

International Journal of

Computational Linguistics & Chinese Language Processing

中文計算語言學期刊

A Publication of the Association for Computational Linguistics and Chinese Language Processing

This journal is included in THCI, Linguistics Abstracts, and ACL Anthology.

Special Issue on “Chinese as a Foreign Language”

Guest Editors: Lung-Hao Lee, Liang-Chih Yu, and Li-Ping Chang

易繫辭曰上古結繩而
治後世聖人易之以書
契百官以治萬民以察
說文敘曰蓋文字者經
藝之本宣教明化之始
前人所以垂後後人所
以識古故曰本立而道
生知天下之至蹟而不
可亂也教化既萌文心
雕龍則謂人之立言因
字而生句積句而成章
積章而成篇篇之彪炳

Vol.20

No.1

June 2015

ISSN: 1027-376X

International Journal of Computational Linguistics & Chinese Language Processing

Advisory Board

- Jason S. Chang*
National Tsing Hua University, Hsinchu
- Hsin-Hsi Chen*
National Taiwan University, Taipei
- Keh-Jiann Chen*
Academia Sinica, Taipei
- Sin-Horng Chen*
National Chiao Tung University, Hsinchu
- Eduard Hovy*
University of Southern California, U. S. A.
- Chu-Ren Huang*
The Hong Kong Polytechnic University, H. K.
- Jian-Yun Nie*
University of Montreal, Canada
- Richard Sproat*
University of Illinois at Urbana-Champaign, U. S. A.
- Keh-Yih Su*
Behavior Design Corporation, Hsinchu
- Chiu-Yu Tseng*
Academia Sinica, Taipei
- Jhing-Fa Wang*
National Cheng Kung University, Tainan
- Kam-Fai Wong*
Chinese University of Hong Kong, H.K.
- Chung-Hsien Wu*
National Cheng Kung University, Tainan

Editorial Board

- Yuen-Hsien Tseng (Editor-in-Chief)*
National Taiwan Normal University, Taipei
- Kuang-hua Chen (Editor-in-Chief)*
National Taiwan University, Taipei
- Speech Processing**
- Yuan-Fu Liao (Section Editor)*
National Taipei University of Technology,
Taipei
- Berlin Chen*
National Taiwan Normal University, Taipei
- Hung-Yan Gu*
National Taiwan University of Science and
Technology, Taipei
- Hsin-Min Wang*
Academia Sinica, Taipei
- Yih-Ru Wang*
National Chiao Tung University, Hsinchu
- Linguistics & Language Teaching**
- Shu-Kai Hsieh (Section Editor)*
National Taiwan University, Taipei
- Hsun-Huei Chang*
National Chengchi University, Taipei
- Hao-Jan Chen*
National Taiwan Normal University, Taipei
- Huei-ling Lai*
National Chengchi University, Taipei
- Meichun Liu*
National Chiao Tung University, Hsinchu
- James Myers*
National Chung Cheng University, Chiayi
- Shu-Chuan Tseng*
Academia Sinica, Taipei
- Information Retrieval**
- Ming-Feng Tsai (Section Editor)*
National Chengchi University, Taipei
- Chia-Hui Chang*
National Central University, Taoyuan
- Chin-Yew Lin*
Microsoft Research Asia, Beijing
- Show-De Lin*
National Taiwan University, Taipei
- Wen-Hsiang Lu*
National Cheng Kung University, Tainan
- Shih-Hung Wu*
Chaoyang University of Technology, Taichung
- Natural Language Processing**
- Richard Tzong-Han Tsai (Section Editor)*
Yuan Ze University, Chungli
- Lun-Wei Ku*
Academia Sinica, Taipei
- Chuan-Jie Lin*
National Taiwan Ocean University, Keelung
- Chao-Lin Liu*
National Chengchi University, Taipei
- Jyi-Shane Liu*
National Chengchi University, Taipei
- Liang-Chih Yu*
Yuan Ze University, Chungli

Executive Editor: *Abby Ho*

English Editor: *Joseph Harwood*

The Association for Computational Linguistics and Chinese Language Processing, Taipei

International Journal of

Computational Linguistics & Chinese Language Processing

Aims and Scope

International Journal of Computational Linguistics and Chinese Language Processing (IJCLCLP) is an international journal published by the Association for Computational Linguistics and Chinese Language Processing (ACLCLP). This journal was founded in August 1996 and is published four issues per year since 2005. This journal covers all aspects related to computational linguistics and speech/text processing of all natural languages. Possible topics for manuscript submitted to the journal include, but are not limited to:

- Computational Linguistics
- Natural Language Processing
- Machine Translation
- Language Generation
- Language Learning
- Speech Analysis/Synthesis
- Speech Recognition/Understanding
- Spoken Dialog Systems
- Information Retrieval and Extraction
- Web Information Extraction/Mining
- Corpus Linguistics
- Multilingual/Cross-lingual Language Processing

Membership & Subscriptions

If you are interested in joining ACLCLP, please see appendix for further information.

Copyright

© The Association for Computational Linguistics and Chinese Language Processing

International Journal of Computational Linguistics and Chinese Language Processing is published four issues per volume by the Association for Computational Linguistics and Chinese Language Processing. Responsibility for the contents rests upon the authors and not upon ACLCLP, or its members. Copyright by the Association for Computational Linguistics and Chinese Language Processing. All rights reserved. No part of this journal may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical photocopying, recording or otherwise, without prior permission in writing form from the Editor-in Chief.

Cover

Calligraphy by Professor Ching-Chun Hsieh, founding president of ACLCLP

Text excerpted and compiled from ancient Chinese classics, dating back to 700 B.C.

This calligraphy honors the interaction and influence between text and language

Contents

Special Issue Articles: Chinese as a Foreign Language

Guest Editorial: Special Issue on Chinese as a Foreign Language. i
Lung-Hao Lee, Liang-Chih Yu, and Li-Ping Chang
Guest Editors

Papers

HANSpeller: A Unified Framework for Chinese Spelling
Correction..... 1
*Jinhua Xiong, Qiao Zhang, Shuiyuan Zhang, Jianpeng Hou, and
Xueqi Cheng*

A Study on Chinese Spelling Check Using Confusion Sets and
N-gram Statistics..... 23
Chuan-Jie Lin, and Wei-Cheng Chu

Automatically Detecting Syntactic Errors in Sentences Written by
Learners of Chinese as a Foreign Language..... 49
Tao-Hsing Chang, Yao-Ting Sung, and Jia-Fei Hong

Automatic Classification of the “De” Word Usage for Chinese
as a Foreign Language..... 65
Jui-Feng Yeh, Chan-Kun Yeh

以「華語學習者語料庫」為本的「了」字句偏誤分析 [The Error
Analysis of “Le” Based on “Chinese Learner Written Corpus”]... 79
*董子昀(Tzu-Yun Tung), 陳浩然(Howard Hao-Jan Chen),
楊惠媚(Hui-Mei Yang)*

Cross-Linguistic Error Types of Misused Chinese Based on
Learners’ Corpora..... 97
Keiko Mochizuki, Hiroshi Sano, Ya-Ming Shen and Chia-Hou Wu

Guest Editorial:

Special Issue on Chinese as a Foreign Language

Lung-Hao Lee*, Liang-Chih Yu⁺, and Li-Ping Chang[#]

Abstract

This introduction paper describes the research trends of Chinese as a second/foreign language along with related studies. We also overview the research papers included in this special issue. Finally, we conclude the findings and offer the suggestions.

Keywords: Computer-Assisted Language Learning, Second Language Acquisition, Learner Corpora, Interlanguage, Mandarin Chinese.

1. Introduction

China's growing global influence has prompted a surge of interest in learning Chinese as a Foreign Language (CFL) and this trend is expected to continue. However, whereas many computer-assisted learning tools have been developed for learning English, support for CFL learners is relatively sparse, especially in terms of tools designed to automatically evaluate learners' responses. For example, while Microsoft Word has integrated robust English spelling and grammar checking functions for years, such tools for Chinese are still quite primitive. Another trend in demanding automated tools for CFL learners is accelerated by the recent progress in online learning technology and platforms, especially the so called MOOC (Massive Open Online Course) where a huge number of learners can enroll in a course. The MOOC idea and platform not only make more people acquainted with online courses, but also demand automatic technology to handle the large volume of assignments and tests that are submitted by the enrolled learners.

* Department of Computer Science and Information Engineering, National Taiwan University, Taiwan
E-mail: lhlee@nlg.csie.ntu.edu.tw

⁺ Department of Information Management & Innovation Center for Big Data and Digital Convergence
Yuan Ze University, Taiwan
E-mail: lcyu@saturn.yzu.edu.tw

[#] Mandarin Training Center, National Taiwan Normal University, Taiwan
E-mail: lchang@ntnu.edu.tw

The author for Correspondence is Li-Ping Chang.

In contrast to the booming research developments for learning English as a foreign language, relatively few studies and tools are available for CFL learners. Chang (1995) proposed a three-step approach that uses the whole context within a sentence for spelling correction. Similar to Chang's approach, Zhang *et al.* (2000) presented an approximate word-matching algorithm to detect and correct Chinese spelling errors with the help of three edit operations: character substitution, insertion, and deletion. Ren *et al.* (2001) tried a hybrid approach that combines a rule-based method and a probability-based method to check Chinese spelling errors. Huang *et al.* (2007) proposed a learning model based on Chinese phonemic alphabet for spelling error check. Wu *et al.* (2010) proposed relative position and parse template language models to detect Chinese errors written by US learners using the NCKU corpus. Yu & Chen (2012) proposed a classifier to detect word-ordering errors in Chinese sentences from HSK learner corpus. Chang *et al.* (2012) proposed a penalized probabilistic First-Order Inductive Learning (pFOIL) algorithm, which integrates Inductive Logic Programming (ILP), First-Order Inductive Learning (FOIL), and a penalized log-likelihood function for error diagnosis. Lee *et al.* (2013) handcrafted a set of linguistic rules with syntactic information to detect grammatical errors. Lee *et al.* (2014) developed a sentence judgment system using both rule-based and n-gram statistical methods to detect grammatical errors in Chinese sentences.

In addition to research papers, several workshops and shared tasks focused on Chinese learning have been organized. For example, Chinese spelling check bakeoffs were organized in annual SIGHAN workshops, that is, the first one was held as part of the SIGHAN-7 in IJCNLP 2013 (Wu *et al.*, 2013); the second version was held in CIPS-SIGHAN joint CLP-2014 conference (Yu *et al.*, 2014); the third evaluation will be held in SIGHAN-8 as a ACL-IJCNLP 2015 workshop (Tseng *et al.*, 2015). The research community has also organized a series of workshops on Natural Language Processing Techniques for Educational Applications (NLP-TEA) to give special attention to researches that have taken computer-assisted Asian language learning into consideration. The first NLP-TEA workshop was held in conjunction with ICCE-2014, accompanying with a shared task on Chinese as a Foreign Language was organized (Yu *et al.*, 2014). The second NLP-TEA will be held as one of ACL-IJCNLP 2015 workshops with a Chinese Grammatical Error Diagnosis shared task (Lee *et al.*, 2015). In summary, all of these academic activities increase the visibility of Chinese educational application research in the NLP community.

This special issue aims at general topics related to CFL research. Topics of interest include, but are not limited to as follows. From engineering perspectives, computer-assisted techniques for Chinese learning are important, such as spelling error check, grammatical error correction, sentence judgment systems, automated essay scoring, educational data mining, and so on. From linguistic perspectives, research areas include second language acquisition and

interlanguage analysis by using learner corpora.

In the rest of this introduction paper, we describe the research paper included in this special issue in Section 2. Finally, we conclude the findings accompanying with suggestions in Section 3.

2. Content of Special Issue

This special issue consists of six research papers, which were reviewed and recommended by at least two experts. We briefly describe them as follows.

The first paper “HANSpeller: A Unified Framework for Chinese Spelling Correction” proposes a framework based on an extended Hidden Markov model and the ranker-based models, along with a rule-based model for Chinese spelling error detection and correction. CLP-2014 CSC datasets are adopted to demonstrate promising performance of their approach.

The second paper “A Study on Chinese Spelling Check Using Confusion Sets and N-gram Statistics” expands the coverage of confusion sets using Shuowen Jiezi and the Four-Corner codes. They also build a two-character confusion set. N-gram statistics are applied with the help of expanded and constructed confusion sets for Chinese spelling error checking. Experimental results show the approach improves the performance achieving by their previous system on SIGHAN 2013 CSC Bake-off.

The third paper “Automatically Detecting Syntactic Errors in Sentences Writing by Learners of Chinese as a Foreign Language” describes how to detect Chinese grammatical errors based on automatically-generated and manually-handcrafted rules. They propose a KNGED algorithm to identify syntactic errors written by CFL learners. NLP-TEA CFL datasets are used to show the effectiveness of their approach.

The fourth paper “Automatic Classification of the “De” Word Usage for Chinese as a Foreign Language” focuses on the usage of morphosyntactic particle “De”. LEM 2 algorithm is adopted for deriving the rule set and then classifying the {的, 得, 地} based on induced rules for correct usages. The method achieves good performance on NLP-TEA CFL datasets.

The fifth paper “The Error Analysis of “Le” Based on Chinese Learner Written Corpus” analyzes the usage and the error types of “Le” made by English-native learners at the beginning and intermediate level based on NTNU learner corpus. The error types include redundancy and mis-selection of *le1*, *le2* and *le(1+2)*. Their findings show *le1* is the most commonly spotted error type, and there is a large number of “*le1*” and “*le(1+2)*” redundant usages. In addition, pedagogical suggestions are also provided.

The sixth paper “Cross-Linguistic Error Types of Misused Chinese Based on Learners’ Corpora” presents the construction of a learner corpus named ‘Full Moon Corpus’ and the tagging system for error annotation. The authors use comparative analysis method to observe

the “*yi* ‘one’ + classifier” phrase by English-native learners and Japanese-native learners and discuss the reasons of ‘overuse’ and ‘underuse’ phenomenon.

3. Conclusions

This paper describes the present research trends of Chinese as a foreign/second language. All research papers included in this special issue are also introduced.

To improve the performance of NLP tools for Chinese learning by machine learning, collecting real learners’ erroneous sentences as much as possible is a challenging issue. The coverage of erroneous types is another. And tagging different corpora using the same format and tag set for learner corpus development is the other difficulty. The best strategy to deal with these problems may be to ally with research teams and to share collected linguistic resources.

Acknowledgments

We would like to thank all of the authors for their support, and our special thanks go to the anonymous reviewers who contributed their valuable wisdom and time to the research community.

References

- Chang, C.-H. (1995). A new approach for automatic Chinese spelling correction. In *Proceedings of Natural Language Processing Pacific Rim Symposium*, 278-283.
- Chang, R.-Y., Wu, C.-H., & Prasetyo, P. K. (2012). Error diagnosis of Chinese sentences using inductive learning algorithm and decomposition-based testing mechanism. *ACM Transactions on Asian Language Information Processing*, 11(1), Article 3 (30 pages).
- Huang, C.-M., Wu, M.-C., & Chang, C.-C. (2007). Error detection and correction based on Chinese phonemic alphabet in Chinese text. In *Proceedings of the 4th Conference on Modeling Decisions for Artificial Intelligence*, 463-476.
- Lee, L.-H., Chang, L.-P., Lee, K.-C., Tseng, Y.-H., & Chen, H.-H. (2013). Linguistic rules based Chinese error detection for second language learning. In *Work-in-Progress Poster Proceedings of 21st International Conference on Computers in Education*, 27-29.
- Lee, L.-H., Yu, L.-C., & Chang, L.-P. (2015). Overview of shared task on Chinese grammatical error diagnosis. In *Proceedings of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications*.
- Lee, L.-H., Yu, L.-C., Lee, K.-C., Tseng, Y.-H., Chang, L.-P., & Chen, H.-H. (2014). A sentence judgment for grammatical error detection. In *Proceedings of the 25th International Conference on Computational Linguistics: System Demonstrations*, 67-70.

- Ren, F., Shi, H., & Zhou, Q. (2001). A hybrid approach to automatic Chinese text checking and error correction. In *Proceedings of 2001 IEEE International Conference on Systems, Man, and Cybernetics*, 1693-1698.
- Tseng, Y.-H., Lee, L.-H., Chang, L.-P., & Chen, H.-H. (2015). Introduction to SIGHAN 2015 bake-off for Chinese spelling check. In *Proceedings of the 8th SIGHAN Workshop on Chinese Language Processing*.
- Wu, C.-H., Liu, C.-H., Harris, M. & Yu, L.-C. (2010). Sentence correction incorporating relative position and parse template language model. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6), 1170-1181.
- Wu, S.-H., Liu, C.-L., & Lee, L.-H. (2013). Chinese spelling check evaluation at SIGHAN bake-off 2013. In *Proceedings of the 7th SIGHAN Workshop on Chinese Language Processing*, 35-42.
- Yu, C.-H., & Chen, H.-H. (2012). Detecting word ordering errors in Chinese sentences for learning Chinese as a foreign language. In *Proceedings of the 24th International Conference on Computational Linguistics*, 3003-3018.
- Yu, L.-C., Lee, L.-H., & Chang, L.-P. (2014). Overview of grammatical error diagnosis for learning Chinese as a foreign language. In *Proceedings of the 1st Workshop on Natural Language Processing Techniques for Educational Applications*, 42-47.
- Yu, L.-C., Lee, L.-H., Tseng, Y.-H., & Chen, H.-H. (2014). Overview of SIGHAN 2014 bake-off for Chinese spelling check. In *Proceedings of the 3rd CIPS-SIGHAN Joint Conference on Chinese Language Processing*, 126-132.
- Zhang, L., Huang, C., Zhou, M., & Pan, H.-H. (2000). Automatic detecting/correcting errors in Chinese text by an approximate word-matching algorithm. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, 248-254.

HANSpeller: A Unified Framework for Chinese Spelling Correction

Jinhua Xiong*, Qiao Zhang*+, Shuiyuan Zhang*+,

Jianpeng Hou*+ and Xueqi Cheng*

Abstract

The number of people learning Chinese as a Foreign Language (CFL) has been booming in recent decades. The problem of spelling error correction for CFL learners increasingly is becoming important. Compared to the regular text spelling check task, more error types need to be considered in CFL cases. In this paper, we propose a unified framework for Chinese spelling correction. Instead of conventional methods, which focus on rules or statistics separately, our approach is based on extended HMM and ranker-based models, together with a rule-based model for further polishing, and a final decision-making step is adopted to decide whether to output the corrections or not. Experimental results on the test data of foreigner's Chinese essays provided by the SIGHAN 2014 bake-off illustrate the performance of our approach.

Keywords: Chinese Spelling Correction, HMM, Ranker-Base Model, Rule-based Model, Decision-making.

1. Introduction

Recent studies have shown that Chinese has become a popular choice for a second language among international college students. More and more people are learning Chinese as a Foreign Language (CFL). It is very difficult, however, for CFL learners to master Chinese because of the intrinsic linguistic features of the Chinese language. When CFL learners write Chinese essays, they are prone to generating a greater number and more diversified spelling errors than native language learners. Therefore, spelling correction tools to support such learners in

* Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
E-mail: xjh@ict.ac.cn, {zhangqiao, zhangshuiyuan}@software.ict.ac.cn

+ University of Chinese Academy of Sciences, Beijing, China

correcting and polishing their Chinese essays is valuable and necessary. For the English language, there are many editing tools that provide spelling check functionality, *e.g.* Microsoft Word's spellchecker. For the Chinese language, however, such tools cannot be found until now.

Spelling correction has been studied for many years on regular text and web search queries. Although these two tasks share many common techniques, they have different concerns. Compared to techniques of web search query spelling correction, where corrections should be presented to search engine users in real-time, more complicated techniques can be applied to spelling correction on regular text to improve the performance, as such a situation has a lower real-time requirement.

In spelling correction of Chinese essays of CFL learners, we face more challenges because of the uniqueness of the Chinese language.

1) Chinese corpora for spelling correction, especially publicly available ones, are rare, compared with English corpora. This impedes work on this practical topic.

2) There are no natural delimiters, such as spaces, between Chinese words, which may result in errors in words splitting, which may cause more splitting errors.

3) The number of error types is more than that of other cases, because CFL learners are prone to different kinds of errors that we cannot imagine as native speakers. There are four major error types that confuse people, as illustrated in Table 1.

Table 1. Examples of spelling error types

Error Types	Misspelled	Corrections
Homophone	一籌莫展 年年有魚 聯合國公布	一愁莫展 年年有餘 聯合國公佈
Near-homophone	好碼差不多一樣	號碼差不多一樣
Similar shape	列如：家庭會變冷漠 如火如荼	例如：家庭會變冷漠 如火如荼
Other errors	每個禮拜 1、3、5	每個禮拜一、三、五
	受了都少苦	受了多少苦
	持續的发展	持續地发展

The first type is the misuse of homophone, which means learners choose the wrong characters with same pronunciation but different meanings. For example, “一籌莫展” may be misspelled as “一愁莫展”. Herein, the second character “籌” is misspelled as “愁,” both with

the same pronunciation (chóu). Another example is “年年有**鱼**” (There will be **fish** every year), which is homophonous with “年年有**余**” (There will be **surpluses** every year). One should take context into account when judging this type of homophone. A single syllable may also have a range of different meanings. The Cihai dictionary lists 149 Chinese characters representing the syllable "yì".

Second, there is the near-homophone error, which means the pronunciations of chosen words are very similar. For CFL learners, difference in diacritical markings may be not enough to distinguish. For example, there is a problem in discriminating pronunciation of the first character in the following sentences, “好**码**差不多一样” and “号**码**差不多一样”.

Besides, some graphically similar Chinese characters are confusing, due to their similar shape. They differ only in subtle aspects. To distinguish between these characters, many aspects, such as sound, meaning, and collocations, should be taken into account. If you do not look carefully, you can hardly distinguish them, *e.g.* “如火如**茶**” and “如火如**荼**,” where the first one is correct, and the second one is wrong.

Finally, some error types usually are caused by grammar rules of Chinese, such as the usage of three confusable words “的,” “地,” and “得”. Moreover, the last two words connect with two different pronunciations in different contexts. Therefore, checking correctness of the usage of these three words is difficult.

The direct reason why these error types are always encountered by CFL learners is that Chinese spelling is not phonetic and each word in a Chinese phrase has its specific meaning. Meanwhile, some other error types can be caused by various Chinese input methods.

4) The Chinese language is continuously evolving. Therefore, correction only based on static corpora is not enough. For example, traditional Chinese and simplified Chinese may have different choices for the same word. In some cases, it is very difficult to distinguish them. Thus, web-based high-quality resources should be considered for decision-making on spelling correction.

To address the above challenges, we propose a unified framework, named HANSpeller, for Chinese essay spelling error detection and correction. Our method combines different methods to improve performance. The main contributions are as follows. (1) An HMM-based approach is used to segment sentences and generate candidates for sentence spelling corrections. (2) Under the unified framework, all kinds of error types can be integrated for candidate generation. We collect some error types that can only be found in CFL learner essays and add them into the candidate generation process. (3) In order to address evolving features of the Chinese language, an online high-quality corpus is collected for training and decision-making and online search engine results also are used in the ranking stage of our model, which can also improve the performance significantly.

The rest of the paper is organized as follows. We discuss related works in Section 2, and we introduce our unified framework approach in Section 3, where we focus on the basic processes of our method. In Section 4, we present the detailed setup of the experimental evaluation and the results of the experiment. Finally, in Section 5, we conclude the paper and explore future directions.

2. Related works

The study of spelling correction has a long history (Kukich, 1992). It is aimed at identifying misspellings and choosing optimal words as suggested corrections. In other words, it contains two subtasks that involve spelling error detection and spelling error correction. In early research, the spelling corrections were mainly devoted to solving non-word errors; such errors were often caused by insertion, deletion, substitution, and transposition of letters in a valid word that result in an unknown word. A common strategy at that time was to rely on a word dictionary or some rules like Levenshtein distance (Levenshtein, 1966). Mangu and Brill (1997) proposed a transition-based learning method for spelling correction. Their methods generated three types of rules from training data, which constructed a high performance and concise system for English.

In these methods, however, the dictionaries and rules were always constructed manually, leading to very high cost. Therefore, statistics generative models were introduced for spelling correction, which made spelling correction step into a new stage. The error model and n-gram language model are two important models (Brill & Moore, 2000). Atwell and Elliott (1987) used n-gram and part-of-speech language models for spelling corrections. Mays *et al.* (1991) used word-trigram probabilities for detecting and correcting real word errors. Brill *et al.* (2000) proposed a new channel model for spelling correction, based on generic string to string edits.

With the development of the Internet, the research and technology on query spelling correction for search engines has been studied intensively. The task of web-query spelling correction shares a lot of technology with traditional spelling correction, but it is more difficult. First, the spelling correction task is faced with more error types, as all kinds of errors may occur in a web environment. In addition, search queries consist of some key words rather than sentences, making some sentence-based methods achieve poor performance. Therefore, many novel ideas have been proposed by researchers. Cucerzan and Brill (2004) presented an iterative process for query spelling check, using a query log and trust dictionary. There, the noisy channel model was used to choose the best correction. Ahmad and Kondrak (2005) used the search query logs to learn a spelling error model, which improves the quality of query spelling check. Li *et al.* (2006) applied a distributional similarity based model for query spelling correction. Gao *et al.* (2010) presented a large-scale ranker-based system for search spelling correction, where the ranker uses web-scale language models and many kinds of

features for better performance, including: surface-form similarity, phonetic-form similarity, entity, dictionary, and frequency features. Suzuki and Gao (2012) proposed a transliteration based character method using an approach inspired by the phrase-based statistical machine translation framework and attained good performance in online spelling correction.

Furthermore, Google and Microsoft have developed some application interfaces for checking spelling. Google (2010) has developed a Java API for a Google spelling check service. Microsoft (2010) provides a web n-gram service.

The above works mainly target the task of English spelling correction. As to Chinese spelling correction, the situation is quite different because English words are separated naturally by spaces, while Chinese words are not. This nature of Chinese makes correction much more difficult than that of English. An early work was by Chang (1995), which used a character dictionary of similar shape, pronunciation, meaning, and input-method-code to deal with the spelling correction task. The system replaced each character in the sentence with a similar character in the dictionary and calculated the probability of all modified sentences based on language model. Zhang (2000) introduced a method that can handle not only Chinese character substitution, but also insertion and deletion errors. They distinguished the way of matching between Chinese and English, thereby largely improving the performance over the work of Chang (1995). Hung and Wu (2008) introduced a method that used manually edited error templates to correct errors. Zheng *et al.* (2011) found the fact that, when people type Chinese Pinyin, there are several wrong types. Then, they introduced a method based on a generative model and the input wrong types to correct spelling errors. Liu *et al.* (2011) pointed out that visually and phonologically similar characters are major factors for errors in Chinese text. Thus, by defining appropriate similarity measures that consider extended Cangjie codes, visually similar characters can be quickly identified.

