meaning which is established through the proximity of a consistent series of collocates." Specifically, the semantic prosody of *jiéguǒ* 'result' is neutral, while *hòuguǒ* 'consequence' always appears in a negative context. This finding was also indicated in the study of Xiao & McEnery (2006). Nevertheless, learners often employed *jiéguǒ* within a negative context, as exemplified in the following instance.

- (1) **Rúguǒ wéile jièjué jī'è de wèntí, fàngqì lǜsèshípǐnde zhòngyàoxìng de huà, zhè zàochéng de jiéguǒ shì kěyǐshuō bùkānshèxiǎng.* (If we abandon the importance of green food to solve the hunger problem, the consequences will be unthinkable.)

In addition, *jiéguǒ* has two usages, functioning as both a conjunction and a noun. Learners may mistakenly assume that *hòuguǒ* also has these two usages, leading to the production of error as shown in (2).

- (2) **Lìngwài, wǒmen dìqiú de zīyuán yóuyú wǒmen méi kǎolǜ yīzhí shǐyòng, xiànzài yě méiyǒu, hòuguǒ wǒmen miànlín quēfá zīyuán.* (Moreover, we have not been mindful in utilizing our planet's resources, which are not unlimited. Consequently, we are faced with resource shortages.)

Six sheets for the instruction of *jiéguǒ* 'result' and *hòuguǒ* 'consequence' were prepared in the classroom. Each sheet contained ten sentences with a single focus. This aided in facilitating easy observation and analysis, ultimately enhancing learners' confidence. The selection of sentences in the corpus followed the criteria of typical usage contexts and functionalities. The order of the teaching materials began with the noun *hòuguǒ*, followed by the noun *jiéguǒ*, and then the conjunction *jiéguǒ* 'therefore'. It was presumed that learners learned the noun usage of *jiéguǒ* before while *hòuguǒ* is introduced as a new word. Hence, the presentation of *hòuguǒ* usage preceded the comparison with the usage of *jiéguǒ*. This sequence intended to first acquaint learners with the distinctions in collocating words when both *hòuguǒ* and *jiéguǒ* function as nouns. Subsequently, the focus shifted to the exclusive function of *jiéguǒ* as a conjunction, demonstrating a typical usage where it indicated an unexpected outcome. This was illustrated by employing the adverb *què* (however) to indicate a contrasting

function. The instructional materials were printed on a single side to facilitate easy comparison and reference for learners, enabling them to readily observe the key differences in usage between *hòuguǎo* and *jiéguǎo*.

In the classroom, the modified five-step instructional design was adopted from Smith (2011): Observation and Analysis, Hypothesis Formulation, Hypothesis Confirmation, Summarization, and Consolidation and Application. The aim was for students to observe, hypothesize, confirm, and revise their understanding of confusable words, followed by consolidation and application of their newfound knowledge. After several cycles of these steps, students were encouraged to summarize their findings, which were then reinforced through practical exercises to solidify and apply their newly acquired knowledge.

3.2. Activities Using Sketch Engine for direct DDL Sessions: *zhíyè* ‘occupation’ and *hángyè* ‘industry’

The preparatory work for direct DDL is similar to the indirect method, the difference lies in the fact that instructors do not need to create paper-based materials; instead, they allow learners to directly operate corpora. In this study, the Sketch Engine was employed as the instructional tool (Kilgarriff et al., 2004). Before achieving the highest efficacy in the classroom, instructors must familiarize themselves with the usage of Sketch Engine and annotate each step to design classroom activities (Chang, 2022).

For illustration, we take *zhíyè* ‘occupation’ and *hángyè* ‘industry’ as examples. In the learner corpora, the authors observed that learners did not tend to misinterpret *zhíyè* as *hángyè*, but they did make the error of using *hángyè* in place of *zhíyè*, as shown in examples (3) and (4).

- (3) *Wǒ cóng wǎnglùshàng kàndào nǐmen de zhāopìn, wǒ hěn yǒu xìngqù, tèbié duì dǎoyóu zhè yī *hángyè*.
(I saw your job advertisement on the internet and I am interested, particularly in the field of tour guiding.)
- (4) *Yīncǐ, wǒ xīwàng néng yìngzhēng guǎnggào shèjìshī zhè yī fèn *hángyè*, wéi guīgōngsī bànrì.
(Hence, I aspire to secure the role of an advertisement designer in your esteemed company.)