Some Chinese spelling checkers have also incorporated word segmentation techniques. Huang *et al.* (2007) used a word segmentation tool (CKIP) to generate correction candidates before detecting Chinese spelling errors. Hung and Wu (2009) segmented the sentence using a bigram language model. In addition, they combined a confusion set and some error templates to improve the results. Chen and Wu (2010) modified the system on the basis of Huang and Wu (2009) using statistic-based methods and a template matching module.

In addition, a hybrid approach has been applied to Chinese spelling correction. Chang *et al.* (2012) used an inductive learning algorithm in Chinese spelling error classification and got better performance than C4.5, maximum entropy, and Naive Bayes classifiers. Hao *et al.* (2013) proposed a Tri-gram modeled-Weighted Finite-State Transducer method integrating confusing-character table, beam search, and A* to correct Chinese text errors. Jin *et al.* (2014) integrated three models, including an n-gram language model, a pinyin based language model, and a tone based language model, to improve the performance of a Chinese checking spelling

error system.

Chinese essay spelling correction as a special kind of spelling correction research effort has been promoted by efforts, such as the SIGHAN bake-offs (Yu *et al.*, 2014; Wu *et al.*, 2013). Huang *et al.* (2014) used a tri-gram language model to detect and correct spelling errors. They also employed a dynamic algorithm and smoothing method to improve the efficiency. Chu and Lin (2014) used a word replacement strategy to generate candidates based on the expanded confusion set. Then, a rule-based classifier and SVM-based classifier were used to locate and correct errors. Gu *et al.* (2014) proposed two systems to solve the Chinese spelling check problem. One was built based on a CRF model, and the other was based on 2-Chars and 3-Chars model. Their experimental results showed that the latter model was better.

Chiu *et al.* (2013) divided the correction task into two subtasks to solve. They used word segmentation to find errors and combined machine translation model to translate the wrong sentences into the appropriate ones. Hsieh *et al.* (2013) developed two error detection systems based on CKIP word segmentation tool and Google 1T uni-gram data, respectively. Jia *et al.* (2013) proposed a single source shortest path algorithm based on the graph model to correct spelling errors.

In our system, we need to detect and correct spelling errors on Chinese essays that always are written by CFL learners. It has some different concerns with query text or query spelling correction. Noting that spelling correction methods require lexicons and/or language corpora, we adopt the method based on statistics combined with lexicon and rule-based methods.

3. A Unified Framework for Chinese Spelling Correction

In this section, we present a unified framework, named HANSpeller, for Chinese spelling correction based on extended HMM and ranking models. The major idea of our approach is to model the spelling correction process as a ranking and decision-making problem.

Figure 1 shows the whole outlined architecture of HANSpeller. It separates the Chinese spelling correction system into four major steps. First is to use the extended HMM model to generate the top-k candidates for the sentences being checked. Then, a ranking algorithm is applied to re-rank the correction candidates for later decision. The third step conducts rule-based analysis for a specific correction task, *e.g.* the correction rule of the usage of three confusable words “的,” “地,” and “得”. Finally, the system makes decision whether to output the original sentence directly or correction results based on the previous output and global constrains.

This framework provides a unified approach for spelling correction tasks, which can be regarded as a language independent framework and can be tailored to different scenarios. To

move to another scenario, you need to prepare a language related corpus, but you do not need to be an expert in that language.

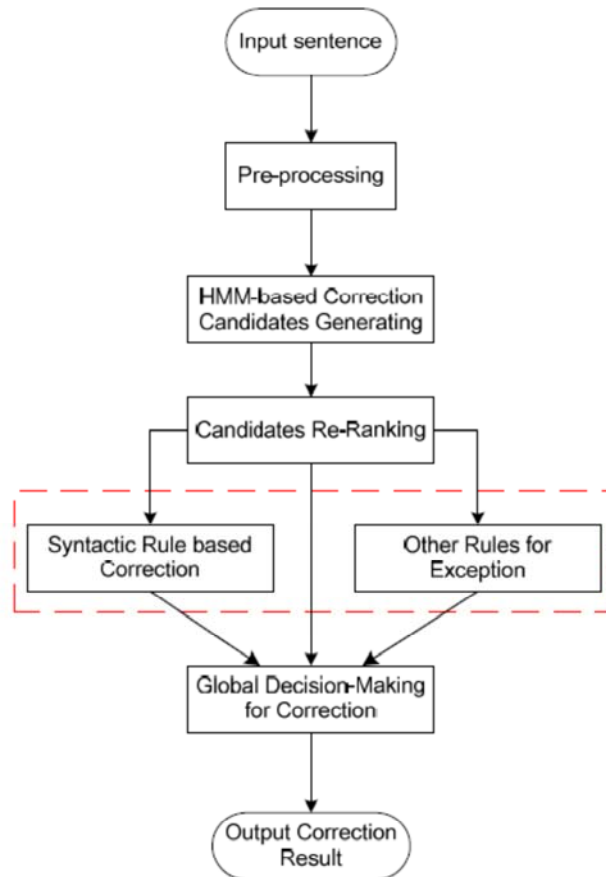


Figure 1. A unified framework (HANSpeller) for Chinese spelling correction.

3.1 Generating Candidates

Generating candidates of spelling correction is the basic part for the whole task, as it determines the upper bound of precision and recall rate of the approach. The HMM method can be used to generate candidates directly, but it faces several challenges when applied to Chinese essay spelling correction. (1) For high-quality spelling correction, the training of HMM is not a trivial task. (2) The long-span dependency in sentences makes a first-order hidden Markov model insufficient to catch contextual information. (3) Too many candidates make the algorithm not efficient enough, and some right corrections may be concealed by the wrong corrections.

To address the above challenges, some extensions have been made to the HMM-based spelling correction approach. First, the HMM-based method is used only for the candidate generation phase, not for final output correction generation. All kinds of possible error transformations will be integrated into the framework of the HMM approach, so as to get a high recall rate. Second, a higher-order hidden Markov model is used to capture long-span context dependency. Third, in order to reduce the number of candidates generated in the process, each word in the sentence only can be replaced with its homophone, near-homophone, or similar-shape word. In addition, a pruning dynamic programming algorithm is adopted to dynamically select the best correction candidates for each round of sentence segmentation and correction.

Figure 2 illustrates the whole process of the candidate generation phase.

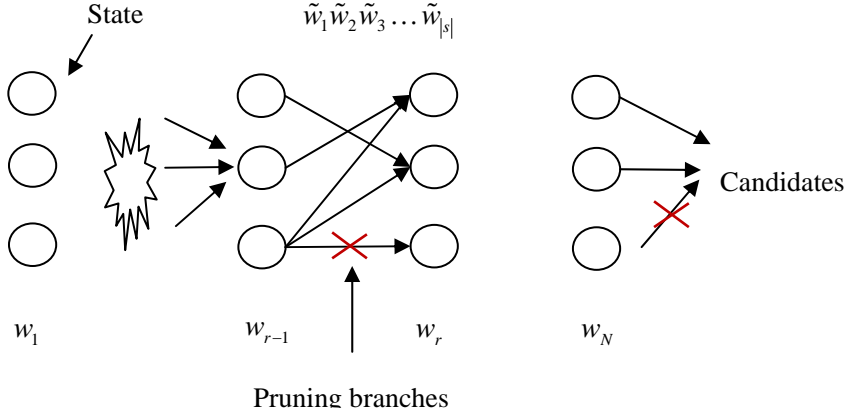


Figure 2. The whole process of generating candidates phase.

During the selection process of state, the edit distance and corrected results are combined to determine the quality of states. Let $S = w_1 w_2 w_3 \dots w_N$ be a sentence needing correction, where each item w_i is a word. C is a state generated from state transition and segmentation of the S 's r -th character, and $\tilde{w}_1 \tilde{w}_2 \tilde{w}_3 \dots \tilde{w}_{|s|}$ is the current corrected results in C . According to the noisy channel model, the occurrence probability of state C can be expressed as follows:

$$\begin{aligned}
 P(C) &= P\left(\tilde{w}_1 \tilde{w}_2 \tilde{w}_3 \dots \tilde{w}_{|s|} | w_1 w_2 w_3 \dots w_r\right) \\
 &= \frac{P(w_1 w_2 w_3 \dots w_r | \tilde{w}_1 \tilde{w}_2 \tilde{w}_3 \dots \tilde{w}_{|s|}) \times P(\tilde{w}_1 \tilde{w}_2 \tilde{w}_3 \dots \tilde{w}_{|s|})}{P(w_1 w_2 w_3 \dots w_r)}
 \end{aligned} \tag{1}$$

As $P(w_1 w_2 w_3 \dots w_r)$ is the same for states in the same level, Equation (1) can be simplified as:

$$P(C) \propto P(w_1 w_2 w_3 \dots w_r | \tilde{w}_1 \tilde{w}_2 \tilde{w}_3 \dots \tilde{w}_{|s|}) \times P(\tilde{w}_1 \tilde{w}_2 \tilde{w}_3 \dots \tilde{w}_{|s|}) \tag{2}$$

$$\log P(C) \propto \log P(w_1 w_2 w_3 \dots w_r | \tilde{w}_1 \tilde{w}_2 \tilde{w}_3 \dots \tilde{w}_{|s|}) + \log P(\tilde{w}_1 \tilde{w}_2 \tilde{w}_3 \dots \tilde{w}_{|s|}) \quad (3)$$

Conceptually, the above formula can be calculated approximately using edit distance and n-gram language model. Symbolically, it can be represented by:

$$\log P(C) \propto \text{editdis} \tan ce(C) + (\log P(\tilde{w}_1) + \log(\tilde{w}_2 | \tilde{w}_1) + \dots \log P(\tilde{w}_{|s|} | \tilde{w}_{|s|-n+1} \dots \tilde{w}_{|s|-1})) \quad (4)$$

In each round of the state generation stage, the best m states are selected according to the above calculated score. The remaining states are screened out to reduce the states' explosive growth, which improves the performance significantly. Finally, each sentence generates k candidates that represent the most likely correction results.

3.2 Ranking Candidates

In the candidate generation phase, top-k best candidates for a sentence are generated, but the HMM-based framework does not have the flexibility to incorporate a wide variety of features useful for spelling correction, such as online search results. Therefore, it is necessary to re-rank the candidates using more rich features, which can improve the precision of spelling correction significantly.

Given the original sentence, our system first generates a list of candidate sentences based on previous results. Then, the candidates in the list are re-ranked at this stage, based on the confidence score generated by a ranker, herein by an SVM classifier. Finally, we choose the top-2 candidates with the highest score to make the final decision.

The features used in our system can be grouped into five categories. They are listed separately in the table below.

Table 2. Five kinds of different features.

Feature Types	Features
Language Model Features	1.Text probability of candidates 2.Text probability of original sentence
Dictionary Features	1.Number of phrases 2.Number of idioms 2.Proportion of phrases 3.Proportion of idioms 3.Phrases and idioms length
Edit Distance Features	1.The number of homophone edit operations 2.The number of near-homophone edit operations 2.Total number of similar-shape edit operations 3.Total edit cost

Segmentation Features	1.The number of single words 2.The number of segmentations of words using MM 3.The number of segmentations of words using CKIP
Web Based Features	1.The search hits proportion of corrected part in title 2.The search hits proportion of corrected part in snippet

Language model features calculate the n-gram text probability of candidate sentences and the original sentence.

The n-gram language probability for a sentence S can be illustrated as the following equation:

$$P(S) = P(w_1, w_2 \dots w_n) = P(w_1)P(w_2|w_1) \dots P(w_n|w_1, w_2 \dots w_{n-1}) \quad (5)$$

Here, $P(w_1)$ is probability of word w_1 appearing in the corpus and $P(w_n | w_1, w_2 \dots w_{n-1})$ is the condition probability, which means the emergence probability of the word w_n under conditions of words $w_1, w_2 \dots w_{n-1}$ appearing.

Dictionary features count the number and proportion of phrases and idioms in candidates after segmentation, according to our dictionaries. In addition, some other factors, e.g. phrase length, are also taken into account.

Here is an example of a traditional Chinese sentence: 根据/联合国/公布/的/数据. The sub-sentence has 4 phrases and 0 idioms, and the proportion of phrases and idioms are 0.8 and 0.0, respectively, based on dictionaries.

Edit distance features compute the edit number and its weight, from the original sentence to candidate sentences. Here, different edit operations are given different edit weights. For example, in our spelling correction system, we give homophone, near-homophone, and similar shape word different edit weights, which are determined by experience.

Segmentation features use the results of the Maximum Matching Algorithm and the CKIP Parser segmentation. In addition, we count the number of single words. As we know, inappropriate candidates containing spelling errors will tend to have more single words after segmentation.

Web based features use Bing or another search engine's search results when submitting the spelling correction part and the corresponding part of the original sentence to the search engine.

“经济持续增长”and its candidate sentence “经济持续增长” would be an example. When you search “经济持续” or “持续增长” and “经济持续” or “持续增长” using Bing, the search engine will return different hits.

In our framework, the re-ranking phase is a must, because the candidates generated by HMM are ordered only by n-gram language probability and edit distance and the optimal state

of the HMM is not necessarily the best candidate. So, we use more features to reorder the candidates to view the candidate sentences according to the actual quality of candidates as much as possible. This step can help to improve the performance of final spelling correction.

In order to verify the effectiveness of re-ranking, we give the performance, whether adopting re-ranking or not, through experiments in the fourth section of the paper.

3.3 Rule-based Correction for Errors

As illustrated in Figure 1, the third step conducts rule-based analysis for a specific correction task. Some common errors still are difficult to distinguish, such as the usage of three confusable words “的,” “地,” and “得”. In order to correct such errors, syntactic analysis must be developed. The following sentence contains an error of Chinese syntax:

今天/我/穿着/刚/买/地/新/衣服。

Here, the character “地” should be corrected to another character “的”. To deal with these kinds of errors, sentence parsing must be done to check and correct such errors before the syntactic rules are applied. We have summarized three rules of usage for “的,” “地,” and “得” according to Chinese grammar as follows.

The Chinese character “的” is the tag of attributes, which generally is used in front of subjects and objects. Words in front of “的” generally are used to modify or restrict things following “的”.

The Chinese character “地” is adverbial marker, usually used in front of predicates (verbs, adjectives). Words in front of “地” generally are used to describe actions following “地”.

The Chinese character “得” makes the complement and generally is used behind predicates. The part follows “得” generally is used to supplement the previous action.

In addition, some other specific rules are needed to improve the final performance, which can be concluded from the test data and corpus.

3.4 Decision-making on Corrections

Through the aforementioned processing steps, we choose the top-2 candidates for each sub-sentence. To make the final decision on spelling correction, some global constraints should be considered, which can be summarized into four categories.

First, the number of errors in sub-sentence candidates should be considered. If there are more than three errors in a sub-sentence, then we do not correct the sub-sentence. Second, we set different weights for different types of spelling errors by experience. For example,

syntactic errors need to be given more weight than others, as these errors are detected by some strong syntactic rules. Then, if the original sub-sentence is in its candidate set, the sub-sentence has a greater probability of being error-free. Finally, the ratio of corrected sentences to the total amount of checked sentences is also one of the factors to consider. This ratio relates to the average error rate of CFL essays.

Let $Candi_{sentence} = \{candi_sub_1, candi_sub_2, \dots, candi_sub_n\}$ be the candidate set of a sentence, and $candi_sub_i$ be the top-2 candidates of its sub-sentence, $Final_Candi = \{final_candi_sub_p, final_candi_sub_{p+1}, \dots, final_candi_sub_q\}$ be the final candidate list of the sub-sentence in the intermediate process, and $Final_Correction = \{final_sub_1, final_sub_2, \dots, final_sub_n\}$ be the final correction result.

According to the constraints above, our rules are summarized as follows.

- 1) Scan each element of $Candi_{sentence}$. If the number of errors of top-2 candidates in $candi_sub_i$ is all more than 3 or the original sub-sentence is in $candi_sub_i$ and ranked first after re-ranking, store the original sub-sentence in $Final_Correction$ and continue scanning $candi_sub_{i+1}$; otherwise, go to Step 2);
- 2) Compute the scores of the top-2 candidates in $candi_sub_i$, and store the candidate with higher score in $Final_Candi$. If the scan is not over, go to Step 1); otherwise, go to Step 3);
- 3) Provide statistics for the total number of errors in $Final_Candi$. If the error quantity is less than the threshold value, then output $Final_Candi$ to $Final_Correction$ and skip to Step 5); otherwise, go to Step 4);
- 4) Sort the $Final_Candi$ according to the score computed in Step 2). Scan $Final_Candi$, output the front part of $Final_Candi$ to $Final_Correction$ according to the global error rate, and the remaining part of $Final_Candi$ is not corrected, go to step 5);
- 5) Output the $Final_Correction$.

In Step 2) above, there is a function to calculate the score of candidate, and the score can be computed as follows:

$$score(candidate) = edit_weight + original_weight - edit_num \quad (6)$$

where $edit_weight$ is the edit weight of the candidate, $original_weight$ is the weight of whether the candidate is original sentence or not, and $edit_num$ is the number of edits in candidate. The weights currently are set by experience. The value of $edit_weight$ is set according to the error type. If the type is homophone or similar shape, $edit_weight$ is set to 0.8, otherwise it is set to 0.5. The value of $original_weight$ is also set by experience. If the candidate is original sentence, it is set to 1, otherwise it is set to 0.75.

On the basis of the above rules, we developed a rule-based classifier to get the final correction result of each sentence.

4. Experiment and Evaluation

4.1 Experimental Setting

In the experiment, 1062 traditional Chinese sentences with/without spelling errors were given, which were from CFL learners' essays. The error types in the sentences mainly resulted from three different categories, being homophone, near-homophone, or similar-shape. The test data was provided by SIGHAN 2014 Bake-off: Chinese Spelling Check Task.

As the test data set was based on traditional Chinese, we must consider building a traditional Chinese corpus to train our model. In our system, we use several corpora, including Taiwan Web as corpus; SogouW dictionary, which is a traditional Chinese dictionary translated from the simplified Chinese dictionary Sogou, a traditional Chinese dictionary of words and idioms; a pinyin table and a cangjie code table of common words; and some Web based resources. The details of the corpora are described below.

(1) Taiwan Web Pages as Corpus

Due to the difference in simplified Chinese and traditional Chinese, although we have a high quality simplified Chinese corpus, we do not translate the simplified corpus into a traditional corpus because the translation process may cause information loss, such as the fact that both“週末” and “周末” in traditional Chinese are translated into “周末” in simplified Chinese. Therefore, we try to find Taiwan webs whose pages contain high quality traditional Chinese text to build the corpus. We gathered pages from the artificially selected pages under the “.tw” domain, containing around 3.2 million web pages, to build the corpus. Then, the content extracted from these pages was used to build a traditional Chinese n-gram model, where n is from 2 to 4.

(2) SogouW Dictionary

SogouW dictionary is built from the statistical analysis of Chinese Internet corpus by Sogou Search Engine. It contains about 150,000 high-frequency words of the Chinese Internet. Nevertheless, words in the corpus are simplified Chinese characters that cannot be used directly. We first translated them into traditional Chinese via Google translation service.

(3) Chinese Words and Idioms Dictionary

As introduced in Chiu *et al.* (2013), we also obtained the Chinese words and Chinese idioms published by the Ministry of Education of Taiwan, which are built from dictionaries and related books. There are 64,326 distinct Chinese words and 48,030 distinct Chinese idioms. We combined these two dictionaries with the SogouW dictionary to build our trie tree dictionary.

(4) Pinyin and Cangjie Code Table

We collected more than 10000 pinyin forms of words commonly used in Taiwan to build

the homophone and near-homophone words table, which will be used in candidate generation phase. In addition, cangjie code can be used to measure the form/shape similarity between Chinese characters. Therefore, we collected cangjie codes to build the table of Similar-form characters.

(5) Web based Resources

We use the online CKIP Parser results to help rank the candidates. For example, the segmentation of “持續下滑” is “特/續/下滑” while “持續下滑” is “持續/下滑”. Thus, the segmentation results of a wrong candidate sentence will have more words than the correct one.

In addition, we use the Bing search results as one feature in the candidate ranking phase, which clearly improves the performance. For example, the sentence “根據聯合國公布的數字” has several candidate sentences, one of which may be “根據聯合國公佈的數字”. If we use Bing to search the error correction part and the corresponding part of the original sentence “聯合國公佈” and “聯合國公布,” the search results will be clear enough to identify the correct candidate sentence, because the first one would be more popular than the second one on the web corpus.

4.2 Evaluation Results and Analysis

To evaluate the method we propose, a Chinese spelling check system was implemented. We have done some experiments to prove the effectiveness of our method for Chinese spelling correction. The task can be divided into two related subtasks. One is error detection and the other one is error correction. Chinese spelling error detection task aims to find out the location of the spelling errors in the sentences. The error correction task aims to correct the error words found in the error detection phase. There are five metrics, used to evaluate the performance of different methods. They are calculated as the following expression:

$$FPR(\text{FalsePositiveRate}) = \frac{FP}{FP + TN} \quad A(\text{Accuracy}) = \frac{TP + TN}{TP + TN + FP + FN}$$

$$P(\text{Precision}) = \frac{TP}{TP + FP} \quad R(\text{Recall}) = \frac{TP}{TP + FN}$$

$$F1 - \text{Score} = \frac{2 * P * R}{P + R}$$

where TP , FP , TN , and FN can be obtained from the confusion matrix in Table 3.

Table 3. Confusion Matrix.

Confusion Matrix		System Results	
		Positive (Error)	Negative (No Error)
Gold Standard	Positive	TP	FN
	Negative	FP	TN

11 competing teams joined the SIGHAN Bake-off 2014 and submitted their final results. These submitted methods are used to evaluate the performance of our proposed framework. NCTU & NTUT used a CRF-based parser and scored with a tri-gram LM; NCYU combined E-HowNet and n-gram models to construct the rule induction; NJUPT developed two CSC systems based on CRF model and 2 Chars & 3-Chars model, respectively; NTHU used a channel model and a character-based language model in the noisy model; SinicaCKIP combined the error template rules and n-gram models for Chinese spelling correction; the SJTU proposed an improved graph model based on a graph model for generic errors and two independently trained models for specific errors. The results of the two subtasks are described in detail in Sections 4.2.1 and 4.1.2.

In addition, we will analyze the effects of several features used in the ranking stage on the final results. The comparative results are introduced in Section 4.2.3.

4.2.1 Chinese Spelling Error Detection

The goal of this subtask is to detect whether a Chinese sentence contains errors or not. If the sentence contains errors, the subtask must point out the location of the error word. Table 4 shows the evaluation results of Chinese spelling error detection.

Table 4. Results of error detection subtask for different methods

Methods	A	P	R	F1
Decision-Making Model [CAS]	0.6149	0.7148	0.3823	0.4982
CRF-Model + N-gram model[NCTU& NTUT]	0.5028	0.5138	0.1055	0.175
Rule Induction [NCYU]	0.6008	0.8543	0.2429	0.3783
CRF-Model + N-gram Model [NJUPT]	0.403	0.3344	0.1959	0.247
Noisy Channel Model [NTHU]	0.4228	0.3677	0.2147	0.2711
Error Template Rule + N-gram Model [SinicaCKIP]	0.5367	0.5607	0.339	0.4225
Graph-Model + CRF-Model [SJTU]	0.5471	0.5856	0.322	0.4156

The above results illustrate that our system significantly outperforms other systems with submitted technique reports to the organizer in this subtask. This is due to our method using the extended HMM to guarantee the recall rate and introducing the re-rank phase combined with rich features to improve the precision.

4.2.2 Chinese Spelling Error Correction

The subtask is based on the task of error detection. The main idea is to correct the errors found in the detection phase. In this stage, each sentence will be corrected and compared to the reference answer. Our system showed good performance in this subtask. The error correction results are shown in Table 5.

Table 5. Results of error correction subtask for different methods

Methods	FPR	A	P	R	F1
Decision-Making Model [CAS]	0.1525	0.5829	0.676	0.3183	0.4328
CRF-Model + N-gram model[NCTU& NTUT]	0.0998	0.4925	0.4592	0.0847	0.1431
Rule Induction [NCYU]	0.0414	0.5885	0.8406	0.2185	0.3468
CRF-Model + N-gram Model [NJUPT]	0.3898	0.3964	0.3191	0.1827	0.2323
Noisy Channel Model [NTHU]	0.3691	0.3823	0.2659	0.1337	0.1779
Error Template Rule + N-gram Model [SinicaCKIP]	0.2655	0.5104	0.5188	0.2863	0.3689
Graph-Model + CRF-Model [SJTU]	0.2279	0.5377	0.5709	0.3032	0.3961

The results show that our system also provides good performance in the correction subtask. This is because it achieves good results in the detection subtask, which is the basis of the correction subtask.

4.2.3 The Influence of Different Ranking Features

In this part, we compare the effects of several features used in the ranking step on the final results. As the dictionary features and segmentation features are closely related, we ignore the comparison of segmentation features. In the experiment, we conducted the test over multiple rounds, where we excluded one kind of feature in each round. The test results are shown in Table 6.

Table 6. The effect of difference ranking features

Features (Excluded)	FPR	Detection-Level			
		A	P	R	F1
Language Model Features	0.2312	0.548	0.564	0.3153	0.4045
Dictionary Features	0.1523	0.5857	0.7068	0.3418	0.4608
Edit Distance Features	0.1726	0.5574	0.7003	0.3339	0.4522
Web Based Features	0.3663	0.5094	0.4401	0.3558	0.3935
None	0.1525	0.6149	0.7148	0.3823	0.4982
Features (Excluded)	FPR	Correction-Level			
		A	P	R	F1
Language Model Features	0.2312	0.5113	0.496	0.2398	0.3233
Dictionary Features	0.1523	0.5584	0.6709	0.2891	0.4041
Edit Distance Features	0.1726	0.5273	0.6612	0.2788	0.3923
Web Based Features	0.3663	0.4586	0.3485	0.2421	0.2857
None	0.1525	0.5829	0.676	0.3183	0.4328

Based on the results above, the language model features and web-based features are the two most important features in the ranking phase on the final results, as the two features mainly reflect the quality of web based corpus.

4.2.4 The Influence of Re-ranking

In this part, we verify the important role of re-ranking in the spelling correction. We correct the sentences in two ways, one only based on HMM and the other adopting re-ranking after generating candidates. Table 7 shows the final results.

Table 7. The correction results of whether adopting re-ranking or not

Error-Detection	A		P	R	F1
With Re-ranking	0.6149		0.7148	0.3823	0.4982
Without Re-ranking	0.4859		0.5156	0.2383	0.3259
Error-Correction	FPR	A	P	R	F1
With Re-ranking	0.1525	0.5829	0.676	0.3183	0.4328
Without Re-ranking	0.2441	0.4407	0.4038	0.1516	0.2205

As illustrated by the above results, re-ranking significantly improves the performance of results both in the error-detection and the error-correction tasks. In the error-detection task, the method with re-ranking outperforms the method without re-ranking with 19.92% improvement in precision and 14.4% improvement in recall rate. In the error-correction task, the precision and recall rate increase by 27.22% and 16.67%, respectively.

5. Conclusion

This paper proposes a unified framework (HANSpeller) for Chinese essay spelling correction based on extended HMM and ranker-based models. An extended HMM is proposed to generate candidate sentences for ranking. A rule-based strategy is used for further correction polishing and for a final decision on whether the output is the correction or not. Our approach was evaluated at the CLP-2014 bake-off on the Chinese spelling correction task, and it displayed good performance, ranking second among 13 teams.

Some interesting future work on Chinese spelling correction would include: (1) collecting and considering more error types in the candidates generating process and (2) how to better deal with the differences between traditional and simplified Chinese.

Acknowledgments

This research was supported by the National High Technology Research and Development Program of China (Grant No. 2014AA015204), the National Basic Research Program of China (Grant No. 2014CB340406), the NSFC for the Youth (Grant No. 61402442) and the Technology Innovation and Transformation Program of Shandong (Grant No.2014CGZH1103).

Reference

- Ahmad, F., & Kondrak, G. (2005). Learning a spelling error model from search query logs. In *Proceeding of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 955-962.
- Atwell, E. S., & Elliot, S. (1987). Dealing with ill-formed English text. *The Computational Analysis of English: A Corpus-Based Approach*, 120-138.
- Bril, E., & Moore, R. C. (2000). An improved error model for noisy channel spelling correction. In *Proceeding of the 38th Annual Meeting on Association for Computational Linguistics*, 286-293.
- Chang, C. H. (1995). A new approach for automatic Chinese spelling correction. In *Proceeding of Natural Language Processing Pacific Rim Symposium*, 278-283.

- Chang, R. Y., Wu, C. H., & Prasetyo, P. K. (2012). Error Diagnosis of Chinese Sentences Using Inductive Learning Algorithm and Decomposition-Based Testing Mechanism. *ACM Transactions on Asian Language Information Processing (TALIP)*, 11(1), 3.
- Chen, Y. Z. (2010). *Improve the detection of improperly used Chinese characters with noisy channel model and detection template* (Doctoral dissertation, Master thesis, Chaoyang University of Technology).
- Chiu, H. W., Wu, J. C., & Chang, J. S. (2013). Chinese Spelling Checker Based on Statistical Machine Translation. In *Proceeding of the Seventh SIGHAN Workshop on Chinese Language Processing (SIGHAN-7)*, 49-53.
- Chu, W. C., & Lin, C. J. (2014). NTOU Chinese Spelling Check System in CLP Bake-off 2014. In *Proceeding of the Third CIPS-SIGHAN Joint Conference on Chinese Language Processing*, 210-215.
- Cucerzan, S., & Brill, E. (2004). Spelling Correction as an Iterative Process that Exploits the Collective Knowledge of Web Users. In *Proceeding of EMNLP*, 293-300.
- Gao, J., Li, X., Micol, D., Quirk, C., & Sun, X. (2010). A large scale ranker-based system for search query spelling correction. In *Proceeding of the 23rd International Conference on Computational Linguistics*, 358-366.
- Google. (2010). *A Java API for Google spelling check service*. <http://code.google.com/p/google-api-spellingjava/>
- Gu, L., Wang, Y., & Liang, X. (2014). Introduction to NJUPT Chinese Spelling Check Systems in CLP-2014 Bakeoff. In *Proceeding of the Third CIPS-SIGHAN Joint Conference on Chinese Language Processing*, 167-172.
- Hao, S., Gao, Z., Zhang, M., Xu, Y., Peng, H., Su, K., & Ke, D. (2013). Automated error detection and correction of chinese characters in written essays based on weighted finite-state transducer. In *Proceeding of the 12th International Conference on Document Analysis and Recognition (ICDAR)*, 763-767.
- Hsieh, Y. M., Bai, M. H., & Chen, K. J. (2013). Introduction to CKIP Spelling Check System for SIGHAN Bakeoff 2013 Evaluation. In *Proceeding of the Seventh SIGHAN Workshop on Chinese Language Processing (SIGHAN-7)*, 59-63.
- Huang, Q., Huang, P., Zhang, X., Xie, W. J., Hong, K., Chen, B. Z., & Huang, L. (2014). Chinese Spelling Check System Based on Tri-gram model. In *Proceeding of the Third CIPS-SIGHAN Joint Conference on Chinese Language Processing*, 173-178.
- Huang, C. M., Wu, M. C., & Chang, C. C. (2007). Error detection and correction based on Chinese phonemic alphabet in Chinese text. In *Modeling Decisions for Artificial Intelligence*, 463-476.
- Hung, T. H. (2009). *Automatic Chinese character error detecting system based on n-gram language model and pragmatics knowledge base* (Doctoral dissertation, Master thesis, Chaoyang University of Technology).
- Hung, T. H., & Wu, S. H. (2008). Chinese essay error detection and suggestion system. In *Taiwan E-Learning Forum 2008*.

- Jia, Z. Y., Wang, P. L., & Zhao, H. (2013). Graph Model for Chinese Spelling Checking. In *Proceeding of the Seventh SIGHAN Workshop on Chinese Language Processing (SIGHAN-7)*, 88-92
- Jin, P., Chen, X., Guo, Z., & Liu, P. (2014). Integrating Pinyin to Improve Spelling Errors Detection for Chinese Language. In *Proceeding of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies*, 455-458.
- Kukich, K. (1992). Techniques for automatically correcting words in text. *ACM Computing Surveys (CSUR)*, 24(4), 377-439.
- Levenshtein, V. I. (1966). Binary code capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, 10(8), 707-710.
- Li, M., Zhang, Y., Zhu, M., & Zhou, M. (2006). Exploring distributional similarity based models for query spelling correction. In *Proceeding of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, 1025-1032.
- Liu, C. L., Lai, M. H., Tien, K. W., Chuang, Y. H., Wu, S. H., & Lee, C. Y. (2011). Visually and phonologically similar characters in incorrect Chinese words: Analyses, identification, and applications. *ACM Transactions on Asian Language Information Processing (TALIP)*, 10(2), 1-39.
- Mangu, L., & Brill, E. (1997). Automatic rule acquisition for spelling correction. In *Proceeding of the 14th International Conference on Machine Learning*, 187-194.
- Mays, E., Damerau, F. J., & Mercer, R. L. (1991). Context based spelling correction. *Information Processing & Management*, 27(5), 517-522.
- Microsoft Microsoft web n-gram services. (2010). <http://research.microsoft.com/web-ngram>
- Suzuki, H., & Gao, J. (2012). A unified approach to transliteration-based text input with online spelling correction. In *Proceeding of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 609-618.
- Wu, S. H., Liu, C. L., & Lee, L. H. (2013). Chinese Spelling Check Evaluation at SIGHAN Bake-off 2013. In *Proceeding of the Sixth International Joint Conference on Natural Language Processing*, 35-42.
- Xiong, J., Zhang, Q., Hou, J., Wang, Q., Wang, Y., & Cheng, X. (2014). Extended HMM and Ranking models for Chinese Spelling Correction. *CLP 2014*, 133-138.
- Yu, L. C., Lee, L. H., Tseng, Y. H., & Chen, H. H. (2014). Overview of SIGHAN 2014 Bake-off for Chinese Spelling Check. In *Proceeding of the Third CIPS-SIGHAN Joint Conference on Chinese Language Processing*, 126-132.
- Zhang, L., Huang, C., Zhou, M., & Pan, H. (2000). Automatic detecting/correcting errors in Chinese text by an approximate word-matching algorithm. In *Proceeding of the 38th Annual Meeting on Association for Computational Linguistics*, 248-254.

Zheng, Y., Li, C., & Sun, M. (2011). Chime: An efficient error-tolerant Chinese pinyin input method. In *Proceeding of International Joint Conference on Artificial Intelligence(IJCAI)*, 22(3), 2551-2256.

A Study on Chinese Spelling Check Using Confusion Sets and N -gram Statistics

Chuan-Jie Lin* and Wei-Cheng Chu*

Abstract

This paper proposes an automatic method to build a Chinese spelling check system. Confusion sets were expanded by using two language resources, Shuowen Jiezi and the Four-Corner codes, which improved the coverages of the confusion sets. Nine scoring functions which utilize the frequency data in the Google Ngram Datasets were proposed, where the idea of smoothing was also adopted. Thresholds were also decided in an automatic way. The final system achieved far better than our baseline system in CSC 2013 Evaluation Task.

Keywords: Chinese Spelling Check, Confusion Set Expansion, Google Ngram Scoring Function.

1. Introduction

Automatic spelling check is a basic and important technique in building NLP systems. It has been studied since 1960s as Blair (1960) and Damerau (1964) made the first attempt to solve the spelling error problem in English. Spelling errors in English can be grouped into two classes: non-word spelling errors and real-word spelling errors.

A non-word spelling error occurs when the written string cannot be found in a dictionary, such as in “*fly fron* Paris*”. The typical approach is finding a list of candidates from a large dictionary by edit distance or phonetic similarity (Mitton, 1996; Deorowicz & Ciura, 2005; Carlson & Fette, 2007; Chen *et al.*, 2007; Mitton, 2008; Whitelaw *et al.*, 2009).

A real-word spelling error occurs when one word is mistakenly used for another word, such as in “*fly form* Paris*”. Typical approaches include using confusion set (Golding & Roth, 1999; Carlson *et al.*, 2001), contextual information (Verberne, 2002; Islam & Inkpen, 2009), and others (Pirinen & Linden, 2010; Amorim & Zampieri, 2013).

Spelling error problem in Chinese is quite different. Because there is no word delimiter

* Department of Computer Science and Engineering, National Taiwan Ocean University
No. 2, Pei-Ning Road, Keelung, 20224 Taiwan
E-mail: (cjlin, wcchu.cse)@ntou.edu.tw

in a Chinese sentence and almost every Chinese character can be considered as a one-character word, most of the errors are real-word errors.

Although that an illegal-character error can happen where writing by hand, i.e. the written symbol is not a legal Chinese character and thus not collected in a dictionary, such an error cannot happen in a digital document because only legal Chinese characters can be typed or shown in computer.

Spelling error problem in Chinese is defined as follows: given a sentence, find the locations of misused characters which result in wrong words, and propose the correct characters.

There have been many attempts to solve the spelling error problem in Chinese (Chang, 1994; Zhang *et al.*, 2000; Cucerzan & Brill, 2004; Li *et al.*, 2006; Liu *et al.*, 2008). Among them, lists of visually and phonologically similar characters play an important role in Chinese spelling check (Liu *et al.*, 2011).

Two Chinese spelling check evaluation projects have been held: Chinese Spelling Check Evaluation at SIGHAN Bake-off 2013 (Wu *et al.*, 2013) and CLP-2014 Chinese Spelling Check Evaluation (Yu *et al.*, 2014), including error detection and error correction subtasks. The tasks are organized based on some research works (Wu *et al.*, 2010; Chen *et al.*, 2011; Liu *et al.*, 2011). Our baseline system participated in both tasks. This paper describes an extended system based on Chinese Spelling Check (shorten as CSC tasks hereafter) 2013 and 2014 datasets.

This paper is organized as follows. Section 2 introduces our baseline system developed during Chinese Spelling Check Task 2013 and 2014. We sought new resources to expand confusion sets as described in Section 3. New scoring functions and threshold decision using Google Ngram frequencies to estimate the likelihood of passages were defined in Section 4. Section 5 shows experimental results with discussions and Section 6 concludes this paper.

2. Baseline System Description

2.1 System Architecture

Figure 1 shows the architecture of our Chinese spelling checking system. A sentence under consideration is first word-segmented. Candidates of spelling errors are replaced by similar characters one by one. The newly created sentences are word segmented again. They are sorted according to sentence generation probabilities measured by word or POS bigram model. If a replacement results in a better sentence, spelling error is reported.

In CSC tasks, the set of similar characters is called a confusion set. More information about confusion sets is given in Section 2.2.

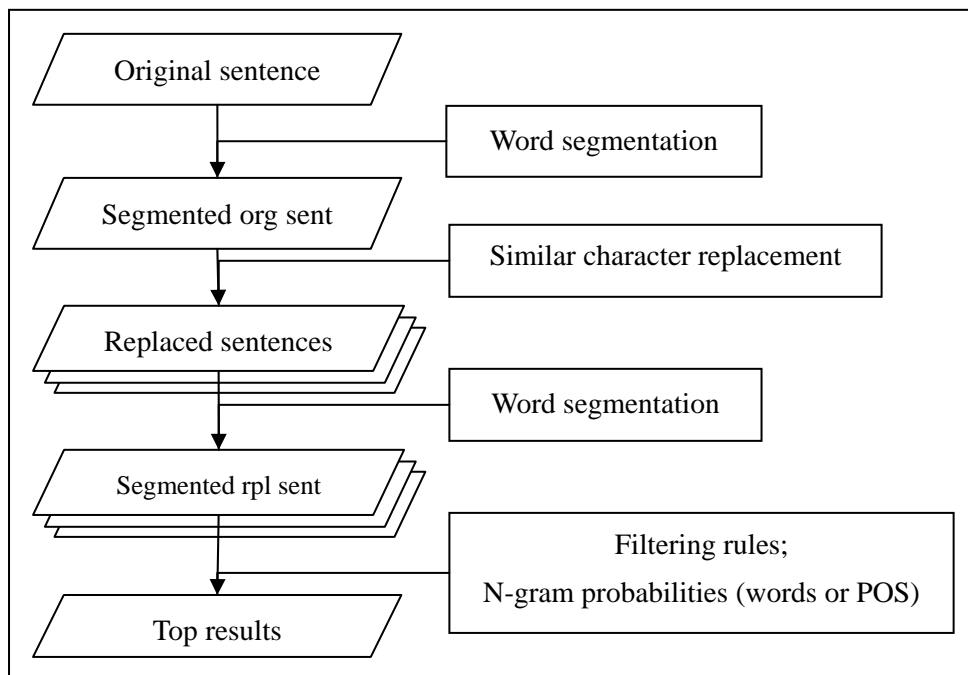


Figure 1. Architecture of NTOU Chinese Spelling Check System

There are two kinds of spelling-error candidates in our system: one-character words and two-character words. Their replacement procedures are different, as described in Section 2.3 and 2.4.

Section 2.5 introduced two rules for filtering out unlikely replacements. *N*-gram probability models in our baseline system are described in Section 2.6. The procedure to decide locations of errors is given in Section 2.7.

2.2 Confusion Sets

In SIGHAN7 Bake-off 2013 Chinese Spelling Check task, the organizers provided six kinds of confusion sets: 4 sets of phonologically similar characters and 2 sets of visually similar characters. The four sets of phonologically similar characters include characters with the same pronunciation in the same tone (同音同調, shorten as SPST hereafter), characters with the same pronunciation but in different tones (同音異調, shorten as SPDT hereafter), characters with similar pronunciations in the same tone (近音同調, shorten as DPST hereafter), and characters with similar pronunciations but in different tones (近音異調, shorten as DPDT hereafter). For example, phonologically similar characters to the character 情 (whose pronunciation is [qing2] and meaning is ‘feeling’) are:

SPST: 擘晴擎[qing2]
 SPDT: 青卿蜻傾輕鯖氫清[qing1] 頃請[qing3] 慶罄磬[qing4]
 DPST: 擒禽噲琴勤秦芹[qin2]
 DPDT: 精經驚睛…京[jing1] 頸景警…井[jing3] 竟靜競徑鏡…敬[jing4]
 今筋斤津…金[jin1] 僅儘錦緊…謹 [jin3] 近進勁盡禁…浸[jin4]
 親侵欽嶽[qin1] 寢[qin3] 沁撤[qin4]

There are two confusion sets of visually-similar characters. The first one is the set of characters with the same radicals (部首) with the same number of strokes (筆劃) (同部首同筆畫數, shorten as RStrk hereafter). For example, the radical of the character 情 is 心 (shown as 忄 inside the character) with 11 strokes. Characters belonging to the radical 心 with 11 strokes are:

RStrk: 惋您悉惇惆悠患怙惚悼悽惘悸惟惜悻悴悵愾惕

The second visually-similar-character set collects characters with similar Cangjie codes (倉頡碼, shorten as CJie hereafter). Cangjie is a well-known code map of Chinese characters. Each Chinese character is encoded by a combination of at most 5 codes representing basic strokes in its visual structure. Characters who have similar Cangjie codes are likely visually similar. Liu *et al.* (2011) considered the information of surface structure and stroke similarity to create this confusion set. For example, the Cangjie code of the character 情 ([qing2], ‘feeling’) is PQMB, where “P 忄” denotes its radical part (忄) and “QMB 青一月” denotes its body part (青). So its similar characters are:

CJie:
 清[EQMB] 晴[AQMB] 倩[OQMB] 猜[KHQMB] 睛[BUQMB]
 靖[YTQMB] 精[FDQMB] 蜻[LIQMB] 鯖[NFQMB] 菁[TQMB]
 請[YRQMB] 青[QMB] 債[OQMC] 漬[EQMC] 嘖[RQMC]
 磧[MRQMC] 積[HDQMC] 績[VFQMC] 蹟[QMQMC] 責[QMBUC]

2.3 One-Character Word Replacement

After doing word segmentation on the original sentence, every one-character word is considered as candidate where error occurs. These candidates are one-by-one replaced by similar characters in their confusion sets to see if a new sentence is more acceptable.

Taking C1-1701-2 in the test set as an example. The original sentence is

...嬰兒個數卻特續下滑...

and it is segmented as

...嬰兒 個數 卻 特 續 下滑...

“卻”，“特” and “續” are one-character words so they are candidates of spelling errors. The confusion set of the character “卻” includes 腳欲叩卸... and the confusion set of the character “特” includes 持時恃峙侍... Replacing these one-character words with similar characters one-by-one will produce the following new sentences.

...嬰兒個數脚特續下滑...
...嬰兒個數欲特續下滑...
...嬰兒個數卻持續下滑... (*correct*)
...嬰兒個數卻時續下滑...
.....

(English meaning: 嬰兒 infant, 個數 number, 卻 but, 腳 foot, 欲 desire,
特 particular, 續 continue, 持續 keep, 時 time, 下滑 decrease)

(Original sentence: infant number but special continue decrease

'but the number of infants particularly continues to decrease')

(Correct sentence: 嬰兒個數卻持續下滑 *'but the number of infants keeps decreasing'*)

2.4 Two-Character Word Replacement

Our observation on the training sets finds that some errors occur in two-character words, which means that a string containing an incorrect character is also a legal word. Examples are “身手” ([shen1-shou3], ‘skills’) versus “生手” ([sheng1- shou3], ‘amateur’), and “人員” ([ren2-yuan2], ‘member’) vs. “人緣” ([ren2-yuan2], ‘relation’).

To handle such kinds of spelling errors, we created confusion sets for all known words by the following method. The resource for creating word-level confusion set is Academia Sinica Balanced Corpus (ASBC for short hereafter, cf. Chen *et al.*, 1996).

For each word appearing in ASBC, each character in the word is substituted with its similar characters one by one. If a newly created word also appears in ASBC, it is collected into the confusion set of this word. Take the word “人員” as an example. After replacing “人” or “員” with their similar characters, new strings 仁員, 壬員, ..., 人緣, and 人韻 are looked up in ASBC. Among them, only 人緣, 人猿, 人文, and 人備 are legal words thus collected in 人員’s confusion set.

For each two-character word, if it has a confusion set, similar words in the set one-by-one substitute the original word to see if a new sentence is more acceptable.

Take ID=00058 in the Bakeoff 2013 CSC Datasets as an example. The original sentence is

... 在教室裡只要人員好...

and it is segmented as

... 在 教室 裡 只要 人員 好...

where “教室”, “只要”, and “人員” are multi-character words with confusion sets. By replacing 教室 with 教士, 教師..., replacing 只要 with 祇要, 只有, and replacing 人員 with 人緣, 人猿..., the following new sentences will be generated.

... 在教士裡只要人員好...
 ... 在教師裡只要人員好...
 ... 在教室裡祇要人員好...
 ... 在教室裡只要人緣好... (correct)
 ... 在教室裡只要人猿好...

(English meaning: 在 in, 教室 classroom, 教士 priest, 教師 teacher, 裡 inside, 只要 as-long-as, 祇要 as-long-as (variant), 人員 member, 人緣 relations, 人猿 ape, 好 good)

(Original Sentence: in classroom inside as-long-as member good
 ‘as long as there are good members in the classroom...’)

(Correct sentence: 在教室裡只要人緣好 ‘in the classroom, as long as you have good relations with the others...’)

2.5 Filtering Rules

Two filter rules are applied before error detection in order to discard apparently incorrect replacements. The rules are defined as follows.

Rule 1: No error in person names

If a replacement results in a person name, discard it. Our word segmentation system performs named entity recognition at the same time. If the replacing similar character can be considered as a Chinese family name, the consequent characters might be merged into a person name. As most of the spelling errors do not occur in personal names, we simply ignore these replacements. Take C1-1701-2 as an example:

...每 位 產 齡 婦 女...

(every QF pregnancy age woman ‘*every woman in the age of pregnancy*’)

“魏” is phonologically similar to “位” and is a Chinese family name. The newly created sentence is segmented as

...每 魏產齡(PERSON) 婦女...

(every Chan-Ling Wei woman: *nonsense*)

where “魏產齡” is recognized as a person name so this replacement is discarded.

Rule 2: Stopword filtering

For the one-character replacement, if the replaced (original) character is a personal anaphora (你 ‘you’ 我 ‘I’ 他 ‘he/she’) or numbers from 1 to 10 (一 二 三 四 五 六 七 八 九 十), discard the replacement. We assume that a writer seldom misspell such words. Take B1-0122-2 as an example:

...我 會 在 二 號 出 口 等 你...

(I will at two number exit wait you ‘*I will wait for you at Exit No. 2*’)

Although “二” is a one-character word, it is in our stoplist therefore no replacement is performed on this word.

2.6 *N*-Gram Probabilities

A basic hypothesis is that a correct replacement will generate a “better” sentence which has higher probability than the original one.

The likelihood of a passage being understandable can be estimated as sentence generation probability by language models. We tried smoothed word-unigram, word-bigram, and POS-bigram models in our baseline system. The training corpus used to build language models is ASBC. As usual, we use log probabilities instead.

Besides applying rules in which the probabilities were compared directly, we also treated them as features to train a SVM classifier which guessed whether a replacement was correct or not.

2.7 Error Detection

In our system, error detection and correction greatly rely on sentence generation probabilities. Therefore, all the newly created sentences should also be word segmented. If a new sentence results in a better word segmentation, it is very likely that the original character is misused and this replacement is correct. But if no replacement is better than the original sentence, it is reported as “no error”.

The detail of our error detection algorithm is delivered here. The original sentence is first divided into several sub-sentences by six sentence-delimiting punctuation marks: comma, period, exclamation, question mark, colon, and semicolon. The following steps are performed on each sub-sentence, referred to as *original passage* hereafter.

1. Divide the original sentence into several passages by the sentence-delimiting punctuation marks
2. Perform word segmentation on the original passages
3. Measure the likelihood of the original passages by language models
4. For each one-character word in each original passage
 - (1) Skip the word if it is a person name or a stopword (filtering rules)
 - (2) Replace the word with its similar characters in the confusion sets to generate un-segmented passages, one new passage for one similar character
 - (3) Perform word segmentation on the new passages
5. For each two-character word in each original passage
 - (1) If the word appears in the two-character confusion set, replace the word with its similar words in the two-character confusion sets to generate un-segmented passages, one new passage for one similar word
 - (2) Perform word segmentation on the new passages

6. Measure the likelihood of the new passages from step 4 and 5 by language models
7. If no new passage has a higher score than its original passage, report “no error” in this original passage
8. Consider only the new passage with the highest score
 - (1) If its score comparing to the original one is not higher than a pre-defined threshold, report “no error” in this original passage
 - (2) Otherwise, report the location and the similar character (or locations of similar characters in a two-syllable similar word) of the replacement which generates this new passage

3. Confusion Set Expansion

In our experience, the confusion sets provided by the task organizers do not cover all the errors. The error coverage of the confusion sets is depicted in Table 1, where TR means training set and TS means test set. The first 9 rows show the coverage of each confusion set, where set 0 to set 5 have been explained in Section 2.2. We can see that the SPST confusion set alone covers 70% of the errors in CSC 2013 datasets but only about half of the errors in CSC 2014 datasets. The second important confusion set is CJie, which covers 30% to 40% of the errors.

The last 10 rows of Table 1 show the coverage of the unions of confusion sets. The union of set 0~5 covers 94.59% of the errors. The union of set 0~3+5 has the same coverage as the union of set 0~5, which suggests that RStrk can be ignored.

In order to achieve better coverage, we used two resources to expand the confusion sets. One is Shuowen Jiezi and the other is the Four-Corner Encoding System.

Table 1. Error Coverage of Confusion Sets (%)

Confusion Set	TR2013	TS2013	TR2014	TS2014
set0: SPST	70.09	72.13	47.92	47.41
set1: SPDT	15.10	17.50	46.52	47.03
set2: DPST	3.70	4.99	5.15	4.68
set3: DPDT	3.70	4.67	8.41	7.71
set4: RStrk	9.12	3.17	0.38	0.88
set5: CJie	40.46	36.18	29.72	31.10
set6: Cor4	14.81	6.89	1.84	1.52
set7: SWen1	17.09	19.24	11.48	12.64
set8: SWen2	18.23	19.64	11.91	12.90

set0+1	74.93	78.23	71.89	72.57
set0+1+2	78.35	83.06	76.55	76.61
set0+...+3	79.20	83.85	81.55	82.05
set0+...+4	87.75	86.94	81.76	82.30
set0+...+5	94.59	93.27	83.86	84.58
set0+...+6	96.01	93.67	84.22	84.70
set0+...+7	97.15	94.54	84.58	85.59
set0+...+8	97.15	94.54	84.60	85.59
set0+1+2+3+5	94.59	93.27	83.86	84.58
set0+1+2+3+5+7	97.15	94.54	84.58	85.59

3.1 Confusion Set from Shuowen Jiezi

Shuowen Jiezi¹ (說文解字) is a dictionary of Chinese characters. Xu Shen (許慎), author of this dictionary, analyzed the characters according to the six lexicographical categories (六書). One major category is phono-semantic compound characters (形聲), which were created by combining a radical (形符) with a phonetic component (聲符). Characters with same phonetic components were collected to expand confusion sets, because they are by definition phonologically and visually similar. For example, the following characters share the same phonetic component “寺” ([si4, ‘temple’]) thus become confusion candidates (their actual pronunciation are given in brackets):

SWen: 侍[si4]持[chi2]恃[shi4]特[te4]時[shi2]...

It happens a phonetic component might not be atomic, which means it also has its own phonetic component. For example, 潔’ s phonetic component is 絜, but 絜’ s phonetic component is 丰. We tried two creation methods. The first one was created by collecting characters with the same phonetic component (referred to as SWen1), and the second one was the closure of SWen1 (referred to as SWen2).