In addition to this finding, the authors utilized the Corpus Of Contemporary Chinese (COCT) to further analyze the critical differences between the two words. Furthermore, the focus of instruction was placed on highlighting that *zhíyè* differed from *hángyè* in that, despite both being nouns, *zhíyè* possessed the additional function of being used as an attributive modifier meaning ‘professional’.

In the classroom, the tasks assigned to students offered clear steps and objectives for their observations. Examples of tasks were as follows:

Task 1: Identify the three most frequent collocations for *zhíyè* and *hángyè*.

The activity involved guiding learners to use concordance functions to observe the usage of *zhíyè* and *hángyè* separately. The instructor employed questioning techniques to facilitate the learners' observations, such as “What are the three most common words that frequently appear together with *zhíyè*?” Similarly, learners were prompted to identify the three most frequently associated terms with *hángyè* and the typical words occurring on its left and right sides. The objective of the instruction was to enable learners to ultimately deduce that *zhíyè* can serve as an attributive modifier for nouns, such as ‘athlete, actor, soldier, assassin,’ etc., expressing the meaning of ‘professional’ or ‘specialized’. On the other hand, *hángyè* did not possess such a usage.

Task 2: Use the Thesaurus Function to differentiate *zhíyè* and *hángyè*.

The second instructional activity involved guiding learners to access the Thesaurus Function and input *zhíyè* and *hángyè* separately. Through the utilization of visualizations and diagrams, learners were led to discover the semantic differences between the two terms. The innermost circle represented words closely related to key concepts, while the outer circles indicated more distant associations. As depicted in Figure 1, the meaning of *zhíyè* is closely related to terms such as skills, and profession. Figure 2 reveals that the terms closely associated with the semantic context of *hángyè* include sector, enterprise, field, and market, etc. By completing the tasks,

learners were guided towards concluding that *zhíyè* and *hángyè* referred to different meanings, and *zhíyè* can be an attributive often to modify the word on the right.

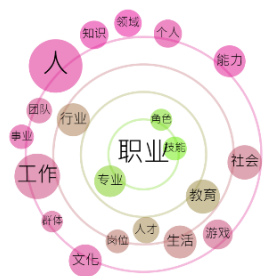


Figure 1. Thesaurus result of *zhíyè*

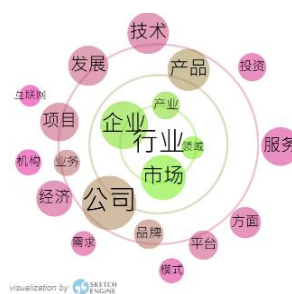


Figure 2. Thesaurus result of *hángyè*

4. Results

The average score for the pretest was 71%. The posttest had an overall average score of 95%, indicating a 24% improvement compared to the pretest. The Wilcoxon signed-rank test showed a statistically significant improvement in applying DDL to learn confusable words (p -value = 0.039).

The results of questionnaires and interviews were divided into three aspects. First, in the realm of learning preferences learners exhibited a pronounced inclination toward the concept of collocation, deeming it advantageous to their learning endeavors. Furthermore, they expressed a preference for the indirect approach due to its perceived impact on learning efficiency and its ability to bolster their confidence in the analysis of genuine linguistic data.

Second, in the evaluation of the course learners provided positive evaluations for it, and they perceived the guidance provided by the teacher in the classroom as highly significant. Learners placed significant emphasis on the pivotal roles of Hypothesis Formulation and Hypothesis Confirmation in both indirect and direct DDL methods. However, there existed varying opinions regarding the instructional arrangement of teaching a set of confusable words within a one-hour session, which may be attributed to the learners' individual learning strategies and personal learning styles.

Third, the experiences of learners using the tool and their interview responses can be summarized as follows. The majority of learners considered the corpus a valuable reference tool; however, learners did not achieve proficiency in utilizing the functions within Sketch Engine after five instructional sessions. It was suggested in the future to contemplate adopting an integrated approach to prevent learner fatigue and mitigate its impact on learning efficiency. For example, an integrated 50-minute class structure could encompass 30 minutes dedicated to indirect DDL, followed by 20 minutes allocated to direct DDL.