Set 7 and 8 in Table 1 represent SWen1 and SWen2. Although they alone do not provide good coverage, unions including SWen sets can cover up to 97.15% errors in CSC 2013 Training set.

Closure set only cover one more error in CSC 2014 Training set. In order not to introduce too much noise, the closure SWen set is not recommended.

¹ <http://zh.wikisource.org/wiki/說文解字>

3.2 Confusion Set from the Four-Corner System

The Four-Corner System² (四角號碼) is an encoding system for Chinese characters. Digits 0~9 represent some typical shapes in character strokes. A Chinese character is encoded into 4 digits which represent the shapes found in its 4 corners. We collect characters in the same Four-Corner codes to expand confusion sets, because they are by definition visually similar. For example, the following characters are all encoded as 6080 in the Four-Corner System (shorten as Cor4 hereafter):

Cor4: 只囚貝足昷是員異買圖圍

Set 6 in Table 1 represents Cor4. Unfortunately unions including Cor4 do not cover more errors than set0~3+5+7. It is hard to say if The Four-Corner System is helpful or not.

3.3 Two-Character Confusion Set Expansion

To make a larger two-character confusion set, unigrams in the Chinese Google Ngram dataset were used instead of ASBC. But some issues should be handles before dataset creation, which are discussed in Section 3.3.1.

3.3.1 Google Ngram Dataset Preprocessing

Chinese Web 5-gram³ is real data released by Google Inc. who collected from all webpages in the World Wide Web which are unigram to 5-grams. Frequencies of these ngrams are also provided. Some examples from the Chinese Web 5-gram dataset are given here:

Unigram

稀釋剂 321928 ('thinner' in Simplified Chinese)

稀釋劑 17260 ('thinner' in Traditional Chinese)

Bigram

蒸发量 超过 869 ('the-amount-of-evaporation has-exceeded' in SC)

蒸發量 超過 69 ('the-amount-of-evaporation has-exceeded' in TC)

Trigram

能量 远 低于 727 ('energy far lower-than' in SC)

能量 遠 低於 113 ('energy far lower-than' in TC)

² 四角號碼列表 <http://code.web.idv.hk/misc/four.php>

³ <https://catalog ldc.upenn.edu/LDC2010T06>

4-gram

张贴 色情 图片 或 116 (‘posting pornographic images or’ in SC)

張貼 色情 圖片 或 73 (‘posting pornographic images or’ in TC)

5-gram

幸好 我们 发现 得 早 182 (‘fortunately we found-it DE early’ in SC)

幸好 我們 發現 得 早 155 (‘fortunately we found-it DE early’ in TC)

There are several issues with regard to using the Chinese Web 5-gram dataset in this task. First, the Chinese Web 5-gram dataset includes both Traditional and Simplified Chinese ngrams, but our experimental datasets are written in Traditional Chinese. To make full use of this dataset, we decide to translate every Simplified Chinese words into Traditional Chinese. Our translation method was simply table-lookup on the Simplified-to-Traditional Chinese word mappings provided by Wikipedia⁴. Note that the translation may not be perfect.

After translation, some ngrams become identical, such as 電視 and 电视 (‘television’) and all the Chinese Google Ngrams shown in the previous examples. Identical words are combined into one entry and their frequencies are merged.

3.3.2 Confusion Set Expansion by Google Ngram

The two-character confusion set in our baseline system was trained from ASBC. We tried to use unigram set in the Chinese Web 5-gram dataset to create a larger two-character confusion set.

The procedure is the same as in the baseline system development: collect all the two-character words in the Chinese Web unigram set, replace each character by its similar characters, collect all the new strings which also appear in the Chinese Web unigram set as the original word’s two-character confusion set.

In CSC 2014 training data, there are cases that both characters in a two-character word are misused, such as 也是 ([ye3-shi4], ‘also’) vs. 夜市 ([ye4-shi4], ‘night market’). We also performed such kind of replacement and collected legal similar words into the two-character confusion set.

4. Passage Likelihood Scoring

In CSC tasks held in 2013 and 2014, we tried bigram probability model to predict errors in sentences. The language generation model was trained from Academia Sinica Balanced

⁴ <http://zh.wikipedia.org/wiki/Wikipedia:繁簡處理>

Corpus. We found the volume and vocabulary of ASBC was not large enough. So we turn to use Chinese Google Ngram dataset.

4.1 Ngram Scoring Functions

Given a sentence (word-segmented, with or without errors) $S = \{w_1, w_2, \dots, w_m\}$, let $Gram(S, n)$ be the set of all n -grams containing in the sentence S , i.e. $Gram(S, n) = \{(w_i, w_{i+1}, \dots, w_{i+n-1}) | 1 \leq i \leq m-n+1\}$. We define **Google Ngram Frequency** $gnf(g)$ of a n -gram to be its frequency count provided in the Chinese Web 5-gram dataset. If it does not appear in that dataset, its value is defined as 0.

Five scoring functions $GS_*(S)$ were used to measure the likelihood of a sentence. Equation 1 is the definition of **raw frequency score** $GS_{raw}(S)$ which sums up the frequencies of all n -grams. Equation 2 and 3 give the definitions of **log frequency score** $GS_{logn}(S, n)$ and $GS_{log}(S)$ which sums up the logarithm of frequencies of all n -grams. Because large frequency tends to dominate the scores and then leads to bias, hopefully logarithm values can provide a moderate scoring. Note that we skip the ngrams which do not appear in the Chinese Web 5-gram dataset when calculating the log frequency score (or in another word, its log score is set to be 0).

$$GS_{raw}(S) = \sum_{n=2}^5 \left(\sum_{g \in Gram(S, n)} gnf(g) \right) \quad (1)$$

$$GS_{logn}(S, n) = \sum_{g \in Gram(S, n)} \log(gnf(g)) \quad (2)$$

$$GS_{log}(S) = \sum_{n=2}^5 GS_{logn}(S, n) \quad (3)$$

It is obvious that matching of a higher gram is more welcome than of a lower gram. To favor higher grams, we define the third scoring function **length-weighted log frequency score** $GS_{len}(S)$ which multiplies the log frequency score with n .

$$GS_{len}(S) = \sum_{n=2}^5 \left(n \times \sum_{g \in Gram(S, n)} \log(gnf(g)) \right) \quad (4)$$

We further tried two average scores where scores of the same n are averaged before summation. Equation 5 and 6 illustrate the logarithm and length-weighted versions, respectively.

$$GS_{\log av}(S) = \sum_{n=2}^5 \left(\frac{1}{|Gram(S, n)|} \times \sum_{g \in Gram(S, n)} \log(gnf(g)) \right) \quad (5)$$

$$GS_{lenav}(S) = \sum_{n=2}^5 \left(\frac{n}{|Gram(S, n)|} \times \sum_{g \in Gram(S, n)} \log(gnf(g)) \right) \quad (6)$$

We also tried a smoothing-like function to handle zero frequency. If a ngram does not appear in the Chinese Web 5-gram dataset, its log score is set to a negative constant ε . The *smoothed log frequency score* $gnf'(g)$ is defined as Equation 6.

$$gnf'(g) = \begin{cases} \text{if } gnf(g) = 0 & \varepsilon \\ \text{otherwise} & \log(gnf(g)) \end{cases} \quad (7)$$

Figure 1 demonstrates the detailed information and steps of compute the values of two of the scoring functions, log frequency score and length-weighted log frequency score, with or without smoothing, by using the first passage of B1-0143-1 as an example. As we can see, the smoothed length-weighted log frequency score can successfully identify the correct answer.

4.2 Threshold Learning

A replacement is considered to be “correct” if the score of the generated new passage is higher than the original’s to a certain degree. As described in Section 2.7, a pre-defined threshold is used to ensure that the new passage is far better than the original passage.

In CSC 2013 and 2014, this threshold was set by consulting classification rules learned by decision tree. In this paper, we try to observe the efficiency of thresholds in a more systematical way as follows.

Two kinds of thresholds were considered. The first one is for the score difference of the scores of the new passage and the original passage. Because the new passage must have a higher score than the original one, this value is always positive. The second one is for the ratio of the score difference to the original passage’s score. Because scores may be negative, we take its absolute value instead, i.e.

$$| (\text{score}_{\text{new}} - \text{score}_{\text{org}}) / \text{score}_{\text{org}} |.$$

B1-0143-1 妳還記得我們在高中在已樣的課嗎

Org, Segmented: 妳 還 記 得 我 們 在 高 中 在 已 樣 的 課 嗎

Rpl1, 妳→你, Segmented: 你 還 記 得 我 們 在 高 中 在 已 樣 的 課 嗎

Rpl2, 樣→聽, Segmented: 妳 還 記 得 我 們 在 高 中 在 已 聽 的 課 嗎

Rpl3, 已→一, Segmented: 妳 還 記 得 我 們 在 高 中 在 一 樣 的 課 嗎

(English meanings: 妳 you(female), 你 you, 還 still, 記 得 remember, 我們 we, 在 in, 高中 high-school, 已 already, 樣 pattern, 聽 listen, 一樣 same, 的 DE, 課 class, 嗎 Qpunc)

(Org: ‘Do you still remember that we were in the patterned class in high school?’)

(Rpl1: ‘Do you still remember that we were in the patterned class in high school?’)

(Rpl2: ‘Do you still remember that we were in the listened class in high school?’)

(Rpl3: ‘Do you still remember that we were in the same class in high school?’)

Google Ngram Information:

Bigram	<i>gnf</i>	log	Trigram	<i>gnf</i>	log
妳 還	337282	12.729	妳 還 記 得	22344	10.014
你 還	27319449	17.123	你 還 記 得	1127456	13.935
還 記 得	8552177	15.962	還 記 得 我 們	264628	12.486
記 得 我 們	756252	13.536	記 得 我 們 在	40942	10.620
我 們 在	24371694	17.009	在 高 中 在	843	6.737
在 高 中	838050	13.639	在 已 聽	61	4.111
高 中 在	100156	11.514	在 一 樣 的	19422	9.874
在 已	1193110	13.992	已 聽 的	1991	7.596
在 一 樣	41218	10.627	聽 的 課	8342	9.029
已 樣	1025	6.932	Trigram with <i>gnf</i> (.)=0		
已 聽	121888	11.710	我 們 在 高 中, 高 中 在 已, 高 中 在 一 樣, 在 已 樣, 已 樣 的, 樣 的 課, 一 樣 的 課, 的 課 嗎		
樣 的	3280256	15.003			
聽 的	5830567	15.579			
一 樣 的	35523054	17.386			
的 課	2695074	14.807			
課 嗎	0	---			

4-gram with <i>gnf</i> (.) > 0	<i>gnf</i>	log	5-gram with <i>gnf</i> (.) > 0	<i>gnf</i>	log
妳 還 記 得 我 們	896	6.798	你 還 記 得 我 們 在	2846	7.954
你 還 記 得 我 們	43508	10.680	還 記 得 我 們 在 高 中	78	4.357
還 記 得 我 們 在	16260	9.696			
記 得 我 們 在 高 中	238	5.472			

Figure 1. (a) Examples of Google Ngram Information in Scoring

List of scores

	GS_{log}	GS_{len}	GS'_{log}	GS'_{len}
Org	201.304	499.469	1.304	-290.531
Rpl1	221.456	575.321	31.456	-164.679
Rpl2	227.394	572.386	57.394	-127.614
Rpl3 (correct)	203.263	513.261	43.263	-126.739

Scoring details:

$$\begin{aligned}
 GS_{Log}(\text{Org}) &= (\log(\text{gnf}(\text{妳 還})) + \log(\text{gnf}(\text{還 記得})) + \dots + \log(\text{gnf}(\text{課 嗎})) + \\
 &\quad (\log(\text{gnf}(\text{妳 還 記得})) + \dots + \log(\text{gnf}(\text{的 課 嗎})) + \\
 &\quad (\log(\text{gnf}(\text{妳 還 記得 我們})) + \dots + \log(\text{gnf}(\text{樣 的 課 嗎})) + \\
 &\quad (\log(\text{gnf}(\text{妳 還 記得 我們 在})) + \dots + \log(\text{gnf}(\text{已 樣 的 課 嗎})) \\
 &= 12.729 + 15.962 + 13.536 + \dots + 15.003 + 14.807 + 0 + \\
 &\quad 10.014 + 12.486 + 10.620 + \dots + 0 + 0 + \\
 &\quad 6.798 + 9.696 + 5.472 + 0 + \dots + 0 + 0 + \\
 &\quad 0 + 4.357 + 0 + \dots + 0 \\
 &= 135.124 + 39.857 + 21.967 + 4.357 = \underline{201.304} \\
 GS_{Log}(\text{Rpl1}) &= (\log(\text{gnf}(\text{你 還})) + \log(\text{gnf}(\text{還 記得})) + \dots + \log(\text{gnf}(\text{課 嗎})) + \\
 &\quad (\log(\text{gnf}(\text{你 還 記得})) + \dots + \log(\text{gnf}(\text{的 課 嗎})) + \\
 &\quad (\log(\text{gnf}(\text{你 還 記得 我們})) + \dots + \log(\text{gnf}(\text{樣 的 課 嗎})) + \\
 &\quad (\log(\text{gnf}(\text{你 還 記得 我們 在})) + \dots + \log(\text{gnf}(\text{已 樣 的 課 嗎})) \\
 &= 139.518 + 43.778 + 25.849 + 12.310 = \underline{221.456} \\
 GS_{Log}(\text{Rpl2}) &= 140.477 + 60.594 + 21.967 + 4.358 = \underline{227.394} \\
 GS_{Log}(\text{Rpl3}) &= 127.208 + 49.731 + 21.967 + 4.358 = \underline{203.263} \\
 GS_{Len}(\text{Org}) &= 135.124 \times 2 + 39.857 \times 3 + 21.967 \times 4 + 4.357 \times 5 = \underline{499.469} \\
 GS_{Len}(\text{Rpl1}) &= 139.518 \times 2 + 43.778 \times 3 + 25.849 \times 4 + 12.310 \times 5 = \underline{575.321} \\
 GS_{Len}(\text{Rpl2}) &= 140.477 \times 2 + 60.594 \times 3 + 21.967 \times 4 + 4.358 \times 5 = \underline{572.386} \\
 GS_{Len}(\text{Rpl3}) &= 127.208 \times 2 + 49.731 \times 3 + 21.967 \times 4 + 4.358 \times 5 = \underline{513.261} \\
 GS'_{Log}(\text{Org}) &= 135.124 - 10 + 39.857 - 10 \times 6 + 21.967 - 10 \times 6 + 4.357 - 10 \times 7 \\
 &\quad (1 \text{ bigram, 6 trigrams, 6 fourgrams, and 7 fivegrams with } \text{gnf}(\cdot) = 0) \\
 &= 125.124 - 20.143 - 38.033 - 65.643 = \underline{1.304} \\
 GS'_{Log}(\text{Rpl1}) &= 139.518 - 10 + 43.778 - 10 \times 6 + 25.849 - 10 \times 6 + 12.310 - 10 \times 6 \\
 &= 129.518 - 16.222 - 34.151 - 47.690 = \underline{31.456} \\
 GS'_{Log}(\text{Rpl2}) &= 140.477 - 10 + 60.594 - 10 \times 3 + 21.967 - 10 \times 6 + 4.358 - 10 \times 7 \\
 &= 130.477 + 30.594 - 38.033 - 65.642 = \underline{57.394} \\
 GS'_{Log}(\text{Rpl3}) &= 127.208 - 10 + 49.731 - 10 \times 4 + 21.967 - 10 \times 5 + 4.358 - 10 \times 6 \\
 &= 117.208 + 9.731 - 28.033 - 55.642 = \underline{43.263} \\
 GS'_{Len}(\text{Org}) &= 125.124 \times 2 - 20.143 \times 3 - 38.033 \times 4 - 65.643 \times 5 = \underline{-290.531} \\
 GS'_{Len}(\text{Rpl1}) &= 129.518 \times 2 - 16.222 \times 3 - 34.151 \times 4 - 47.690 \times 5 = \underline{-164.679} \\
 GS'_{Len}(\text{Rpl2}) &= 130.477 \times 2 + 30.594 \times 3 - 38.033 \times 4 - 65.642 \times 5 = \underline{-127.614} \\
 GS'_{Len}(\text{Rpl3}) &= 117.208 \times 2 + 9.731 \times 3 - 28.033 \times 4 - 55.642 \times 5 = \underline{-126.739}
 \end{aligned}$$

Figure 1. (b) Details of Scoring Steps

A threshold is trained in the steps as follows. Under a scoring function, all replacements are sorted according to the score difference (or ratio). Largest values are ranked higher. Since each replacement is known to be “correct” or “incorrect”, precision, recall, and F-score at each rank can be decided. Choose the difference (or ratio) which achieves the highest F-score as the threshold.

Best F-scores under different scoring functions, smoothing strategies, and training data are shown in Table 2(a) and 2(b), where the first columns represent scoring functions introduced in Section 4.1. Meanings of labels in the second rows are as follows:

OL: no smoothing, at most one error report at one location

OP: no smoothing, at most one error report at one passage

ML: smoothing, at most one error report at one location

MP: smoothing, at most one error report at one passage

Table 2. Best F-Scores Achieved by Threshold Tuning

(a) Threshold Tuning on CSC 2013 Training Set

F-score	Difference				Ratio			
	OL	OP	ML	MP	OL	OP	ML	MP
GS_{raw}	3.23	3.23	---	---	3.39	2.36	---	---
$GS_{logn(2)}$	0.43	0.43	1.11	1.18	0.55	0.61	0.76	0.94
$GS_{logn(3)}$	10.74	10.27	22.25	22.22	6.18	7.49	12.68	17.09
$GS_{logn(4)}$	15.16	15.28	33.81	33.12	10.85	12.09	17.85	19.59
$GS_{logn(5)}$	10.28	9.63	21.38	21.96	9.79	9.66	11.50	13.02
GS_{log}	6.67	6.74	33.78	35.78	3.36	4.19	20.69	25.87
GS_{logav}	26.60	28.25	30.92	33.16	20.32	25.62	24.58	30.35
GS_{len}	9.93	9.86	42.75	44.06	4.83	5.50	25.52	31.34
GS_{lenav}	27.38	28.34	30.06	33.74	19.53	24.51	26.05	29.34

(b) Threshold Tuning on CSC 2014 Training Set

F-score	Difference				Ratio			
	OL	OP	ML	MP	OL	OP	ML	MP
GS_{raw}	3.31	2.82	---	---	3.08	2.73	---	---
$GS_{logn}(2)$	1.50	0.94	1.62	1.07	1.52	0.85	1.52	0.89
$GS_{logn}(3)$	7.17	6.84	10.61	9.66	5.81	6.13	7.71	8.75
$GS_{logn}(4)$	10.82	10.90	14.31	14.43	10.14	11.65	10.72	11.41
$GS_{logn}(5)$	7.89	7.73	8.73	8.35	9.44	9.08	6.32	5.38
GS_{log}	6.20	5.99	17.19	16.56	3.99	4.35	12.20	14.03
GS_{logav}	13.86	15.13	13.98	15.79	13.03	15.12	13.38	15.65
GS_{len}	7.98	7.66	22.07	21.65	5.04	5.65	14.60	16.93
GS_{lenav}	14.03	15.35	14.11	15.84	12.91	15.14	13.61	15.52

As we can see in Table 2, smoothing and logarithm did improve the performance. Using thresholds of score differences was better than using thresholds of ratios. Among the 9 scoring functions, length-weighted log frequency score GS_{len} outperformed other functions. However, averaging at each n level harmed the performance.

To our surprise, bigram model $GS_{logn}(2)$ was not very useful. However, 4-gram model $GS_{logn}(4)$ alone could achieve pretty good performance. Moreover, the characteristics of CSC 2013 training set and CSC 2014 training set are quite different. F-cores on CSC 2014 data sets were much lower.

5. Experiments

5.1 Datasets

Four benchmarks are used to evaluate our systems: the training set and test set in Chinese Spelling Check Evaluation at SIGHAN Bake-off 2013 (Wu *et al.*, 2013), and the training set and test set in CLP-2014 Chinese Spelling Check Evaluation (Yu *et al.*, 2014). They are referred to as CSC 2013 and 2014 datasets in this paper. Number of topics and errors containing in these datasets are listed in Table 3.

Table 3. Number of Topics and Errors in CSC 2013 and 2014 Datasets

Dataset	#Topics	#Errors
CSC 2013 Training	350	351
CSC 2013 Test	1000	1464
CSC 2014 Training	3434	5280
CSC 2014 Test	531	791

5.2 Evaluation Metrics

There are two subtasks in CSC Task: error detection and error correction. Error detection subtask evaluates the correctness of detected error locations. Error correction subtask evaluates the correctness of locations and proposed corrections.

The metrics are evaluated in both levels by the following metrics:

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

$$\text{F1-Score} = 2 * \text{Precision} * \text{Recall} / (\text{Precision} + \text{Recall})$$

Note that the unit of “correctness” is topic. It only counts the topics whose errors are all successfully corrected with no false alarm.

5.3 Experimental Results

All combinations of system settings have been evaluated on all the datasets. Table 4 shows the runs achieving the best F1-scores according to each subtask, dataset, and scoring functions. The labels of system settings are defined as follows (cf. Section 3.2):

Ranking and threshold setting

diff: ranking by the score difference

ratio: ranking by the score ratio

Smoothing Strategy

O: no smoothing

M: smoothing

Detection unit

N: at most one error in one topic, no threshold

Q: at most one error in one topic, filtered by threshold

P: at most one error in one passage, filtered by threshold

L: at most one error at each location, filtered by threshold

More precisely, Table 4(a)~4(d) shows the experimental results of *error detection* evaluated on CSC 2013 training set, CSC 2013 test set, CSC 2014 training set, and CSC 2014 test set, respectively. Table 4(e)~4(h) shows the experimental results of *error correction* evaluated on CSC 2013 training set, CSC 2013 test set, CSC 2014 training set, and CSC 2014 test set, respectively.

Almost all results support similar conclusions as we made in Section 4.2: the best system uses the smoothed length-weighted log frequency score, ranking by score differences without threshold ($GS_{len,diff,M,N}$). Thresholds are not helpful except on CSC 2014 test set.

Table 4. Experimental Results on CSC2013 and 2014 Datasets

(a) Error-Detection, CSC2013 Training Set

Scoring	System	P	R	F	Acc
GS_{raw}	ratio,O,N	100.00	7.71	14.32	7.71
$GS_{logn(2)}$	diff,M,N	100.00	9.71	17.71	9.71
$GS_{logn(3)}$	diff,M,N	100.00	30.00	46.15	30.00
$GS_{logn(4)}$	diff,M,N	100.00	30.00	46.15	30.00
$GS_{logn(5)}$	diff,M,N	100.00	18.57	31.33	18.57
GS_{log}	diff,M,N	100.00	42.00	59.15	42.00
GS_{logav}	diff,M,N	100.00	37.71	54.77	37.71
GS_{len}	diff,M,N	100.00	46.57	63.55	46.57
GS_{lenav}	diff,M,N	100.00	36.00	52.94	36.00

(b) Error-Detection, CSC2013 Test Set

Scoring	System	P	R	F	Acc
GS_{raw}	ratio,O,N	100.00	4.80	9.16	4.80
$GS_{logn(2)}$	diff,M,N	100.00	5.10	9.71	5.10
$GS_{logn(3)}$	diff,M,N	100.00	18.40	31.08	18.40
$GS_{logn(4)}$	diff,M,N	100.00	18.20	30.80	18.20
$GS_{logn(5)}$	diff,M,Q	100.00	11.90	21.27	11.90
GS_{log}	diff,M,N	100.00	25.90	41.14	25.90

GS_{logav}	diff,M,N	100.00	24.80	39.74	24.80
GS_{len}	diff,M,N	100.00	28.80	44.72	28.80
GS_{lenav}	diff,M,N	100.00	25.00	40.00	25.00

(c) Error-Detection, CSC2014 Training Set

Scoring	System	P	R	F	Acc
GS_{raw}	ratio,M,N	98.21	4.80	9.16	4.80
$GS_{logn}(2)$	diff,M,N	97.22	3.06	5.93	3.05
$GS_{logn}(3)$	diff,M,N	99.31	12.64	22.42	12.63
$GS_{logn}(4)$	diff,M,N	99.38	13.98	24.51	13.97
$GS_{logn}(5)$	diff,M,N	98.72	6.73	12.60	6.72
GS_{log}	diff,M,N	99.52	18.29	30.90	18.27
GS_{logav}	diff,M,N	99.47	16.37	28.11	16.35
GS_{len}	diff,M,N	99.59	21.40	35.23	21.38
GS_{lenav}	diff,M,N	99.46	15.96	27.50	15.94

(d) Error-Detection, CSC2014 Test Set

Scoring	System	P	R	F	Acc
GS_{raw}	ratio,M,Q	5.40	5.46	5.43	4.90
$GS_{logn}(2)$	diff,M,Q	6.45	3.01	4.11	29.66
$GS_{logn}(3)$	diff,M,Q	17.28	9.79	12.50	31.45
$GS_{logn}(4)$	diff,M,Q	14.88	14.88	14.88	14.88
$GS_{logn}(5)$	diff,M,Q	8.85	8.85	8.85	8.85
GS_{log}	diff,M,N	17.94	20.72	19.23	12.99
GS_{logav}	ratio,M,Q	19.21	18.27	18.73	20.72
GS_{len}	diff,M,Q	25.63	19.21	21.96	31.73
GS_{lenav}	diff,M,Q	19.63	17.89	18.72	22.32

(e) Error-Correction, CSC2013 Training Set

Scoring	System	P	R	F	Acc
GS_{raw}	diff,O,N	100.00	2.86	5.56	2.86
$GS_{logn}(2)$	diff,M,L	100.00	0.86	1.70	0.86

$GS_{logn(3)}$	diff,M,N	100.00	20.29	33.73	20.29
$GS_{logn(4)}$	diff,M,N	100.00	23.71	38.34	23.71
$GS_{logn(5)}$	diff,M,N	100.00	15.71	27.16	15.71
GS_{log}	diff,M,N	100.00	32.57	49.14	32.57
GS_{logav}	diff,M,N	100.00	30.57	46.83	30.57
GS_{len}	diff,M,N	100.00	41.71	58.87	41.71
GS_{lenav}	diff,M,N	100.00	30.57	46.83	30.57

(f) Error- Correction, CSC2013 Test Set

Scoring	System	P	R	F	Acc
GS_{raw}	ratio,O,N	100.00	0.90	1.78	0.90
$GS_{logn(2)}$	ratio,M,N	100.00	0.50	1.00	0.50
$GS_{logn(3)}$	diff,M,N	100.00	12.50	22.22	12.50
$GS_{logn(4)}$	diff,M,N	100.00	14.80	25.78	14.80
$GS_{logn(5)}$	diff,M,Q	100.00	10.00	18.18	10.00
GS_{log}	diff,M,N	100.00	19.20	32.21	19.20
GS_{logav}	diff,M,N	100.00	20.10	33.47	20.10
GS_{len}	diff,M,N	100.00	23.60	38.19	23.60
GS_{lenav}	diff,M,N	100.00	20.70	34.30	20.70

(g) Error- Correction, CSC2014 Training Set

Scoring	System	P	R	F	Acc
GS_{raw}	diff,O,N	95.38	1.81	3.54	1.80
$GS_{logn(2)}$	ratio,M,N	83.33	0.44	0.87	0.44
$GS_{logn(3)}$	diff,M,N	98.68	6.52	12.24	6.52
$GS_{logn(4)}$	diff,M,N	99.10	9.61	17.52	9.60
$GS_{logn(5)}$	diff,M,N	98.13	4.57	8.74	4.57
GS_{log}	diff,M,N	99.26	11.76	21.04	11.75
GS_{logav}	diff,M,N	99.21	11.04	19.86	11.03
GS_{len}	diff,M,N	99.42	15.03	26.11	15.01
GS_{lenav}	diff,M,N	99.22	11.12	20.01	11.11

(h) Error- Correction, CSC2014 Test Set

Scoring	System	P	R	F	Acc
GS_{raw}	ratio,O,Q	2.90	2.82	2.86	4.05
$GS_{logn(2)}$	diff,M,Q	1.28	0.56	0.78	28.44
$GS_{logn(3)}$	diff,M,Q	11.39	6.03	7.88	29.57
$GS_{logn(4)}$	diff,M,Q	11.55	11.11	11.32	12.99
$GS_{logn(5)}$	diff,M,Q	6.20	6.03	6.11	7.44
GS_{log}	diff,M,P	14.75	8.47	10.77	29.76
GS_{logav}	diff,M,Q	15.03	12.43	13.61	21.09
GS_{len}	diff,M,Q	21.28	15.07	17.64	29.66
GS_{lenav}	diff,M,Q	15.62	13.56	14.52	20.15

By observing the text in the benchmarks, it seems that the sentences in CSC 2014 datasets were written by non-Chinese-native speakers. It means that (1) even the corrected sentences may not be natural enough, so ngram model cannot predict successfully; (2) some errors are so common that appear in many sentences, so hand-crafted rules may be more successful.

6. Conclusion

In this paper, we proposed two resources to expand confusion sets which improved the error coverage up to 97.17% in CSC training set. We also proposed a method to build a larger two-character confusion set. Nine scoring functions using Google Ngram frequency information were also introduced. Among them, length- weighted log frequency score greatly improved our baseline system on CSC 2013 datasets.

Although that the methods proposed in this paper do not perform well enough on CSC 2014 datasets, we still think that our method can cooperate with hand-crafted rules (as top CSC systems did in CSC 2014), which becomes our future work.

References

- de Amorim, R.C., & Zampieri, M. (2013). Effective Spell Checking Methods Using Clustering Algorithms. *Recent Advances in Natural Language Processing*, 7-13.
- Blair, C. (1960). A program for correcting spelling errors. *Information and Control*, 3, 60-67.
- Carlson, A., & Fette, I. (2007). Memory-Based Context-Sensitive Spelling Correction at Web Scale. In *Proceedings of the 6th International Conference on Machine Learning and Applications*, 166-171.

- Carlson, A., Rosen, J., & Roth, D. (2001). Scaling up context-sensitive text correction. In *Proceedings of the 13th Innovative Applications of Artificial Intelligence Conference*, 45-50.
- Chang, C.H. (1994). A pilot study on automatic chinese spelling error correction. *Journal of Chinese Language and Computing*, 4, 143-149.
- Chen, K.J., Huang, C.R., Chang, L.P., & Hsu, H.L. (1996). Sinica Corpus: Design Methodology for Balanced Corpora. In *Proceeding of the 11th Pacific Asia Conference on Language, Information and Computation*, 167-176.
- Chen, Q., Li, M., & Zhou, M. (2007). Improving Query Spelling Correction Using Web Search Results. In *Proceedings of the 2007 Conference on Empirical Methods in Natural Language (EMNLP-2007)*, 181-189.
- Chen, Y.Z., Wu, S.H., Yang, P.C., Ku, T., & Chen, G.D. (2011). Improve the detection of improperly used Chinese characters in students' essays with error model. *International Journal of Continuing Engineering Education and Life-Long Learning*, 21(1), 103-116.
- Cucerzan, S., & Brill, E. (2004). Spelling correction as an iterative process that exploits the collective knowledge of web users. In *Proceedings of EMNLP*, 293-300.
- Damerau, F. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7, 171-176.
- Deorowicz, S., & Ciura, M. G. (2005). Correcting Spelling Errors by Modelling Their Causes. *International Journal of Applied Mathematics and Computer Science*, 15(2), 275-285.
- Golding, A., & Roth, D. (1999). A winnow-based approach to context-sensitive spelling correction. *Machine Learning*, 34(1-3), 107-130.
- Islam, A., & Inkpen, D. (2009). Real-word spelling correction using googleweb 1t 3-grams. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP-2009)*, 1241-1249.
- Li, M., Zhang, Y., Zhu, M.H., & Zhou, M. (2006). Exploring distributional similarity based models for query spelling correction. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, 1025-1032.
- Liu, W., Allison, B., & Guthrie, L. (2008). Professor or screaming beast? Detecting words misuse in Chinese. *The 6th edition of the Language Resources and Evaluation Conference*.
- Liu, C.L., Lai, M.H., Tien, K.W., Chuang, Y.H., Wu, S.H., & Lee, C.Y. (2011). Visually and phonologically similar characters in incorrect Chinese words: Analyses, identification, and applications. *ACM Transactions on Asian Language Information Processing*, 10(2), 1-39.
- Mitton, R. (1996). *English Spelling and the Computer*. Harlow, Essex: Longman Group.
- Mitton, R. (2008). Ordering the Suggestions of a Spellchecker Without Using Context. *Natural Language Engineering*, 15(2), 173-192.

- Pirinen, T., & Linden, K. (2010). Creating and weighting hunspell dictionaries as finite-state automata. *Investigationes Linguisticae*, 21.
- Verberne, S. (2002). *Context-sensitive spell checking based on word trigram probabilities*, Master thesis, University of Nijmegen.
- Whitelaw, C., Hutchinson, B., Chung, G.Y., & Ellis, G. (2009). Using the Web for Language Independent Spellchecking and Autocorrection. In *Proceedings Of Conference On Empirical Methods In Natural Language Processing (EMNLP-2009)*, 890-899.
- Wu, S.H., Chen, Y.Z., Yang, P.C., Ku, T., & Liu, C.L. (2010). Reducing the False Alarm Rate of Chinese Character Error Detection and Correction. In *Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP 2010)*, 54-61.
- Wu, S.H., Liu, C.L., & Lee, L.H. (2013). Chinese Spelling Check Evaluation at SIGHAN Bake-off 2013. In *Proceedings of the 7th SIGHAN Workshop on Chinese Language Processing (SIGHAN'13)*, 35-42.
- Yu, L.C., Lee, L.H., Tseng, Y.H., & Chen, H.H. (2014). Overview of SIGHAN 2014 Bake-off for Chinese Spelling Check. In *Proceedings of the 3rd CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP'14)*, 126-132.
- Zhang, L., Zhou, M., Huang, C.N., & Pan, H.H. (2000). Automatic detecting/correcting errors in Chinese text by an approximate word-matching algorithm. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, 248-254.

Automatically Detecting Syntactic Errors in Sentences Written by Learners of Chinese as a Foreign Language

Tao-Hsing CHANG^{*}, Yao-Ting SUNG⁺ and Jia-Fei HONG[#]

Abstract

This paper proposed a method that can automatically detect syntax errors in Chinese sentences. The algorithm for identifying syntax errors proposed in this study is known as KNGED, which uses a large database of rules to identify whether syntax errors exist in a sentence. The rules were generated either manually or automatically. This paper further proposed an algorithm for identifying the type of error that a sentence contained. Experimental results shown that the false positive rate and F1-measure of the proposed method for detecting syntax errors in Chinese sentences are 0.90 and 0.65.

Keywords: Syntactic Errors, Chinese Grammar, Chinese Written Corpus.

1. Introduction

The teaching of languages has always been an important area of research and a commercially viable market. An important topic of research is the means by which the linguistic abilities of learners can be enhanced efficiently. This is especially so for learners of foreign languages, who have to learn the target language within a limited time period while being in a non-immersive learning environment, unlike the ample time they had for learning their native language. Contrastive linguistics is a tool that can be used to improve the efficiency of learning a foreign language effectively. Since most learners would already have well-developed capabilities in their native language, pointing out and analyzing the differences between the native and foreign language can help learners to understand the differences between the two, thereby facilitating the conversion from the former to the latter.

^{*} Department of Computer Science and Information Engineering, National Kaohsiung University of Applied Sciences, Taiwan

E-mail: changth@gm.kuas.edu.tw

⁺ Department of Educational Psychology and Counseling, National Taiwan Normal University, Taiwan

E-mail: sungtc@ntnu.edu.tw

[#] Department of Applied Chinese Language and Culture, National Taiwan Normal University, Taiwan

E-mail: jiafeihong@ntnu.edu.tw

However, simply understanding the differences between two languages does not mean that a person can make the conversion from one to the other effectively in real-life usage situations. In comparative linguistics, two phenomena often appear in the patterns of language usage. First, since the types and quantity of differences are substantial, learners may not necessarily notice each and every difference between the native and foreign languages when using the latter. Second, when learners are not familiar with the linguistic differences, they become susceptible to the phenomenon of language transfer.

An example is the use of suffixes that signal tenses of English verbs, which has no parallel in the grammar of Chinese verbs. Although learners of English are aware that they need to pay attention to the tenses of verbs, they often make the mistake of using the wrong tense. Learners must keep practicing to become familiar with the relevant linguistic differences. During the learning process, teachers must also point out the errors committed. Only then can learners internalize the differences and gain the ability to use the foreign language. Unfortunately, the labor costs of making these corrections are high. In the existing educational model where one teacher is often responsible for teaching many students, it is not possible for him/her to conduct intensive practices for all students, nor correct the errors of each individual student.

To overcome this issue, many studies have proposed the concept of “automatic detection of learners’ errors during language usage.” These methods mainly employ detection models that target word or syntax errors. Many useful methods have already been proposed for the automatic detection and correction of English syntax errors. Some of these rely on having an excellent grammar parser. If the parser is unable to deconstruct a sentence completely and convert it to a parsing tree, then some syntax errors in this sentence will fail to be detected and corrected. However, it is difficult to apply such a concept to the issue of identifying Chinese syntax errors for two main reasons. First, it is difficult to identify the limits of a single sentence. For English sentences, the contents between two periods can be treated as a syntactic structure and unit of analysis. For Chinese sentences, a segment that ends with a comma can be a sentence with a complete syntactic and semantic structure, just a clause, or even a phrase. Second, the Chinese language contains many more syntactical changes, making it difficult for learners to distinguish between correct and erroneous usage. Hence, using a grammar parser for learning Chinese is not as effective as using one for learning English. These reasons make the detection and correction of errors in Chinese sentences more difficult than in those of English.

We believe that the identification of patterns in syntax errors is a possible solution. Common syntax errors usually involve part of a sentence rather than its overall structure. This situation is particularly pronounced for syntax errors committed by learners of a second language, the root cause of which is the phenomenon of language transfer. The following is an

example of an error that is often committed by Korean students when writing Chinese sentences.

Erroneous: “他來台北一年讀書了” (He has been in Taipei a year for studying.)

Correct: “他來台北讀書一年了” (He has been studying in Taipei for a year.)

In Korean, a temporal noun is always placed before the verb. As a result, many continue to do so when writing Chinese sentences, thus committing errors. If this and other commonly made errors can be compiled and sorted into general categories, further analysis can be done to determine the identification rules for each category of errors. If part of a sentence contains a grammatical structure that may be flagged by an identification rule, then that structure is likely to be erroneous. When sufficient identification rules have been compiled, a comparison of written sentences with the rules base will highlight those with syntax errors. Statistical methods can also be used to analyze the large number of sentences contained in learners' corpora to identify frequently occurring grammatical structures. The larger the corpus, the bigger the number of identification rules that can be generated, which in turn help to detect more errors.

The main aim of this paper is to propose a method that can automatically detect syntax errors in Chinese sentences and then state the type of error that has been committed. In terms of framework, this method employs learners' writing corpora as the basis and two methods to generate rules for identifying syntax errors. In the first method, linguistic experts generate rules by examining corpora through a system; the second method uses formulas to establish rules automatically through the application of statistical methods to corpora. After establishing the rules, we applied them to determine whether a sentence was erroneous. For erroneous sentences, we further proposed an algorithm for identifying the type of error that the sentence contained.

The organization of the rest of this paper is as follows: an analysis of related studies and their impact on our research motivation is done in Section 2; the corpora used in this study are listed in Section 3, with detailed explanations of a learners' corpus that has been specially created to identify erroneous sentences written by those for whom Chinese is a second language; manually identified rules created by this study are also introduced in the section, together with the method of using formulas to automatically establish identification rules; the proposed algorithm for automatic identification of erroneous sentences is also explained in the section; the effectiveness of the proposed approach is illustrated in Section 4; and Section 5 is the conclusion.

2. Related Works

Syntax errors are usually classified as belonging to either the category of “language form” or “surface structure.” The former uses the language subsystems as the framework by which to classify the type of error. Specifically, this refers to errors in parts of speech (POS), syntax and semantics. The latter uses the structural method to classify the type of error, that is, by comparing the erroneous and correct forms. Surface structure errors are generally divided into four types: omissions, erroneous additions, overpresentations and misorders (Dulay, Burt, & Krashen, 1982; James, 1998).

Many analytical studies have been done on errors made by learners. One of the most famous English learners’ corpora is the Cambridge Learner Corpus (CLC), with as many as 16 million words having been tagged as erroneous. The three most common types of errors include wrong selection of words, wrong prepositions, and wrong qualifiers (Nicholls, 2003). After 200 learners for whom English is a second language had taken writing ability tests, Donahue (2001) analyzed their performance and compared his findings with the linguistic errors made by native English speakers as proposed by Connors and Lunsfor (1988). Donahue found that the most common types of errors made by non-native versus native English speakers were different. For the former, these included mistakes in the use of commas or words, as well as omission of words.

In recent years, common syntax errors made by learners for whom Chinese is a second language have become a popular research topic. Wang (2011) indicates that for Chinese language learners who are native English speakers, the most common syntax errors include the omission of language elements, wrong word order, and structural errors. Cheng, Yu & Chen (2014) used the corpus of the Chinese Proficiency Test (HSK), which comprised 35,884 erroneous sentences in total, to analyze the types of syntax errors. The study found that the most common problems involved wrong word order, as well as omission of adverbial elements and predicates.

With the development of natural language processing technologies over the past decade, various researches have been done and tools for the automatic detection of English syntax errors have been proposed. The most common types of errors detected by these studies involve prepositions (Eeg-Olofsson & Knuttson, 2003; Tetreault & Chodorow, 2008; Gamon *et al.*, 2009; De Felice & Pulman, 2009; Dale, Anisimoff, & Narroway, 2012; Ng *et al.*, 2013), articles (Gamon *et al.*, 2009; Dale & Kilgarriff, 2011; Ng *et al.*, 2013), and qualifiers (Dale *et al.*, 2012; Ng *et al.*, 2013).

These tools automatically detect errors in the learners’ usage of qualifiers, articles, and prepositions, and then correct learners’ grammatical errors. By using these tools, foreign language learners in mastering the correct grammar and are useful for the improvement of

writing skills (Chodorow *et al.*, 2012; Leacock *et al.*, 2010). However, there have been very few studies on learners' corpora for the automatic detection of Chinese grammatical errors. Cheng *et al.* (2014) and Yu & Chen (2012) had used the Chinese sentences included in the HSK corpus for dynamic composition to develop detection techniques for errors in word order. For the method proposed by Lee *et al.* (2014), other than the HSK corpus for dynamic composition, the study had also included manual rules for common Chinese erroneous sentences when developing their system for detecting various errors in sentence construction and grammar.

Three conclusions can be derived from the aforementioned literature review. First, most studies have classified the types of syntax errors in terms of grammar or form, for example, omission of prepositions and redundancy of articles. Second, for the identification of errors, automatic detection methods make use of either manually established rules or statistical models. The identification results of the rule-based method detects some error types well, but most error types are such that this method does not capture them (Lee *et al.*, 2013). On the other hand, the statistical approach requires a considerably large learners' corpus to be effective. Third, there are very few learners' corpora for Chinese learners, and methods involving the use of statistical models to generate rules for identifying errors are even rarer.

3. Method

The algorithm for identifying syntax errors proposed in this study is known as KNGED, which uses a large database of rules to identify whether syntax errors exist in a sentence. The rules were generated either manually or automatically, the details of which will be elaborated upon in Subsections 3.2 and 3.3 respectively. Data sets of erroneous sentences had to be used during the rule-generating process. This study made use of two such data sets to generate identification rules for syntax errors: (i) dry run data (hereinafter referred to as TEA1-DRY) from the Shared Task on Grammatical Error Diagnosis for Learning Chinese as a Foreign Language (hereinafter referred to as NLPTEA1-CFL), which was organized by the 1st Workshop on Natural Language Processing Techniques for Educational Applications; and (ii) the Chinese Written Corpus (CWC) that we had developed, which will be described in detail in the next subsection.

3.1 Chinese Written Corpus

The CWC comprises 1,147 essays divided into two data subsets, with a total of approximately 750,000 words. Within each data set are essays on the same topic written by different authors who are expatriates learning Chinese in one of 11 Chinese language center of 11 universities in Taiwan. This group of authors had very diverse linguistic backgrounds; the total number of different native languages in it was 37. The texts were collected and compiled between

September 2010 and June 2013. Each essay was graded by two trained raters using the criteria from the *Chinese Composition Scoring Standard* developed by Hsiung et al. (2014). These criteria reference the classification structure of ACTFL (2012) and are prescribed for rating Chinese essays written by expatriates for whom Chinese is a second language. Specifically, writing abilities are rated as “distinguished,” “superior,” “advanced,” “intermediate,” or “novice.” The latter three grades are in turn subdivided into “high,” “medium,” and “low,” yielding 11 levels in total.

Each Chinese sentence of every essay in the CWC had undergone tagging for segmentation and POS based on WECA system (Chang, Sung, & Lee, 2012), followed by the correction of errors by trained taggers. Forty-eight POS tags were used, including the 46 simplified tags for Chinese POS as defined in CKIP (1993), the verb nominalization tag *Nv*, and the unknown POS tag *b*. Each sentence had been checked by the taggers for syntax errors. If found, the position and type of error were tagged accordingly, together with the corrected sentence. The main types of errors included erroneous additions/errors of redundancy, omissions, incorrect word order, and erroneous word selection.

3.2 Automatic Machine-generated Rules

The assumptions for our proposed method were based on two pieces of observed information. First, some of the erroneous positions and terms within a sentence are related to the preceding or subsequent word or POS. Second, most errors will occur repeatedly if the corpus is sufficiently large. Hence, the proposed method first examines all the possible patterns for syntax errors that can be generated by an erroneous sentence. Next, each pattern is individually checked to see if it appears in any other sentences within the corpus. A pattern is treated as a candidate rule if it occurs more than once. The following sentence is an example:

這些 地方 是 在 日本 (These places are located in Japan)

Neqa Na SHI P Nc

The tags below the sentence are the POS of each word. In the corpus, the “是” (are) character in the sentence was marked as being an error of the redundant type. Based on the aforementioned assumptions, all 32 possible combinations based on the word “是,” its POS tag “SHI,” and the preceding or subsequent word or POS tag are listed in Figure 1.

The symbol “+” in the figure indicates that the preceding/subsequent word/POS tag is immediately adjacent to the erroneous position, while the symbol “>” indicates that the preceding/subsequent word/POS tag is not immediately adjacent to the erroneous position. Each combination is treated as a candidate identification rule. The corrected pattern

corresponding to the combination is denoted as correction rule. For instance, the correction rule for candidate rule “Na + SHI + P” is “Na + P”.

The 32 candidate rules can be subjected to a further conditional test. A recurring pattern r is an identification rule if the following conditions are met:

$$\text{FreqInErr}(r) \geq p \text{ and } \text{Reliability}(r) \geq k, \text{ where } \text{Reliability}(r) = \text{FreqInCol}(re)/\text{FreqInCor}(r)$$

$\text{FreqInErr}(r)$ represents the number of times that rule r applies to the erroneous sentences which are identified by rule r . $\text{FreqInCor}(x)$ represents the number of corrected sentences in the corpus that complies with the rule r . re represents the correction rule for rule r . Parameters p and k are thresholds obtained during the experiment.

- | | |
|--------------------|----------------------|
| (1) 這些 > 是 + 在 | (17) Neqa > 是 + 在 |
| (2) 這些 > 是 + P | (18) Neqa > 是 + P |
| (3) 這些 > 是 > 日本 | (19) Neqa > 是 > 日本 |
| (4) 這些 > 是 > Nc | (20) Neqa > 是 > Nc |
| (5) 地方 + 是 + 在 | (21) Na + 是 + 在 |
| (6) 地方 + 是 + P | (22) Na + 是 + P |
| (7) 地方 + 是 > 日本 | (23) Na + 是 > 日本 |
| (8) 地方 + 是 > Nc | (24) Na + 是 > Nc |
| (9) 這些 > SHI + 在 | (25) Neqa > SHI + 在 |
| (10) 這些 > SHI + P | (26) Neqa > SHI + P |
| (11) 這些 > SHI > 日本 | (27) Neqa > SHI > 日本 |
| (12) 這些 > SHI > Nc | (28) Neqa > SHI > Nc |
| (13) 地方 + SHI + 在 | (29) Na + SHI + 在 |
| (14) 地方 + SHI + P | (30) Na + SHI + P |
| (15) 地方 + SHI > 日本 | (31) Na + SHI > 日本 |
| (16) 地方 + SHI > Nc | (32) Na + SHI > Nc |

Figure 1. Examples of machine-generated candidate identification rules

If the value of p is large, it indicates that more erroneous sentences contain the possible rule and hence, it should be included in the database of identification rules. In other words, the possible rule r should not be a random product that appears after the combinations have been listed. If the value of k is large, it indicates that a smaller ratio of false alarms will be generated when the possible rule r is used to identify erroneous sentences. In other words, the accuracy rate of identification will be higher. Using the 32 rules in Figure 1 as an example, if p and k are both set at 2, only 11 of the rules will be included as identification rules for errors

(please refer to Figure 2). When an identification rule for errors is included in the rules database, its corresponding correction rule will also be included.

- | | |
|---------------------|----------------------|
| (9) 這些 > SHI + 在 | (28) Neqa > SHI > Nc |
| (10) 這些 > SHI + P | (29) Na + SHI + 在 |
| (12) 這些 > SHI > Nc | (30) Na + SHI + P |
| (14) 地方 + SHI + P | (31) Na + SHI > 日本 |
| (25) Neqa > SHI + 在 | (32) Na + SHI > Nc |
| (26) Neqa > SHI + P | |

Figure 2. Rules from Fig. 1 that are added to the rule base after screening

Theoretically, the length of a rule extracted using this method need not be restricted to one preceding/subsequent word/POS tag. However, since there are many erroneous sentences, the possible rules that can be generated will be too numerous, making the computation process too time consuming. Therefore, in terms of the format of the rule, this study only considered the immediately preceding/subsequent word/POS tag. Given this premise, the automatic machine-generated method only generated rules for two types of errors: redundancy and omission. Moreover, these rules were produced based on CWC.

In addition, we observed that many examples of the selection type of error involved the wrong use of a unit, for example, “一個公車” (a bus) instead of “一輛公車.” So, we compiled all the units that are used with each noun from the Sinica corpus (Chen, Huang, Chang, & Hsu, 1996). Since each noun can be matched with more than one type of unit, all units that can be used were included in the database of units. If one of the patterns “Neu + Nf + Na” or “Neu + Nf > DE + Na” appears in a sentence, the words corresponding to the two POS—Nf and Na—will be treated as the unit and designated noun respectively. The pair formed by the unit and designated noun of this pattern is then sent to the database of units for checking. If the pair has not appeared previously, it means that an error of the selection type has been detected. The correct pair of unit and designated noun is then treated as the rule for correction.

3.3 Manually-generated Rules

All manually-generated rules are established by linguistic experts through the following four steps. First, the experts observed the erroneous sentences in TEA1-DRY and then listed the candidate rules for identifying and correcting syntax errors. Next, they used an inspection program to analyze whether each syntactic rule is correct. The program would indicate the number of sentences that satisfy the three separate conditions stipulated in the CWC: (i) the number of erroneous sentences that complied with a rule identifying wrong syntax; (ii) the number of corrected sentences that complied with the rule for correction; and (iii) the number of corrected sentences that complied with the rule for identification.

An effective rule for identifying and correcting grammatical errors must generate as many results as possible under the first and second conditions, but as few results as possible under the third condition. If more sentences satisfy the first condition, it means that the rule can identify more of the erroneous sentences. On the other hand, if more sentences comply with the third condition, it means that the rules for error identification will wrongly treat more of the correct sentences as being erroneous. Hence, the smaller the number of sentences identified under the third condition, the better are the results. If many sentences satisfy the second condition, it means that the rules for correction are common and correct forms of usage, thus their general presence in the corpus. Consequently, the likelihood of the rules for correction being effective will also be higher.

The format of the manually-generated identification and correction rules is similar to the machine-generated rules, although there is no restriction on the number of preceding/subsequent words/POS. Hence, the former has a higher accuracy rate for detection. However, non-limitation on the number of preceding/subsequent words/POS also resulted in rules with sequential errors. Eight hundred and forty manually-generated identification rules were used in this study, which could be broken down into the following types: 90 missing, 73 redundant, 51 selection, and 626 wrong order. Since the proposed method for automatic machine-generated rules could not generate rules with disorder errors, the number for this type of manually generated rules far exceeded the other types.

3.4 Detection of Erroneous Sentences and Algorithm for Detected Types of Errors

After setting up the rule base generated by machine and manually, each test sentence was compared with the rules to determine if it was erroneous and if so, the type of error and rules for correction. Since one sentence could be simultaneously identified by multiple rules, we designed an algorithm shown in Figure 3 to identify the most likely error.