5. Conclusions

In light of learners' expressed preference for the indirect DDL, it is advisable for instructors aiming to implement DDL instruction to initiate their pedagogical endeavors with the indirect DDL approach. This initial step serves the purpose of acquainting learners with the five-step procedural framework, encompassing observation and analysis, hypothesis formulation, hypothesis revision, summarization, and consolidation and application. The arrangement can release learners' pressure to learn the corpus skills and data-driven steps simultaneously. Such an approach serves to enhance learners' proficiency and confidence in the analysis of linguistic concordance before progressing to the direct DDL method. Moreover, instructors should allocate ample time for constructive discussions and reflective activities to facilitate a comprehensive and enriched learning experience.

The study aimed to equip instructors with effective instructional design strategies for using DDL to teach Chinese confusable words, thereby enhancing the teaching and learning efficiency of these words. Despite the study's small sample size, the mitigation of this limitation was achieved through the application of a

comprehensive course design and rigorous research methodology. Based on the post-test results, it was evident that learners, following instruction on confusable word pairs, exhibited an average pass rate improvement of 24%. In light of these course outcomes and learners' feedback, this study recommends that educators explore the application of data-driven learning in various facets of language instruction, including vocabulary, grammar, and writing pedagogy.

Acknowledgements

We would like to thank the research support from the National Science and Technology Council (NSTC 112-2410-H-002-060), with assistance from Ms. Chun-ting Chou in conducting statistical significance tests.

References

- Boulton, A. (2010). Data-driven learning: Taking the computer out of the equation. *Language learning*, 60(3), 534-572.
- Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning*, 67(2), 348-393. <https://doi.org/10.1111/lang.12224>
- Chang, Li-ping. (2022). The preliminary study of corpus literacy training for in-service Chinese language teachers. *Journal of Chinese Language Teaching*, 19(4), 83-124.
- Chujo, K., Anthony, L. and Oghigian, K. (2009). DDL for the EFL classroom: Effective uses of a Japanese-English parallel corpus and the development of a learner- friendly, online parallel concordancer. In M. Mahlberg, V. González-Díaz, and C. Smith (Eds.), *Proceedings of 5th Corpus Linguistics Conference 2009*, University of Liverpool, UK. <http://ucrel.lancs.ac.uk/publications/cl2009>
- Johns, T. (1990). From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. *CALL Austria Newsletter* (July 1990) (pp. 14-34).
- Kilgarriff, A., Rychlý, P., Smrž, P., & Tugwell, D. (2004, July 6-10). The Sketch Engine. Paper presented at XI EURALEX International Congress, Lorient, France.
- Lee, H., Warschauer, M., & Lee, J. H. (2019). The effects of corpus use on second language vocabulary learning: A multilevel meta-analysis. *Applied Linguistics*, 40(5), 721-753.
- Louw, B. (2000). Contextual prosodic theory: Bringing semantic prosodies to life. In C. Heffer and H. Sauntson (Eds.), *Words in context: In honour of John Sinclair* (pp. 48-94). University of Birmingham, Birmingham.
- Mizumoto, A., & Chujo, K. (2015). A meta-analysis of data-driven learning approach in the Japanese EFL classroom. *English Corpus Studies*, 22, 1-17.
- Smith, S. (2011). Corpus-based tasks for learning Chinese: a data-driven approach. *The Asian Conference on Technology in the Classroom Official Conference Proceedings 2011* (pp. 48-59).
- Wang, P., Hsu, C., Long, S. & Liles, X. (2020). Designing Data-Driven Learning Activities for the Chinese as a Second Language Classroom. *Journal of Chinese Language Teaching*, 17(3), 103-137.
- Xiao, R., & McEnery, T. (2006). Collocation, semantic prosody, and near synonymy: A cross-linguistic perspective. *Applied Linguistics*, 27(1), 103-129.
- Yeh, M., & Zhang, X. (2018). Corpus-based instruction: Teaching discourse-linking jiu (就) in storytelling. *Chinese as a Second Language*, 53(1), 1-23.
- Yoon, H. & Hirvela, A. (2004) ESL student attitudes toward corpus use in L2 writing. *Journal of Second Language Writing*, 13(4), 257-283.