KNGED (sentence S , integer y)

Begin

maximum = 0;

rule-pointer = null;

Tag the segmentation and POS of the sentence using WECA_n;

for every identification rule r_i for the selection error type in the rule base

if sentence S contains any structure that can be identified by r_i

then tag the erroneous portion of sentence S and **return** the corrected sentence;

for every identification rule r_i for the disorder error type in the rule base

if sentence S contains any structure that can be identified by r_i

then tag the erroneous portion of sentence S and **return** the corrected sentence;

for every identification rule r_i for the redundant and missing error types in the rule base

{ **if** sentence S contains any structure that can be identified by r_i

then if r_i is the redundant error type

then {

if Reliability(r_i) > maximum

then

 maximum = Reliability(r_i);

 rule-pointer = r_i ;

 }

else if (Reliability(r_i) * y) > maximum

then {

 maximum = Reliability(r_i) * y ;

 rule-pointer = r_i ;

 }

}

Tag the erroneous portion of sentence S with the rule identified by the rule-pointer and **return** the corrected sentence;

return sentence S is the correct sentence;

End.

Figure 3. Proposed KNGED algorithm for the detection and correction of syntax errors

The methods for generating identification rules for different types of errors vary, and so does their effectiveness. We applied the various types of identification rules to the TEA1-DRY data set and then analyzed their effectiveness. We found that the identification rules for the selection type of errors had a much higher degree of accuracy compared to the rules for the other types of errors. This is because the identification of errors in the use of units is completely based on the vocabulary, resulting in a relatively lower rate of error. Thus, under the proposed algorithm, once a sentence has been identified as having an error of the selection type, that type of error would be ascribed to the sentence first. On the other hand, results for the wrong order type of error all arise from manually generated rules, hence the relatively lower rate of accuracy. Nevertheless, they are still more accurate than the identification rules for the redundant and missing types of errors. Thus, when a sentence is identified as having the wrong order type of error but not that of selection, it should first be ascribed the former type of error.

The value for sentences that have not been identified under the selection and wrong order types of errors but have been identified under the missing and redundant types is calculated based on the reliability value for each rule as shown in Formula (1). Compared to the rule for the omission error it is easier for the rule for the redundant type to achieve a higher value in terms of reliability. Hence, if a sentence complies with an identification rule for the redundant type of error and another for the omission type, the reliability value of the former must be several times greater than that for the latter (*i.e.*, the y value of the algorithm). It is only in this situation that the identification results for the redundant type of errors are adopted. Otherwise, the sentence should be treated as the omission type of errors.

4. Experimental Results

The formal run data provided by NLPTEA1-CFL (Yu, Lee, & Chang, 2014) was used to evaluate the effectiveness of the proposed method. The data consist of 1,750 sentences. A half of these sentences have no grammatical errors while each of the remainder only contain one grammatical error. The number of sentences with error type redundant, missing, disorder, and selection is 279, 350, 120, and 126, respectively. Three indicators for evaluating the performance of our proposed method are defined as follows:

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP})$$

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN})$$

$$\text{F1} = 2 * \text{Precision} * \text{Recall}/(\text{Precision} + \text{Recall})$$

where TP refers to the number of sentences for which the error type was correctly detected, FP

refers to the number of sentences with no errors that were nevertheless identified as erroneous, and FN refers to the number of sentences with errors that were not detected or detected but ascribed the incorrect error type. Since the assessment facets for recall and precision are different, the F1-measure was used as the overall indicator of assessment effectiveness. In NLPTEA1-CFL, the evaluation is divided into detection level and identification level. In detection level, the proposed method only grouped test sentences into two types: correct or incorrect. In identification level, the proposed method should clearly identifies test sentences to be one of four error types: Redundant, Missing, Disorder, and Selection.

The performance of KNGED based on the three assessment indicators is shown in Table 1. Since the performance of KNGED is affected by the parameter settings, Table 1 also shows the calculation results for KNGED's effectiveness under various parameters settings. When the parameter settings for KNGED-1 were $p = 1$, $k = 2$, $y = 50$, the number of rules generated for the redundant and omission types of errors was 53,834 and 3,781, respectively. When the parameter settings for KNGED-2 were $p = 1$, $k \approx \infty$ (i.e. $FreqInCor(r)=0$), $y = 50$, the numbers of rules generated for the same two types of errors were 10,114 and 145. The parameter settings for KNGED-3 were $p = 1$, $k = 2$, $y = 1$. Because the parameter p and k of KNGED-3 were the same as of for KNGED-1, the numbers of rules generated for the same two types of errors were also 53,834 and 3,781 respectively.

Table 1. Comparison of results for different parameter settings for the previous experiment

Submission		KNGED-1	KNGED-2	KNGED-3
False positive rate		0.9040	0.2686	0.9040
Detection Level	Precision	0.5015	0.5164	0.5015
	Recall	0.9326	0.2880	0.9326
	F1	0.6523	0.3698	0.6523
Identification Level	Precision	0.2600	0.2555	0.2505
	Recall	0.3257	0.0926	0.3097
	F1	0.2892	0.1359	0.2770

In detection level, the F1-measure values of KNGED-1 and KNGED-3 were the highest and far exceeded the effectiveness of KNGED-2. The main reason is because the parameter settings of KNGED-2 resulted in only few rules in the rule base, causing the recall to decrease. It can thus be seen that the setting of parameter values have considerable impact on effectiveness. In addition, the performance of three parameter settings of KNGED do not perform well in identification level. The main reason is the inclusion of many invalid rules in

the rules database. It causes the accuracy to decrease.

A comparison between the effectiveness of manually-generated identification rules and machine-generated rules under KNGED-1 is shown in Table 2. In KNGED-1, the machine-generated rules do not contain the disorder type of errors, whereas the numerical variations between the various types of errors for manually-generated identification rules are large. Thus, we cannot deduce arbitrarily which method was better. However, it can be seen from Table 2 that it is insufficient to only employ manually-generated rules to identify grammatical errors. On the other hand, Table 2 also shows that the machine-generated rules of KNGED-1 are effective even all rules are simple bi-gram or tri-gram patterns.

Table 2. Comparison of effectiveness between manually-generated rules and machine-generated rules under KNGED-1

Rules		Manually-generated	Machine-generated
Detection Level	Precision	0.5217	0.5019
	Recall	0.3978	0.9399
	F1	0.4514	0.6543
Identification Level	Precision	0.1429	0.2697
	Recall	0.0608	0.3445
	F1	0.0853	0.3025

Since the information in NLPTEA1-CFL includes the language proficiency level for each sentence, we tested the effectiveness of KNGED-1 at detecting syntax mistakes by authors at different proficiency levels. The results are shown in Table 3. The language proficiency levels were in line with the grading standards of the Common European Framework of Reference for Languages (CEFR). The A1 and C2 grade represents the lowest and highest level of proficiency. It can be seen that the KNGED-1 for identifying erroneous sentences by writers with poor capabilities were more effective than that with good proficiency. This may be because for the writers with good proficiency, the erroneous structures that they make and the related causes are more complex, such that it was inadequate to use simple rules for identification.

Table 3. KNGED-1 identification results of erroneous sentences produced by writers of different CEFR linguistic proficiency levels

Level of CEFR		A2	B1	B2	C1
Detection Level	Precision	0.5104	0.5005	0.4971	0.5263
	Recall	0.9111	0.9342	0.9399	1.0000
	F1	0.6543	0.6518	0.6503	0.6897
Identification Level	Precision	0.2849	0.2683	0.2162	0.2500
	Recall	0.3481	0.3419	0.2623	0.3000
	F1	0.3133	0.3006	0.2370	0.2727

5. Conclusion and Future Work

We made several discoveries based on the processes and results of this experiment. First, although manually-generated rules are more complex than those generated automatically using formulas, their accuracy rates are not necessarily higher. Through manipulation of parameter settings, automatic generation can actually result in more reliable identification rules. Second, automatic generation leads to many rules that have not been manually proposed. This means that the use of machines to determine identification rules is a feasible method. Integrating these two points of view, if the effectiveness of search rules in programs can be significantly enhanced, then it is actually feasible to have a fully automatic system to identify syntax errors by writers for whom Chinese is a second language.

There are several areas in which the proposed method can be further improved. First, the contents of the CWC were the main basis for establishing the rules. Currently, this corpus is still at the expansion phase. As the contents become increasingly enriched, the effectiveness of the system should improve correspondingly. Second, for automatic machine-generated rules, only the immediately preceding/subsequent words/POS are currently considered for rules to identify the redundant and missing types of errors. If the effectiveness of screening the possible rules can be improved, more precise rules will be generated, thereby further enhancing the system's performance.

Third, the heuristic algorithm that we have proposed is unable to handle the issue of one sentence having multiple errors. In terms of practical application, it is very important to develop an algorithm that is able to identify sentences with multiple syntax errors. Fourth, many selection and word order types of syntax errors are related to context rather than syntactic hierarchy. The proposed method has already included the generation of identification rules for erroneous usage of units, which is context-related. Subsequently, further in-depth analysis can be made for other patterns of errors under this category. This will facilitate the

extraction of methods to generate identification rules for errors that are based on or related to context.

Acknowledgement

This work is supported in part by the Ministry of Science and Technology, Taiwan, R.O.C. under the Grants MOST 103-2511-S-151-001. It is also partially supported by the “Aim for the Top University Project” and “Center of Learning Technology for Chinese” of National Taiwan Normal University (NTNU), sponsored by the Ministry of Education, Taiwan, R.O.C. and the “International Research-Intensive Center of Excellence Program” of NTNU and Ministry of Science and Technology, Taiwan, R.O.C. under Grant MOST 104-2911-I-003-301.

References

- ACTFL Proficiency Guidelines 2012 - Writing. (2012). Retrieved August 25, 2014, from <http://actflproficiencyguidelines2012.org/writing>
- Chang, T. H., Sung, Y. T., & Lee, Y. T. (2012). A Chinese word segmentation and POS tagging system for readability research. In *Proceedings of SCiP 2012*, Minneapolis, MN.
- Chen, K. J., Huang, C. R., Chang, L. P., & Hsu, H. L. (1996). Sinica corpus: Design methodology for balanced corpora. *Language*, 167-176.
- Cheng, S. M., Yu, C. H., & Chen, H. H. (2014). Chinese Word Ordering Errors Detection and Correction for Non-Native Chinese Learners. In *Proceedings of COLING 2014*, 279-289, Dublin, Ireland.
- Chodorow, M., Dickinson, M., Israel, R., & Tetreault, J. R. (2012). Problems in Evaluating Grammatical Error Detection Systems. In *Proceedings of COLING 201*, 611-628, Mumbai, India.
- CKIP, (1993). *Analysis of Syntactic Categories for Chinese*. CKIP Tech. Report#93-05, Sinica, Taipei,
- Connors, R. J., & Lunsford, A. A. (1988). Frequency of formal errors in current college writing, or ma and pa kettle do research. *College Composition and Communication*, 39(4), 395-409.
- Dale, R., Anisimoff, I., & Narroway, G. (2012). HOO 2012: A report on the preposition and determiner errorcorrection shared task. In *Proceedings of the 7th Workshop on Building Educational Applications Using NLP*, 54-62, Montréal, Canada.
- Dale, R., & Kilgarriff, A. (2011). Helping our own: The HOO 2011 Pilot Shared Task. In *Proceedings of the 13th European Workshop on Natural Language Generation*, Nancy, France.
- De Felice, R., & Pulman, S. (2009). Automatic detection of preposition errors in learner writing. *CALICO Journal*, 26(3), 512-528.

- Donahue, S. (2001). Formal errors: Mainstream and ESL students. *Presented at the 2001 Conference of the Two-Year College Association (TYCA)*; cited by Leacock et al. 2010.
- Dulay, H. C., Burt, M. K., & Krashen, S. D. (1982). *Language Two*. New York: Oxford University Press.
- Eeg-Olofsson, J., & Knutsson, O. (2003). Automatic grammar checking for second language learners: the use of prepositions. In *Proceedings of the 14th Nordic Conference in Computational Linguistics*, Reykjavik, Iceland.
- Gamon, M., Leacock, C., Brockett, C., Dolan, W. B., Gao, J. F., Belenko, D., & Klementiev, A. (2009). Using Statistical Techniques and Web Search to Correct ESL Errors. *CALICO Journal*, 26(3), 491-511.
- Hsiung, Y. W., Lee, H. H. & Sung, Y. T. (2014). Examining the ACTFL writing assessment rating scale for L2 Chinese learners, *Journal of Chinese Language Teaching*, 11(4), 105-133.
- James, C. (1998). *Errors in Language Learning and Use: Exploring Error Analysis*. London: Addison Wesley Longman.
- Lee, L. H., Chang, L. P., Lee, K. C., Tseng, Y. H., & Chen, H. H. (2013). Linguistic Rules Based Chinese Error Detection for Second Language Learning. In *Proceedings of ICCE 2013*, 27-29.
- Lee, L. H., Yu, L. C., Lee, K. C., Tseng, Y. H., Chang, L. P., & Chen, H. H. (2014). A Sentence Judgment System for Grammatical Error Detection. In *Proceedings of COLING 2014*, 67-70, Dublin, Ireland.
- Leacock, C., Chodorow, M., Gamon, M., & Tetreault, J. (2010). *Automated Grammatical Error Detection for Language Learners*. Morgan and Claypool Publishers.
- Ng, H. T., Wu, S. M., Wu, Y., Hadiwinoto, C., & Tetreault, J. (2013). The conll-2013 shared task on grammatical error correction. In *Proceedings of the 17th Conference on Computational Natural Language Learning*.
- Nicholls, D. (2003) The Cambridge learner corpus: error coding and analysis for lexicography and ELT. In *Proceedings of the Corpus Linguistics 2003 Conference*, 572-581, Lancaster, UK.
- Tetreault, J., & Chodorow, M. (2008). The ups and downs of preposition error detection in ESL writing. In *Proceedings of the 22nd International Conference on Computational Linguistics*, 865-872, Manchester, UK.
- Wang, Z. (2011). *A Study on the Teaching of Unique Syntactic Pattern in Modern Chinese for Native English-Speaking Students*. Master Thesis. Northeast Normal University.
- Yu, C. H., & Chen, H. H. (2012). Detecting word ordering errors in Chinese sentences for learning Chinese as a foreign language. In *Proceedings of COLING 2012*, 3003-3018, Bombay, India.
- Yu, L. C., Lee, L. H., & Chang, L. P. (2014). Overview of Grammatical Error Diagnosis for Learning Chinese as a Foreign Language. In *Proceedings of ICCE 2014*, 42-47, Nara, Japan.

Automatic Classification of the “De” Word Usage for Chinese as a Foreign Language

Jui-Feng Yeh* and Chan-Kun Yeh*

ABSTRACT

This paper proposed a word usage classification for “De” in Chinese as a secondary language by rule induction algorithm. Learning of Chinese characters and tone adaption are both essential and hard tasks for non-native speakers. The frequent terms, defined in morphosyntactic particle “De” with three characters {的, 得, 地}, is hard to learn for foreign learners due to the similar pronunciation and meaning. This investment illustrates a data-driven algorithm to classify the usages about the morphosyntactic particle “De” in Chinese learning. Rule induction is one of the most important techniques to learn the knowledge from data. Since regularities hidden in data are frequently expressed in terms of rules, rule induction is one of the fundamental tools for natural language processing and obtains a significant improvement in character selection. By the automatic rule induction process, 32 rules are adopted here to classify the character usage in morphosyntactic particle “De.” According to the experimental results, we find the proposed method can provide good enough performance to classify the character usages for morphosyntactic particle “De.”

Keywords: Rule Induction, Natural Language Processing, Secondary Language Learning, Classifier, Word Usage.

1. Introduction

To learn Chinese as a foreign or second language is to study of the Chinese languages by non-native speakers and new learners. Increasing interested peoples in China learning from those outside has led to a corresponding interest in the study of Chinese as their second language, the official languages of mainland China and Taiwan. However, the learning of Chinese both within and outside China is not a recent phenomenon. Westerners started learning different Chinese languages in the 16th century. Within China, Mandarin became the

* Department of Computer Science and Information Engineering, National Chiayi University, Chiayi city, Taiwan, ROC

E-mail: Ralph@mail.ncyu.edu.tw

official language in early 20th century. According to the analysis of Summer Institute for Linguistics (SIL), there are near to seven thousands languages over the world nowadays. Among these languages, the top five languages are Chinese, English, Spanish, Bengali and Indian by their population sizes. As the first and second languages, Chinese occupies 14.8 percents populations to be the most used language. China's growing global influence has prompted a surge of interest in learning Mandarin Chinese as a foreign language (CFL), and this trend is expected to continue. Therefore, the population to learn Chinese as the second language is increasing in the latest decades (Simpson, 2000). Compared to the alphabetic language, Chinese is more complex and hard to understand for non-native speakers due to its several thousand characters and complicated sentence structures. Due the historical evolution of Chinese is deep and far, there are some word usage is susceptible to the corresponding allusions. Therefore, it is hard for the second language learners without the Chinese cultural background to understand, handle and use with skill the Chinese words very well. Actually, there are many whereas many computer-assisted learning tools have been developed for learning English, support for CFL learners is relatively sparse, especially in terms of tools designed to automatically evaluate learners' responses.

Computer technologies are used to assist in language learning, the so-called Computer-Assisted Language Learning (CALL), has been invested in the latest decades. An investigation was proposed to the adoption of information and communication technology (ICT) for teachers of Chinese as a foreign language (CFL) in US universities (Lin *et al.*, 2014). Yang (2011) emphasized an online situated language learning environment, for supporting the students, the teachers, and the teaching assistants (TAs) to communicate synchronously and asynchronously in and after class. Chen and Liu (2008) proposed a web-based synchronized multimedia lecture system based on WSML for the learners to learn Chinese as second language. They also compared the Web-CALL, IWiLL, and BRIX based systems for evaluating the proposed systems in Chinese learning/teaching (Chen & Liu, 2008). A user-centered design approach for learn Chinese as second language was invested for evaluating the web usability in (Huang *et al.*, 2010). Lu *et al.* (2014) suggested the curriculum content design in learning Chinese as a second language.

However, Chinese is rated as one of the most difficult languages to learn for people whose native language is English, together with Arabic, Japanese and Korean. There are many difficulties for foreigners to lean Chinese as their second language mainly caused by the special character set and tones in Chinese. Pronunciation cannot be obtained from its character directly. Although there are three aspects: text shape, pronunciation, meaning within one Chinese character. However, there are differences in pronunciation among the similar characters. Therefore, it is hard for the foreign learners to spell the correct Chinese words. For preventing the word segmentation error confusing the word boundaries, Bai *et al.* (2013) used

“De” Word Usage for Chinese as a Foreign Language

the inter-word spacing effects on the acquisition of new vocabulary for readers. One character with different pronunciations and meanings is hard to understand for non-native learners. Compared to other languages, the information of the Chinese character is overloaded. Besides, the number of the Chinese character is too large for a novice especially for the character with server usages. Tone is not easy to control in Chinese characters. The four tones are hard to enunciate for the non-native learners with a toneless source language. For example, the pitch trajectories for the secondary and third tones are one of the main obstacles for the learners. Accented pronunciation confuses the learners to obtain the standard. The pronunciation in Chinese is usually influenced by the speakers' own dialect, since the speaker has learned the dialects before they use the Mandarin. The usages of mandarin usually are affected by the dialects significantly such as Wu, Hokenese, Haka, and Cantonese. The complex structure of the Chinese character makes the hinder for nonnative learners. Reading and writing are main learning activities and they are cross validation for assessment of the achievements to use the Chinese characters. However, the complex structure and too many strokes make it more difficult to understand the reading and writing for learners. The flexible grammar rules in Chinese are not easy to learn for nonnative speakers. Confucius has described the Chinese as “a language without solid grammar (文無定法)” since two thousand years ago. The flexibility in syntax makes Chinese to be one of the most various languages. The rich rhetoric in Chinese make it is interesting and hard to understand the grammatical rules. The influence by ancient writings, the word usage is more complex in Chinese. That is to say, the literary language used in ancient China and preserved today for formal occasions or conspicuous display. Without the culture background, the foreigner learners are not able to obtain the meaning and pronunciation about word preciously.

The part of a word to which inflectional endings are attached, they are usually seen in alphabetical languages. Stem provides a good extension for word usage for language learners. However, the stems are hard to be obtained for non-native speaker, since the Chinese word with complex structure. The lexicon is hard to use for new learners. Actually, the design of Chinese lexicon aims at the user who is experienced in Chinese usage especially for the populations in home country. It is not friend for new learners. This makes it hard to study Chinese by oneself. For removing the barrier of learning Chinese as second language, more efforts are invested in Chinese character learning. Learning Chinese, which consists of more than ten thousands of characters composed of hundreds of basic writing units, presents such a challenge of orthographic learning for non-native speakers at the beginning stages of learning. A classroom was designed to extend previous research on how to support orthographic learning in (Chang *et al.*, 2014). Chuang and Ku (2011) invested the effect of computer-based multimedia instruction with Chinese character recognition for foreign learners. Chen *et al.*, (2013) proposed an approach for investigating the a radical-derived Chinese character

teaching strategy on enhancing Chinese as a Foreign Language (CFL) learners' Chinese orthographic awareness based on statistical data from the Chinese Orthography Database Explorer established and used as an auxiliary teaching tool. Hsiao *et al.* (2013) designed and developed a Chinese character handwriting diagnosis and remedial instruction (CHDRI) system to improve the CFL learners' ability in Chinese character writing. The CFL learners were given two tests based on the CHDRI system. One test focused on Chinese character handwriting to diagnose the CFL learners' errors in the stroke order and their knowledge of Chinese characters, while the other test focused on the spatial structure of Chinese characters (Hsiao *et al.*, 2013). Looi *et al.* (2009) Explored interactional moves in a CSCL environment for Chinese language learning. Chang *et al.* (2012) presented approach for error diagnosis of Chinese sentences for Chinese as second language (CSL) learners. A penalized probabilistic First-Order Inductive Learning (pFOIL) algorithm is presented for error diagnosis of Chinese sentences. The pFOIL algorithm composed with three parts: inductive logic programming (ILP), First-Order Inductive Learning (FOIL), and a penalized log-likelihood function for error diagnosis (Chang *et al.*, 2012). Chinese is a tonal language; tone and pronunciation acquisition also plays an essential role for CSL learners. There are some research efforts were made for listening and speaking diagnosis (Hao, 2012; Chu *et al.*, 2014; Chun *et al.*, 2015; Hsiao *et al.*, 2015).

Since the learning for Chinese is not easy for non-native speakers. This drives us to the question what is the one to help the foreign learners. Indeed, the characters those are frequently used and mistake for each other usually confuse the foreign Chinese learners. The second language learners for Chinese usually are in the state of confusion about the usage of "De" (Jiang *et al.*, 2012). Shi and Li (2002) analyzed the causal relationship between the establishment of classifier system and the grammatical issues of the particle "De". Yip and Rimmington (2004) described that "De" is required to be present in the relative clause as modifier contexts for Chinese as second language (CSL) learners. Waltraud (2012) analyzed the in subordinate subordinator "De" in Mandarin Chinese. Paul (2012) compared the difference of "De" in Chinese and French. Li (2012) also compared the usage between "De" in Chinese and "E" in Taiwanese. This paper invested an automatic rule induction algorithm for classification of the usages of the morphosyntactic particle "De." The confusing set about the morphosyntactic particle "De" is defined as the character set {的, 得, 地} in Chinese. Herein, the automatically classification about the morphosyntactic particle "De" is further defined as the process to decide which character is correct for using in Chinese. That is to say, we want to help the non-native learner to know which one is correct in the morphosyntactic particle "De" in Chinese.

This paper is organized as follows. Section 2 describes the rule induction algorithm used for classify the usage of morphosyntactic particle "De" in detail. In Section 3, we analyze the

performance in experimental results of the proposed methods. Finally, Section 4 will illustrate the findings and draw the conclusion of this paper.

2. Rule Induction for Morphosyntactic Particle “De”

Using the basic ideas of rough set theory, learning from examples module version 2 (LEM 2) is adopted as the rule induction algorithm based on corpus with semantic tagging. As we known, LEM 2 is one of rule induction methods in LERS data mining system, the flow chart is illustrated in Figure 1.

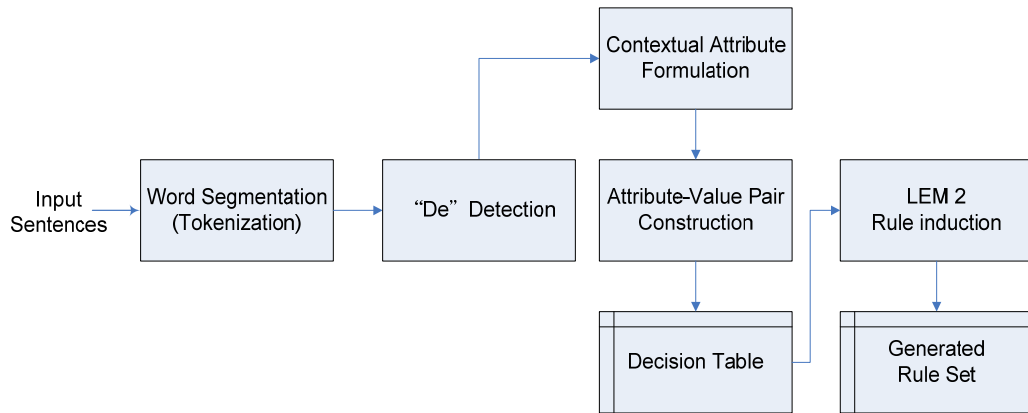


Figure 1. The LEM 2-based rule induction for the morphosyntactic particle “De.”

This paper adopted the LEM 2 algorithm to natural language processing especially for Chinese information processing. For each input Chinese sentence, word segmentation is applied to obtain the word level tokens with part-of-speech (POS) tagging. The detection process for “De” is further used to select the sentence with “De.” The sentences without “De” are dropped in the post-processing here. For extracting the linguistic feature to decide which morphosyntactic particle is used in the sentence, the contextual attributes are defines according to word and part-of-speech based n-grams. The contextual attributes accompanied with morphosyntactic particle {的, 得, 地} to constructing the attribute-value pairs. All the attribute-value pairs gathered in training data are fed into LEM 2 rule induction algorithm to generate the rule set. Therefore, the rule set can be further used to decide the usage of the morphosyntactic particle {的, 得, 地}. Herein, the proposed method is divided into two parts, decision table construction and LEM 2 rule induction algorithm, are described dentally in Section 2.1 and 2.2 respectively.

2.1 Decision Table Construction

Since the decision table is defined as a form for blocks of attribute-value pairs, the attribution plays an essential role in rule induction using LEM 2 algorithm. However, the sentence in natural language is not structural and fitting to the format of attribute. It is noteworthy how to transform the natural language into the attribute. That is to say, proposition extracted from sentence is one of the important issues for attribute. Herein, the contextual information surrounding the morphosyntactic particle “De” is used to form the attributes as shown in Figure 2.

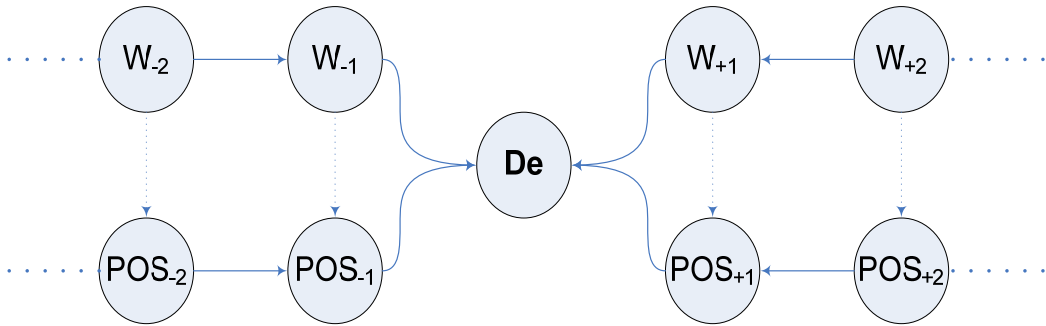


Figure 2. The contextual attribute formulation using the word with POS information surrounding the target word “De.”

Each sentence representing one case and the independent variables are called attributes. As shown in Figure 2, the surrounding words with part-of-speech will combined with their relative position for particle “De” will combine into considering to form the attributes. The values are defined as one of the single-character words {的, 得, 地} in particle set. Each attribute-value pair represents one sample of knowledge about a decision table or a property of cases. These attribute-value pairs and the corresponding blocks serve as a basis for rule induction. Similar to N-gram models, the utility of the proposed contextual features is closely linked with the observation window size. As we know, the longer word sequence can provide more information for predicting the next word in N-gram models. This phenomenon leads us to find the near optimal window size for the “De” classifier. However, we have observed the empirical results of the larger windows size. Here, the relative positions from -2 to +2 are included in the windows for obtaining the contextual attribute, because the performance is near to those by the larger windows size. This condition not to conform to our expectation and the reason should be the limitation of the training corpus. For the example shown in Figure 3, the related information in decision table is illustrated in Table 1. The sentence containing the word sequence “欣賞(enjoy) 美麗(beautiful) 的(De) 一幅(a) 畫(picture)” is illustrated as the case 1 shown in Figure 3. Basically, each case is obtained from one sentence. Actually, the number of cases is dependent on the number of the particle “De” in the sentence. An

“De” Word Usage for Chinese as a Foreign Language

example “特別(special) 的(De1) 愛(Love) 給(give/for) 特別(special) 的(De2) 您(you)” with two particles, the cases 2 and 3 is obtained from the same sentence. The cases 4 and 5 in Table 1 show the examples for “得(De)” and “地(De)” separately.

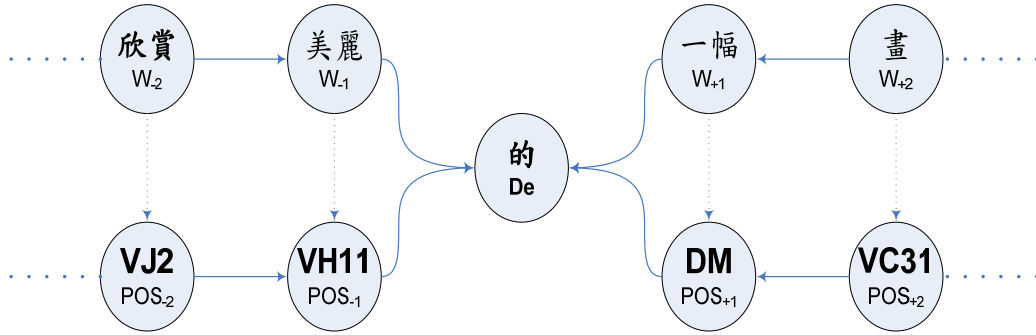


Figure 3. The contextual attribute formulation for the sentence containing “欣賞(enjoy) 美麗(beautiful) 的(De) 一幅(a) 畫(picture).”

Table 1. A decision table for the decision of “De.”

case	Attributes								Decision
	W ₋₂	POS ₋₂	W ₋₁	POS ₋₁	W ₊₁	POS ₊₁	W ₊₂	POS ₊₂	De
1	欣賞	VJ2	美麗	VH11	一幅	DM	畫	VC31	的
2	-	-	特別	VH11	愛	Nad	給	VD1	的
3	給	VD1	特別	VH11	您	Nhac	-	-	的
4	-	-	快樂	VH21	不得了	VH11	-	-	得
5	暗自	Dh	悲傷	VH21	低聲	Dh	哭泣	VA4	地
.				.					.
.				.					.
.				.					.

2.2 LEM 2 Rule Induction Algorithm

Rule induction is important to find relationships between blocks defined by condition attributes and the blocks defined by the decision attributes. In Chinese, particles usually connect adverb, adjective, verb and noun from the observations. U and A denote the set of all cases and the set of all attributes in decision table. Independent variables are treated as attributes. The target variable depended on attributes is called as decision. A function $f(\cdot)$ is defined for mapping the direct product of U and A into the set of all values. These

terminologies are defined in rough set theory and LEM 2. The fundamental idea of rough set theory is aimed at using the blocks in decision table to explain the rule induction. Q is one of the nonempty subsets of A . the indiscernibility relation $IND(Q)$ is defined as follows.

$$(x, y) \in IND(Q) \Leftrightarrow f(x, a) = f(y, a) \text{ for } \forall a \in Q, \quad (1)$$

Since the indiscernibility relation $IND(Q)$ is an equivalence relation. The elementary set of Q , denoted by $[x]_Q$, is defined as the equivalence classes of $IND(Q)$. $IND(Q)$ can be used to obtain the idea of blocks of attribute-value pairs. The intersections of blocks are shown in equation (2).

$$[x]_Q = \cap \{(a, v) | a \in Q, f(x, a) = v\}. \quad (2)$$

This investment adopted the rule induction algorithm to explore the search space of attribute-value pairs. Lower approximation for concept is defined as conditional probability is one. The probability of the upper approximation is greater than zero. According to lower and upper approximations, the concept is further divided into three areas: positive region, boundary region, and negative region. LEM 2 explores the search space of attribute-value pairs and finds a local covering and then converts it into a rule set. The algorithm is shown in (Grzymala-Busse, 2005).

3. EXPERIMENTAL RESULTS

For evaluating the proposed method, the LEM 2 algorithm is adopted for classifying the usage of particle “De” for the learners in Chinese as the second language. We first induce the rule set using the word and its corresponding part-of-speech information by the general training corpus. Furthermore, the evaluation set using the test corpus gathered from the non-native speakers for Chinese. The confuse matrix, precision rate, recall rate and F1 measures are applied for assessment of the proposed method. Here, we illustrate data preparation, evaluation metrics, experimental results and discussion in the following sections in detail.

Table 2. The rule set induced by the proposed method.

rule	Attributes								Decision
	W ₋₂	POS ₋₂	W ₋₁	POS ₋₁	W ₊₁	POS ₊₁	W ₊₂	POS ₊₂	De
1				VC		P	到		的
2				VC		VCL	到		的
3				VK		VH	不得了		的
4				VH		VH	不得了		的
5				V-		Dfa			的

“De” Word Usage for Chinese as a Foreign Language

6		V-	V-	(V_2)'	的
7		VH	Nv	努力	的
8		Dfa	V-		的
9		D	V-		的
10		VK	VJ		得
11		VK	VK		得
12		D	VJ		得
13		D	VK		得
14			Na		得
15			Neu		得
16			Nv		得
17			Nc		得
18		VH	VA		得
19		VC	VA		得
20		VA	VA		得
21	(沒)'		VJ		得
22		Nc-(好)	VC		得
23	(漂亮)'		VE		得
24		VH	D	就	得
25		VH	VL		地
26		V-	Dfa		地
27		VC	Nh		地
28		VC	Nv		地
29		VHC	Nh		地
30		VHC	Nv		地
31			Na		地
32			Neu		地

3.1 Data Preparation and Evaluation Metrics

Since the goal of this paper aims at the usage classification for morphosyntactic particle “De,” two corpuses, CYCCDC (Yeh *et al.*, 2014) and FinalTest_SubTask2 in shared-task on Chinese Grammatical Error Diagnosis (CGED) (Yu *et al.*, 2014), are employed for training corpus and test corpus respectively. CYCCDC is a conversational dialogue corpus form daily life. The recorded speech is collected and annotated as text transcript. Considering of the learners’ usage in real life and learning about the capabilities in listening, speaking, reading and writing, CYCCDC is used for building the rule set. The test file FinalTest_SubTask2 is provided for evaluating the Chinese grammatical error diagnosis. The sentence is gathered from the learner for Chinese as the second language. However, the number of the sentence containing error usage about the morphosyntactic particle “De” is not large enough. The character in the morphosyntactic particle “De” character set {的,得,地} is randomly re-assigned as the character from the same character set to form our test corpus.

The goals of this approach are to detect whether an input sentence contained error usage of the morphosyntactic particle “De” and to identify the correct character/word. Table 3 shows a contingency table of the related hypothesis.

Table 3. Contingency table for the usage classification of the morphosyntactic particle “De.”

Hypothesis		Condition		
		Positive	Negative	Total
Outcome	Positive	True Positive TP	False Positive FP	P
	Negative	False Negative FN	True Negative TN	N
	Total	TP+FN	FP+TN	P+N

There are three metrics were used to assessing the proposed method: precision rate, recall rate and F1 measure, they are formulated as equations (3), (4) and (5) separately.

$$Precision\ rate = \frac{TP}{P}, \quad (3)$$

$$Recall\ rate = \frac{TP}{TP + FN}, \quad (4)$$

$$F1\ measure = \frac{2 \times Precision\ rate \times Recall\ rate}{Precision\ rate + Recall\ rate}. \quad (5)$$

3.2 Experimental Results and Discussion

Tables 4 and 5 show the evaluation results of confusion matrices of the usage classification about the morphosyntactic particle “De” in frequency count and percentage separately. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. This part aims at finding the confusion status among the words {的, 得, 地} by the rule set induced from LEM 2 algorithm. From the observation about confusion matrices, the correction rate of “的” is the highest compared to the “得” and “地.” Due to the occupation ratio of “的” is higher than the other two particles in training corpus and test set, the miss rate about “的” is less than 10 percentages. This is excellent output for practice use. However, there are many false alarm errors about “的” cause the accuracies of “得” and “地” is not good enough. Many induction rules resulted from the training cases in corpus focus on the “的”. This condition makes the false alarm and reduce the precision of the other particles “得” and “地.”

Table 4. Count-based Confusion matrix for of the morphosyntactic particle “De.”

	的	得	地
的	1560	69	49
得	67	46	7
地	45	6	36

Table 5. Correction percentage confusion matrix for of the morphosyntactic particle “De.”

	的	得	地
的	0.929781	0.041120	0.029201
得	0.558333	0.383333	0.058333
地	0.517241	0.068965	0.413793

Tables 6 illustrates the performance measure about morphosyntactic particle “De” including the metrics precision rate, recall rate and F1 measure. From this result, we can find that the performance of “的” achieve the best performance compared to those of “得” and “地.” This is affected by the occupation ratio of particle significantly. Besides, According to the observation of the outcome data, we find that the characters ‘的,’ ‘得’ and ‘地’ maybe part of the word with multiple characters such as “目的,” “得意” and “土地”. This condition cause the performance dramatically reduced. These errors usually come from the wrong word segmentation and the characters.

Table 6. The performance measure of the proposed method using precision rate, recall rate and F1 measure.

	<i>Precision rate</i>	<i>Recall rate</i>	<i>F1</i>
的	0.9459 (70/74)	0.5932 (70/118)	0.7292
得	0.4242 (28/66)	0.3590 (28/78)	0.3889
地	0.5686 (29/51)	0.4915 (29/59)	0.5273

4. CONCLUSIONS

In this paper, we focus on rule induction on the usage of morphosyntactic particle “De” for the Chinese as the second language learners. The attributes that were formed from the surrounding words and the corresponding part-of-speech are adopted for attribute-value pairs. The training data is fed into the rule induction process. Here, LEM 2 algorithm is adopted here for deriving the rule set to classify {的,得,地} in this investment. The main contribution of this paper aims at the attribute-value pair formulation from the sentence in natural language. Considering of the contextual information, the position and part-of-speech of the surrounding words are used to form the independent variables. More than thirty rules are induced by LEM 2 algorithm. According to the observation about experimental results, we found the proposed method is workable and its performance is good enough in practice. We illustrate the confusion matrix and performance measure based on precision and recall rates. By this approach, the Chinese as second language learners can obtains the desired help in the usage of morphosyntactic particle “De”.

ACKNOWLEDGEMENTS

The authors thank the anonymous reviewers for their helpful suggestions. This research was funded by the National Science Council, Taiwan (R.O.C.) (Contract No. NSC 100-2622-E-415-001-CC3). Funding was also received from the Ministry of Education, Taiwan (R.O.C.).

REFERENCES

- Bai, X., Liang, F., Blythe, H. I., Zang, C., Yan, G., & Liversedge, S. P. (2013). Interword spacing effects on the acquisition of new vocabulary for readers of Chinese as a second language. *Journal of Research in Reading*, 36(S1), S4-S17.
- Chang, L. Y., Xu, Y., Perfetti, C. A., Zhang, J., & Chen, H. C. (2014). Supporting Orthographic Learning at the Beginning Stage of Learning to Read Chinese as a Second Language. *International Journal of Disability, Development and Education*, 61(3), 288-305.

“De” Word Usage for Chinese as a Foreign Language

- Chang, R. Y., Wu, C. H., & Prasetyo, P. K. (2012). Error diagnosis of Chinese sentences using inductive learning algorithm and decomposition-based testing mechanism. *ACM Transactions on Asian Language Information Processing (TALIP)*, 11(1), 3.
- Chen, H. Y., & Liu, K. Y. (2008). Web-based synchronized multimedia lecture system design for teaching/learning Chinese as second language. *Computers & Education*, 50(3), 693-702.
- Chen, H. C., Hsu, C. C., Chang, L. Y., Lin, Y. C., Chang, K. E., & Sung, Y. T. (2013). Using a radical-derived character e-learning platform to increase learner knowledge of Chinese characters. *Language Learning & Technology*, 17(1), 89-106.
- Chu, H. T., Tsai, W. S., & Lee, S. Y. (2014). Design of a Cloud Service for Learning Chinese Pronunciation. *Journal of Electronic Science and Technology*, 1, 011.
- Chuang, H. Y., & Ku, H. Y. (2011). The effect of computer-based multimedia instruction with Chinese character recognition. *Educational Media International*, 48(1), 27-41.
- Chun, D. M., Jiang, Y., Meyr, J., & Yang, R. (2015). Acquisition of L2 Mandarin Chinese tones with learner-created tone visualizations. *Journal of Second Language Pronunciation*, 1(1), 86-114.
- Grzymala-Busse, J. W. (2005). Rule induction. *Data Mining and Knowledge Discovery Handbook*, 277-294. Springer US.
- Hao, Y. C. (2012). Second language acquisition of Mandarin Chinese tones by tonal and non-tonal language speakers. *Journal of Phonetics*, 40(2), 269-279.
- Hsiao, H. S., Chang, C. S., Chen, C. J., Wu, C. H., & Lin, C. Y. (2013). The influence of Chinese character handwriting diagnosis and remedial instruction system on learners of Chinese as a foreign language. *Computer Assisted Language Learning*, 1-19.
- Hsiao, H. S., Chang, C. S., Lin, C. Y., Chen, B., Wu, C. H., & Lin, C. Y. (2015). The development and evaluation of listening and speaking diagnosis and remedial teaching system. *British Journal of Educational Technology*.
- Huang, C. K., Hsin, C. O., & Chiu, C. H. (2010). Evaluating CSL/CFL website usability: A user-centered design approach. *Journal of Educational Multimedia and Hypermedia*, 19(2), 177-210.
- Jiang, Y., Wang, T., Lin, T., Wang, F., Cheng, W., Liu, X, Wang C. & Zhang, W. (2012, June). A rule based Chinese spelling and grammar detection system utility. *IEEE International Conference on System Science and Engineering (ICSSE) 2012*, 437-440).
- Li, Y. H. A. (2012). de in Mandarin ↔ e in Taiwanese. *Studies in Chinese Linguistics*, 33(1), 17-40.
- Lin, C. Y., Huang, C. K., & Chen, C. H. (2014). Barriers to the adoption of ICT in teaching Chinese as a foreign language in US universities. *ReCALL*, 26(01), 100-116.
- Looi, C. K., Chen, W., & Wen, Y. (2009). Exploring interactional moves in a CSCL environment for Chinese language learning. In *Proceedings of the 9th international conference on Computer supported collaborative learning*, Vol. 1, 350-359.

- Lu, C. C., Hsieh, G., Chung, Y., Liu, C., Shih, C. W., & Hsu, W. L. (2014). Suggestions on curriculum content design in learning Chinese as a second language. *International Conference on Information Society (i-Society) 2014*, 339-340.
- Paul, W. (2012). Why Chinese *de* is not like French *de*: A critical analysis of the predicational approach to nominal modification. *Studies in Chinese Linguistics*, 33(3), 182-210.
- Shi, Y., & Li, C. N. (2002). The establishment of the classifier system and the grammaticalization of the morphosyntactic particle *de* in Chinese. *Language sciences*, 24(1), 1-15.
- Simpson, A. 2000. On the status of modifying *DE* and the structure of the Chinese *DP*. In S.W. Tang and L.L. Chen-Sheng (eds.), *On the Formal Way to Chinese Languages*. Stanford: CSLI, 74-101.
- Yang, Y. F. (2011). Engaging students in an online situated language learning environment. *Computer Assisted Language Learning*, 24(2), 181-198.
- Yeh, J. F., Lu, Y. Y., and Tan Y. S. (2014). CYCCDC: A Chiayi Chinese Conversation Dialogue Corpus. In *Proceedings of the 22nd International Conference on Computers in Education, workshop proceedings*, 7-12.
- Yip, P.-C., & Rimmington, D. (2004). *Chinese: A comprehensive grammar*. New York: Routledge.
- Yu, L.-C., Lee, L.-H., and Chang, L.-P. (2014). Overview of grammatical error diagnosis for learning Chinese as a foreign language. In *Proceedings of the 1st Workshop on Natural Language Processing Techniques for Educational Applications*, 42-47.
- Waltraud, P. A. U. L. (2012). *The in subordinate subordinator de in Mandarin Chinese: Second take*. The Attributive Particle in Chinese (tentative title) [Frontier in Linguistics Series](ed. Sze-Wing Tang), Beijing:Peking University Pres.

以「華語學習者語料庫」為本的「了」字句偏誤分析¹

The Error Analysis of “Le” Based on “Chinese Learner Written Corpus”

董子昀*、陳浩然*⁺、楊惠媚*

Tzu-Yun Tung, Howard Hao-Jan Chen and Hui-Mei Yang

摘要

「了」為中文常見的時貌標記和句尾虛詞，但其表動作完成的語義，可能使華語學習者將其泛化為過去時標記，而成為學習難點。本文以「臺師大華語學習者語料庫」為本，分析初級 A2 和中級 B1 英語母語者學習「了」字句時的使用情形和偏誤類型，有以下幾點發現：（1）初級 A2 和中級 B1 學習者都較難掌握作為時貌標記的「了₁」，而較容易掌握作為句尾虛詞的「了₂」。（2）不論初級 A2 或中級 B1 都有「了₁」過度泛化的情況。（3）此二級學習者在「了₂」及「了₁₊₂」的使用上，皆多為冗餘偏誤。對於「了」字句的教學，本文與鄧守信（1999）所主張的建議一致，「了」字句之教學上可由簡入繁，先介紹「了₂」和相關句型，再介紹「了₁」和相關句型。同時，本文亦歸納出「了」字句教學時應釐清的重要觀念和建議的教學順序，期能提供華語教材編寫的參考方向。

關鍵詞：「了」字句、偏誤分析、華語學習者語料庫、華語教學

¹本研究感謝教育部「邁向頂尖大學計畫」與科技部「跨國頂尖研究中心計畫」(MOST 104-2911-I-003-301),以及國立臺灣師範大學「華語文與科技研究中心」支持。

* 10610 台北市大安區和平東路一段 162 號 國立台灣師範大學 英語學系

Department of English, National Taiwan Normal University, No. 162, Sec. 1, He-ping E. Rd., Taipei, Taiwan R.O.C. 10610

E-mail: {60021031L, hjchen}@ntnu.edu.tw; huimei.yang2009@gmail.com

⁺本研究通訊作者

The author for correspondence is Howard Hao-Jan Chen.

Abstract

The function word “*le*” in Chinese serves as both a sentence final particle (*le*₁) and an aspect marker (*le*₂). As an aspect marker indicating the completion of action, “*le*” has been observed to be frequently misused by learners of Chinese, among which the overgeneralization of “*le*” a past-tense marker is the most glaring. Based on “NTNU Chinese learners’ written corpus”, we analyzed the usage and the error types of “*le*” made by English-speaking learners at the beginning (A2) and the intermediate level (B1). The results show that both A2 and B1 learners acquire *le*₂ before *le*₁, and in terms of error analyses, *le*₁ is the most commonly spotted error type and there is a large number of redundancy of the use of *le*₂ and *le*(1+2). Therefore, in a similar vein with Teng (1999), this current study sides with the proposition that the use of *le*₂ along with its associated sentence patterns should be taught prior to that of *le*₁. Pedagogical implication as well as the suggestion of the editing of CFL textbooks are also provided.

Keywords: “*le*,” Error Analysis, Chinese Learner Written Corpus, Chinese Teaching

1. 引言

「了」為中文常見的時貌標記和句尾虛詞，廣泛使用於中文語句中，如：「我吃了飯再走」和「他回家了」。「了」可用以標示動作完成，可能導致華語學習者將「了」視為過去時標記。但事實上，「了」不僅能用於表過去時間的句子，也能用於表現在和未來事件的句子；另一方面，在過去時的句中，除了「了」可出現外，中文更常以時間狀語來表達過去時。對母語為非分析型語言的華語學習者來說，這種非一對一的對應關係也許為一大學習難點。

以往針對「了」字句的研究中，主要以實驗性質收集語料，而研究所分析的語料量較屬小量，或較具針對性質而非學習者在自然語境下所產出的語言。因此，晚近學習者語料庫的建構，使之成為外語習得研究有力的工具之一。因其收集了許多外語學習者產出的語句，能讓研究者以大量的真實語料為本，進行有系統的量化和質化分析，進而了解外語習得的歷程和影響因素。學習者語料庫可是歷時性，用以記錄同一個學習者在不同學習階段的語言使用；也可是共時性，用以呈現不同學習者在同一階段的語言使用。

本文使用「華語學習者語料庫」，為一共時性的學習者語料庫，含納了 A2、B1、B2 和 C1 四級不同程度華語學習者參加「華語文能力測驗 TOCFL」² 寫作考試作文。學習者的程度由入門至精通劃分為 A2、B1、B2、C1 四級，而本文選擇分析的語料為 A2 和 B1 的考生作文，希望能了解初級和中級華語學習者在學習「了」字句時，是否有相

² 「華語文能力測驗 TOCFL」由臺灣國家華語測驗推動工作委員會針對母語非華語人士研發而成，於 2003 年 12 月正式對外開辦考試。

同或相異的使用情形和偏誤類型。因語料庫中學習者的母語背景繁多，本文因篇幅限制且為了更針對性地觀察某一語系背景學習者偏誤情形，本文以學習者母語為英語背景者為本次研究對象，希望能將偏誤系統性分類，嘗試找出造成英語為母語學習者偏誤句的影響因素，以提供對華語教材編纂和華語教學之建議。

2. 文獻探討

2.1 「了」之句法與語義功能研究

現代漢語的「了」依其結構位置來看，可以有以下三種情形：(1) 出現於動詞後；(2) 出現於句末；(3) 同時出現於動詞後和句末。根據呂叔湘（1980）的分類，可依其位置及功能區分為「了₁」、「了₂」和「了₁₊₂」三類。「了₁」用於句中動詞後，表示動作完成；「了₂」用於句末，表示情況有變化或即將出現變化；而「了₁₊₂」用於句末動詞後，表示動作完成且情況有改變。此外，呂叔湘（1980）也依「了」在句法上的搭配分成六大類句型：(1) 動+了₁+賓；(2) 動+賓+了₂；(3) 動+了₁+賓+了₁；(4) 動+了₁/了₂/了₁₊₂；(5) 形+了₂；(6) 名詞/數量詞+了₂。

金立鑫(1998)認為「了₁」是表示完成貌，亦可兼表「過去近時」的意義，而「了₂」則是事件實現後的狀態延續到某一參照時間的混合標記，表示「現在」的意義。而劉勛寧(1988)則認為應該將「了₁」動詞當作「實現」的標記，它的語法意義是表明動詞、形容詞和其他謂詞形式的詞義所指處於事實的狀態下。史有為(2002)所提出的看法和劉勛寧(1988)相似，其認為「了₁」實際上說明的是針對具體事例的整個動作行為的達成。這樣的說法側重於過程達成後的延續狀態，與呂叔湘(1980)和金立鑫(1998)認為「了₁」表示「完成」的看法有所不同。

而從篇章功能的角度來看，「了₁」有「前後排序」和「高峰標記」兩大篇章功能。（Chang, 1986；Chu, 1999；屈承熹, 2003）當「動作完成」的核心語義虛化，「了₁」可用以標記兩事件發生的先後順序。其二，若一事件包含數個局部動作，要表示整個事件完成，「了₁」只標示在最重要的動詞後，通常為最後一個動詞。除了標示整個事件完成，「了₁」更有組織篇章的功能，能結合鬆散的子句，標示出信息高峰。若將所有動詞均加上「了₁」，語法雖無誤，卻會造成結構鬆散和信息焦點分散。另一方面，當「了₂」「狀態改變」的核心語義虛化後，將有「敘述段落」的篇章功能，標示敘述告一段落。（屈承熹, 1999）若一語段中的每句話均加上「了₂」，各表「敘述段落」，句子間將缺乏連貫性，整個語段也會失去整體性。

此外，「了₁」和「了₂」也有句法分布上的限制。首先，根據屈承熹（1999），「了₁」有兩大句法限制，其一為「了₁」常出現在主句，而不出現在從句，即便該事件或動作已完成。陳俊光（2008）進一步指出，此句法限制實源自於篇章功能。因「了₁」的篇章功能為「高峰標記」，所標記的信息焦點多由主句表達，因此「了₁」往往出現在主句而非從句。「了₁」的第二個句法限制為若說話類或思想類動詞（如：說、告訴、想）後，接有所說或所想的內容時，即便動作已完成，也不能加上「了₁」。陳俊光（2008）

仍認為此句法限制與篇章功能相關。因「了₁」另一篇章功能為「前後排序」，用以標記兩動作或事件發生的先後順序；然而在「告訴他人什麼內容」或「想起什麼內容」中，「動詞」和「內容」的出現並沒有清楚的時間順序區別，而是兩者同時出現，因此與「了₁」前後排序的篇章功能相衝突。與「了₁」相似，「了₂」也常用於主句，而不用於從句中。（湯廷池, 1999）至於原因為何，陳俊光（2008）也歸因於「了₂」敘述段落的篇章功能。因「了₂」可用於表示敘述告一段落，後面不應再加上其他句子，而無法使用於主語子句或副詞句等從句中。

除了上述的觀點以外，沈家煊(1995)、陸方喆(2014)和管韻(2010)都提及了「了」的「界現性」問題。Li & Thompson (1981)指出有界事件(bounded event)是「了」出現的條件。陸方喆(2014)則認為「了」核心語義是「+出現/實現」以及「+有界」，而詞尾「了₁」和句尾的「了₂」只是不同句法環境下的語義變體，前者表示動作的出現，後者表示事件或狀態的出現。管韻(2010)也認為「了₁」表示說話者的視點體(viewpoint)在事件的外部，完整地描述一個具有起迄點的事件，具有「+有界性」和「+動態性」的語義。本文由主要的前人文獻中，針對「了」句法及語義功能之定義整理呈現於下表 1：

表1. 「了」之句法與語義功能對照表

文獻	了 ₁	了 ₂	了 ₁₊₂
呂叔湘 (1980)	出現於動詞後，表示動作完成。	出現於句末，表示情況有變化或即將出現變化。	同時出現於動詞後和句末，表示動作完成且情況有改。
金立鑫 (1998)	表示完成貌，亦可兼表「過去近時」的意義。	表示事件實現後的狀態延續到某一參照時間的混合標記表示「現在」的意義。	
屈承熹 (1999)； 陳俊光 (2008)	具有表示「高峰標記」的篇章功能，而所標記的信息焦點多由主句表達，因此往往出現在主句而非從句。	具有「敘述段落」的篇章功能，可用於表示敘述告一段落，後面不應再加上其他句子，往往出現在主句而非從句。	
陸方喆 (2014)	有界事件，表示動作的出現。	有界事件，表示事件或狀態的出現。	
管韻 (2010)	加接於動詞後，其後名詞需為有定成份。表示動作的完成，具有有界性和動態性的語義。	位於賓語後，其前詞語較無限制，表示當前情況的相關性。	

2.2 「了」字句之習得與偏誤分析

關於華語學習者「了」字句的習得，鄧守信（1999）曾使用“Chinese interlanguage corpus”，分析了9名英語母語者。在919筆語料中，正確使用率達82.7%，鄧守信將此高正確率歸因於學生在中文系國家學習，且每週至少接受20小時的課室教學。在初步分析中、高級學習者的語料後，其研究得出相似的正确使用率，因而推斷「了」字句的習得很快就進入「高原期」，進步有限，且要持續一段時間後才有可能達到巔峰。而初級學習者連續9個月的語料分析顯示，偏誤率逐月下降。鄧守信指出學習者在初期可能過度使用「了」字句，而後才漸漸習得「了」字句的非必要性。

根據呂叔湘（1980）的分類，鄧守信（1999）指出學習者「動+了₁+賓」類的偏誤最多，並推斷由於此類「了₁」表示動作完成，學習者可能直接將之與英文過去簡單式的時態標記相提並論，因而產生病句。另外，「動+了₁+賓+了₂」類的使用率和錯誤率都偏低，鄧守信推斷可能由於此類結構的複雜度更甚，可表示英文過去簡單式或現在完成式的概念，造成學生不常使用。而「動+賓+了₂」、「形+了₂」、「名詞／數量詞+了₂」類的偏誤率也皆偏低，顯示學生較容易掌握標示情況改變語義的「了₂」。

王媚和張艷榮（2007）則分析了俄羅斯留學生「了」字句的使用偏誤，從留學生平時作文中選取有代表性的偏誤，分類比較後也發現學生將「了」泛化為過去式的標誌。因此若母語使用過去時形式，常會在相對應的漢語動詞後加上「了」而發生誤用。而教師在教學中應該說明漢語的時體意義，有時是利用語意表達，與側重以形態變化來表達的俄語有所區別。

另一方面，肖靜和王惠蓮（2009）使用「HSK 動態作文語料庫」，針對母語為英語的外國留學生的作文進行分析，發現留學生的「了」字句偏誤主要受到兩個因素影響。首先是「語際干擾」，學生受到母語的負遷移，把英語的過去時和完成態對應到中文的「了」。因此若要表示過去發生的動作或事件，就會泛化使用「了」。「語內干擾」為第二個因素，由於漢語與英語不同，缺乏形態變化而著重意念，學習者無法掌握「了」字複雜的性質、用法和使用條件，而造成偏誤。

針對學習漢語一年以上的日本學生，孫海平（2010）分析了他們的寫作和造句練習，並歸納出與肖靜和王惠蓮（2009）相似的偏誤原因：母語負遷移的影響，和「了」本身複雜的使用規則。孫海平（2010）提出的第三個原因為漢語教材和教學，認為在教材中「了」字句缺乏系統且合理的編排，常誤導學生將「了」視為過去時態標記。且在中國國內的對外漢語教材，通常只在初級介紹「了」字句，在中高級卻缺乏深入和強化的教學，不利於學生學習。

孔令達（1993）曾研究中文母語者之「了」字句習得，並指出90名12個月到五歲大的孩童在兩歲六個月前先習得句末「了₂」，之後才習得動詞後「了₁」，而「了₁」和「了₂」共同出現的句子在三歲六個月後才最後習得。鄧守信（1999）進而指出外語學習者也是先習得「了₂」再習得「了₁」，強調第一語言習得和第二語言習得的相似度，並主張華語教材應先教「了₂」再教「了₁」。由於「了₂」易學，應盡早教授；而「了₁」

則應該在學習者對基本動作動詞和過去時間狀語有所了解後再行教授。

綜合上述前人文獻，本文發現前人研究對於「了」字句的偏誤較缺乏藉自然語料中系統性的分析，因此本文希望不只能了解學習者的使用情形，還能針對其偏誤進行分類，並試圖找出可能的原因。另外，前人文獻中也較缺乏以學習者語料庫為本的研究。有些研究僅以教師的課堂觀察為基礎，而沒有大量且系統性的語料分析。因此本文希望能使用「華語學習者語料庫」，歸納出大量學習者的習得概況和歷程。

3. 研究方法

3.1 研究語料

本研究藉臺灣師範大學自 2006 年開始建置之「華語學習者語料庫」³作為研究材料。該語料庫內容為書面語語料，而語料來源有二：一是 2006 年至 2010 年 TOCFL 寫作考試的考生作文，總計約一百萬字；二是 2006 年至 2010 年臺灣師範大學國語教學中心的寫作練習，總計約兩百萬字。目前該語料庫開放線上檢索，其檢索介面如下圖 1 所示。使用者可單選、複選或全選不同面向來篩選語料，能從學習者的母語、語料來源、作文文體、作文功能、字數、語料類別、作文級分和學習者的 CEFR 級別等面向來選擇需要的語料。此外，還可選擇關鍵詞的詞性（包含不考慮和其他十個詞性選項）、語料排序（根據關鍵詞前一個詞或後一個詞，或一檔案編號順序排序），以及每頁顯示筆數。

母語	來源	文體	功能	字數	類別	級分	CEFR LEVEL
[全選]	[全選]	[全選]	[全選]		[全選]	[全選]	[全選]
土耳其語 中文 日語 立陶宛語 匈牙利語 印尼語 吉里巴斯語 西班牙語 希伯來語 波蘭語 法語	紙本 電子	信件 便條 看圖作文 記敘文 圖表分析 應用文 議論文	支撐觀點 拒絕邀請 表達不滿意 表達可能性 表達同情 表達偏好 表達喜好 表達感受 表達意圖 表達道德感 表達意願	4 字 至 1874 字	比賽 預試 練習平臺 練習卷 獎學金	0 1 1.5 2 2.5 3 3.5 4 4.5 5 未評分	A2 B1 B2 C1
查詢詞彙 <input type="text"/> 詞性 不考慮 排序 後面詞排序 每頁筆數 十筆							
<input type="button" value="查詢"/>							

圖 1. 「華語學習者語料庫」檢索介面

本文取「臺師大華語學習者語料庫」中 2006 年至 2010 年 TOCFL 寫作考試總計約一百萬字的考生作文作為研究語料(張莉萍, 2013)，該語料庫已依學習者的母語和 CEFR 級別來篩選語料。考生程度依 CEFR 級別由低至高分為 A2、B1、B2、C1 等四級。本文檢視了各級別之語料，發現語料庫中以 A2 及 B1 的語料量較多，因此本文以程度為 A2 和 B1 的學習者語料為研究內容，期藉著語料庫文本能更進一步了解初級和中級華語學習者「了」的學習狀況。另外，華語學習者語料庫中涵蓋不同母語背景學習者，然而因

³ 「華語學習者語料庫-TOCFL 寫作語料」<http://kitty.2y.idv.tw/~hjchen/cwrite/main.cgi>

「華語學習者語料庫-MTC 寫作語料」<http://kitty.2y.idv.tw/~hjchen/cwrite-mtc/main.cgi>

篇幅關係，本次研究以母語背景為英語學習者為主，一方面是英語為母語學習者語料量較多，另一方面亦希望能自分析探討英語母語者是否會將「了」泛化為過去時的標誌而造成偏誤。期待未來能更進一步探究其他語言背景學習者之偏誤情況。

3.2 語料庫使用方法

於「臺師大華語學習者語料庫」線上檢索系統選定母語為英語和 CEFR 級別為 A2 及 B1，並輸入查詢詞彙「了」之後分別得到符合檢索條件的語料筆數為 385 筆及 797 筆，關鍵字「了」以紅色字體標出並置於搜尋結果中央（見下圖 2）。若需要更多某筆語料的背景訊息，可點選該語料左方檔案欄內的編號，即可連結至語料的原始畫面，取得更多前後文和學習者母語、語料來源、作文文體、作文功能、字數、語料類別、作文級分和學習者的 CEFR 級別等所有資訊。

符合筆數 385 筆	
第 19 / 頁	
檔案	內容
0800	·可是我回來山下，就發現我的車子被小偷偷走了了/語助詞!/? 因為我真的不知道怎麼辦，所以我在路上站著擔心
0671	上課了。上個夏天你有沒有上班?我希望你錢夠了/語助詞!/? 再見。
0813	去。我們五點半到了，我很高興可是非常累了/語助詞!/?
0749	帶水。所以晚上她覺得不舒服。他說「可能感冒了/語助詞!/?」。他還沒看一個人。他不放心!是
0107	·我要過跟他談談，可是老師告訴我「別說了/語助詞!/?」。林美美只笑我。下了課以後，我們跟
2992	李大明，你以後再到夜店明天就不必來我的課了/語助詞!/? 大明怕了，就答?老師不會再晚到學校來

圖 2. 「華語學習者語料庫」中「了」檢索結果

4. 研究結果與討論

本文取得 A2 級考生語料(以下簡稱 A2 語料)共 385 筆,扣除整句意義不明和「了」(liao)字用法共 15 筆語料後,分析的有效語料為 370 筆。其中,偏誤用法為 38 筆,約佔 10.3%,顯示學習者已能掌握相當程度的「了」字用法。而再將「了」細分為「了₁」、「了₂」和「了₁₊₂」後,分析結果顯示,A2 級學習者使用最多的是「了₁」,共有 177 筆語料,其次是「了₂」(114 筆)和「了₁₊₂」(79 筆)。而以偏誤率來看,學習者使用「了₁」的偏誤率也最高,為 13.0%,其次為「了₂」,偏誤率為 9.6%,而「了₁₊₂」的偏誤率僅 5.1%。詳細分布情形如表 2。

表 2. A2 語料正誤分布

	偏誤用法數		正確用法數		總數
了 ₁	23	13.0%	154	87.0%	177
了 ₂	11	9.6%	103	90.4%	114
了 ₁₊₂	4	5.1%	75	94.9%	79
總數	38	10.3%	332	89.7%	370

B1 級考生語料(以下簡稱 B1 語料)則共計 797 筆,扣除整句意義不明和「了」(*liao*)字用法共 34 筆語料後,分析的有效語料為 763 筆。其中,偏誤用法為 136 筆,約佔 17.8%,偏誤率較 A2 級稍微提升,不過仍偏低,並接近鄧守信(1999)所研究的中初級華語學習者正確使用率(82.7%)。將「了」細分為「了₁」、「了₂」和「了₁₊₂」後,B1 級學習者使用最多的仍是「了₁」,共有 402 筆語料,其次也為「了₂」(242 筆)和「了₁₊₂」(119 筆)。而以偏誤率來看,學習者使用「了₁」的偏誤率也最高,為 23.6%。「了₁₊₂」和「了₂」的偏誤率則依序為 13.4%和 10.3%,詳細分布情形如表 3。

表 3. B1 語料正誤分布

	偏誤用法數		正確用法數		總數
	筆數	百分比	筆數	百分比	
了 ₁	95	23.6%	307	76.4%	402
了 ₂	25	10.3%	217	89.7%	242
了 ₁₊₂	16	13.4%	103	86.6%	119
總數	136	17.8%	627	82.2%	763

此外,本文根據前人文獻中呂叔湘(1980)所區分的六種「了」字句型,分別統計 A2 級學習者和 B1 級學習者在六種句型中的偏誤筆數和偏誤率,如下表 4 所示(斜線前是偏誤筆數,後是總筆數,括號內是偏誤率):

表 4. A2 及 B1 語料六大「了」字句型偏誤

	偏誤句數	
	A2 語料	B1 語料
1. 動+了 ₁ +賓	23/170 (13.5%)	76/367 (20.7%)
2. 動+賓+了 ₂	7/64 (10.9%)	13/100 (13%)
3. 動+了 ₁ +賓+了 ₂	0/6 (0%)	0/16 (0%)
4. 動+了 ₁ /了 ₂ /了 ₁₊₂	7/102 (6.9%)	37/204 (18.1%)
5. 形+了 ₂	0/21 (0%)	9/62 (14.5%)
6. 名詞/數量詞+了 ₂	1/7 (14.3%)	1/14 (7.1%)
總數	38/370 (10.3%)	136/763 (17.8%)

根據表 4，A2 級學習者最常使用的是第一類：動+了₁+賓，共 170 筆；其次為 102 筆語料的第四類：動+了₁/了₂/了₁₊₂，再來依序是第二類：動+賓+了₂（64 筆）和第五類：形+了₂（21 筆）。第三和六類則鮮少為學習者使用。以偏誤率來看，第六類偏誤率雖然最高，但整類語料僅七筆，偏誤語料僅一筆。若排除第六類，則第一類的偏誤率最高，為 13.5%，其次才是第二類（10.9%）。

B1 級學習者最常使用的是第一類：動+了₁+賓，共 367 筆；其次為 204 筆語料的第四類：動+了₁/了₂/了₁₊₂，再來依序是第二類：動+賓+了₂（100 筆）和第五類：形+了₂（62 筆）。第三和六類則鮮少為學習者使用。以偏誤率來看，第一類的偏誤率最高，為 20.7%，其次為第四類（18.1%）。

對 A2 級和 B1 級學習者而言，「了₁」都較難掌握，而「了₂」則相對較為容易，此發現與鄧守信（1999）相似，且依據鄧守信（1999）建議在教學上可多就「了₂」和相關句型先讓學習者多加以演練，再介紹「了₁」和相關句型，最後介紹「了₁」和「了₂」共同存在的句型，期能由簡入繁，逐步引導學習者學習「了」字句。

以下將進一步針對學習者的偏誤分門別類，依序探討「了₁」、「了₂」和「了₁₊₂」的偏誤。在每類偏誤的標題後，學習者的偏誤數會包含在括弧中。學習者例句中的「了」字偏誤也會特別標記。而 A2 級和 B1 級學習者加起來，偏誤數仍低於三筆的類別，將不在本文討論範圍；因筆數太少，較難看出學習者整體的學習狀況。首先為 A2 級學習者的偏誤分類：

4.1 A2語料之「了₁」偏誤

本文進一步觀察 A2 學習者「了₁」的 23 筆偏誤中，發現主要可歸納為 3 類偏誤句型：1. 「了₁」過度泛化；2. 與動詞後補語混淆；3. 「了₁」多餘，下文將逐一介紹各個偏誤句型例句。

4.1.1 「了₁」過度泛化（9例）

- (1) (誤) *他想到<了>很特別的想法。
(正) 他想到很特別的想法。

根據王媚和張艷榮（2007），表示動作完成並不一定要用「了₁」，結果補語有時也可表示動作完成。在例句中，結果補語「到」已表示了動作「想」的完成，因此「了₁」在句中為非必要。學習者可能是將「了₁」泛化為過去時的標誌，而加上「了₁」。趙立江（1997）也觀察到此類例句，認為學習者受到其母語英語的影響，而在過去發生的事件中盡可能使用「了」，在趙立江的研究中，學習者在這類的偏誤中使用率在前兩個階段高達 81%和 50%，在第三個階段才降為 16%。

4.1.2 與動詞後補語混淆 (2例)

- (2) (誤) *恭喜你找<了>工作！
 (正) 恭喜你找到工作！

與上述相關，王媚和張艷榮 (2007) 指出，在表示持續狀態的動詞後，若要加上表動作完成的「了₁」，動詞後應先加上結果補語賦予界限意義。例如例 2，應為「恭喜你找到工作」。

4.1.3 「了₁」多餘 (7例)

- (3) (誤) *李台生告訴那個男生天氣真熱，他也還得走<了>五個公里的路。
 (正) 李台生告訴那個男生天氣真熱，他也還得走五個公里的路。
- (4) (誤) *我喜歡吃台灣菜可是現在我胖了一點所以我決定<了>每天去運動。
 (正) 我喜歡吃台灣菜可是現在我胖了一點所以我決定每天去運動。

「了₁」放於動作後表示動作完成，但在例句 3 中，走路的動作尚未完成，因此，不應加「了₁」。另一項「了₁」多餘的情況是，若謂語動詞後帶著賓語小句，則動詞後一般不加「了₁」(李大忠, 1996; 肖靜和王惠蓮, 2009; 孫海平, 2010)。王媚和張艷榮 (2007) 提到，若語義表達的重點放在由動詞或主謂短語等擔任的謂詞性賓語，而不強調謂語動詞的完成，則謂語動詞後一般不加「了₁」。而在例句 4 中，重點為「每天去運動」此賓語小句，因此「決定」後應不加「了₁」。

4.2 A2語料之「了₂」偏誤

在 A2 語料之「了₂」偏誤中，本文整理了主要有 2 類的偏誤句型，請見下文。

4.2.1 「了₂」多餘 (7例)

- (5) (誤) *我還記得我第一次在台灣上中文課<了>。因為我是外國人所以我剛來的時候跟台北一點也不熟，常常迷路了。
 (正) 我還記得我第一次在台灣上中文課。因為我是外國人所以我剛來的時候跟台北一點也不熟，常常迷路了。

在例句 5 中，句中涵意是回憶初次上課的過程。那時還未上課，狀態沒有改變，因此不能加上核心語義為「+狀態改變」的「了₂」。

4.2.2 與「了₁」混淆 (3例)

- (6) (誤) *叫了他以後，他就停車<了>幫我。
 (正) 叫了他以後，他就停了車幫我。

在句 6 中，應該用「了₁」，改為「他就停了車幫我」，強調「停」的動作完成。但學習者用了「了₂」，因此為病句。

4.3 A2語料之「了₁₊₂」偏誤

A2 學習者主要是「了₁₊₂」多餘之偏誤情況。在 A2 的語料中有 2 例。請見例句 7。

- (7) (誤) *每天，他六點半就醒來，要不然一定上課遲到<了>。
 (正) 每天，他六點半就醒來，要不然一定上課遲到。

在本句中，「遲到」的動作並沒有完成，狀態也無改變，因此若加上「了₁₊₂」則為冗餘之偏誤。

4.4 B1語料之「了₁」偏誤

臺師大華語學習者語料庫中母語為英語之 B1 學習者在「了₁」的常見偏誤主要有 4 類：1. 「了₁」過度泛化；2. 與動詞後補語混淆；3. 「了₁」多餘；4. 與「得」混淆。

4.4.1 「了₁」過度泛化 (43例)

- (8) (誤) *在你的生日會上，我看到<了>一個男生，叫志華。
 (正) 在你的生日會上，我看到一個男生，叫志華。

根據王媚和張艷榮(2007)，結果補語有時也可表示動作完成。例句 8 中，結果補語「到」已表示了動作「看」的完成，因此「了₁」在句中為非必要。

4.4.2 與動詞後補語混淆 (2例)

(9) (誤) *如果我學<了>常常對媽媽說「謝謝」,她一定會很高興。

(正) 如果我學會常常對媽媽說「謝謝」,她一定會很高興。

王媚和張艷榮(2007)也指出,在表示持續狀態的動詞後,應先加上結果補語賦予界限意義。而例句9中,應在「學」之後加上「會」,而後加不加「了₁」並不影響句意。

4.4.3 「了₁」多餘 (40例)

(10) (誤) *不過在實際上我還沒決定<了>我的專業。

(正) 不過在實際上我還沒決定我的專業。

(11) (誤) *所以只需要跑<了>半天,就到學校!

(正) 所以只需要跑半天,就到學校!

(12) (誤) *每次看到買帽子的攤子都回十分興奮,但每次不敢買因為總有朋友在我背上潑<了>冷水的說,「你不適合帶」的這一句話。

(正) 每次看到買帽子的攤子都回十分興奮,但每次不敢買因為總有朋友在我背上潑冷水的說,「你不適合帶」的這一句話。

(13) (誤) *我看到了一些餐廳,很快就發現<了>台灣的用餐文化跟國外截然不同!

(正) 我看到了一些餐廳,很快就發現台灣的用餐文化跟國外截然不同!

若動詞前有否定副詞「沒(有)」,表示動作沒有實現或完成,因此動詞後不可加「了₁」。(李大忠,1996;趙立江,1997;尚靜和王惠蓮,2009;孫海平,2010)在例句10中,「決定」的動作尚未完成,因此不需加「了₁」。

例句11是由兩個動作組成的事件,要標示事件完成,「了₁」只需置於最重要的動詞,也就是最後一個動詞「到」之後。錯誤放置「了₁」將標示出錯誤的信息高峰,因此「跑」後的「了₁」應刪除。(陳俊光,2008)。

根據屈承熹(1999),「了₁」常出現在主句。在例句12中,「了₁」出現在子句(從句)「潑了冷水的」中,造成病句。而根據陳俊光(2008),此句法限制與「了₁」的篇章功能「高峰標記」相關,因為信息焦點多出現在主句表達,「了₁」往往也出現在主句而非從句。

王媚和張艷榮（2007）指出，若語義表達的重點不在謂語動詞的完成，而是強調由動詞或主謂短語等擔任的謂詞性賓語，則謂語動詞後一般不加「了₁」。在例句 13 中，重點為「台灣的用餐文化跟國外截然不同」，因此「發現」後不加「了₁」。

4.4.4 與「得」混淆（5例）

(14) (誤) *希望你這幾天你的日子過<了>很好。

(正) 希望你這幾天你的日子過得很好。

例句 14 正確講法應為「過得很好」，在王媚和張艷榮（2007）的研究也記載了相似的偏誤，並建議應教授學生「動詞+得+形容詞」的格式以避免此類錯誤。

4.5 B1語料「了₂」與「了₁₊₂」之偏誤

B1 語料中「了₂」和「了₁₊₂」英語為母語學習者最常出現的偏誤是「了₂」（24 例）及「了₁₊₂」（13 例）之冗餘，下文示例部分例句。

(15) (誤) *我在櫃台等一下店員來幫我，不過沒有人要來<了>。

(正) 我在櫃台等一下店員來幫我，不過沒有人要來。

(16) (誤) *沒有辦法<了>，我的人生最快樂的時刻是能夠吃好吃的食物。

(正) 沒有辦法，我的人生最快樂的時刻是能夠吃好吃的食物。

(17) (誤) *如果我們的時間夠<了>，我們可以到台南去。

(正) 如果我們的時間夠，我們可以到台南去。

在例句 15 中，「沒有人來」代表狀態沒有改變，因此句尾不需加「了₂」。

例句 16 裡，學習者將表示「敘述段落」的「了₂」誤置於語段中間，造成整體不連貫。若將「了₂」置於語段最後，改為「我的人生最快樂的時刻是能夠吃好吃的食物了」，才能發揮「了₂」的篇章功能，標示段落的結束。

根據湯廷池（1999）表示，「了₂」常用於主句，而不用於從句中。陳俊光（2008）也認為這是因為「了₂」擁有表示「敘述段落」的篇章功能。在例句 17，學習者將「了₂」置於從句中，錯誤標記了敘述段落，而造成病句。

- (18) (誤) *吃飯的時候希望我們可以變熟<了>。
(正) 吃飯的時候希望我們可以變熟。
- (19) (誤) *我先跟老師說<了>，再決定換班了。
(正) 我先跟老師說，再決定換班了。
- (20) (誤) *雖然他古代時逝世<了>，但他所做的事情天長地久對全球人民影響很深遠。
(正) 雖然他古代時逝世，但他所做的事情天長地久對全球人民影響很深遠。

在上文例句 18 中，「變熟」的動作還沒有完成，狀態也無改變，因此視為「了₁₊₂」之冗餘。例句 19 的事件包含兩個動作：「說」和「換班」，要標示事件完成，「了₁₊₂」只需置於最重要的動詞，也就是最後一個動詞「換班」之後。「說」之後的「了₁₊₂」標示出錯誤的信息高峰，造成病句。同時，句中的「了₁₊₂」也違反了「了₂」「敘述段落」的篇章功能，誤置於語段中間，造成整體不連貫。反觀「再決定換班了」中的「了」，才真正發揮了標示敘述段落結束的功能。

「了₁」和「了₂」都常用於主句而非從句中(湯廷池, 1999; 屈承熹, 1999)，陳俊光(2008)並進一步指出此項句法限制源自於「了₁」和「了₂」的篇章功能。例句 20 傳遞的是「他所做的事情天長地久對全球人民影響很深遠」這件事，此句為信息焦點和敘述段落的完結的主句，因此在從句「雖然他古代時逝世了」加上「了」即為病句，應刪除「了」。

綜上所述，A2 級學習者和 B1 級學習者的偏誤類型各有異同，本文整理成表以供比對，請見下表 5：

表 5. A2 和 B1 語料偏誤句數比對

	偏誤類型	A2 語料例數	B1 語料例數
「了 ₁ 」偏誤	1. 「了 ₁ 」過度泛化	9	43
	2. 與動詞後補語混淆	2	2
	3. 「了 ₁ 」多餘	7	40
	4. 與「得」混淆	0	5
「了 ₂ 」偏誤	1. 「了 ₂ 」多餘	10	24
	2. 與「了 ₁ 」混淆	3	0
「了 ₁₊₂ 」偏誤	1. 「了 ₁₊₂ 」多餘	2	13

由上表可得知，A2 和 B1 級學習者都常過度泛用「了₁」，顯示學習者可能將「了₁」泛化為過去時的標誌，傾向於以「了₁」表示事件完成，而不知道結果補語有時也可表示動作完成，並賦予表示持續狀態的動詞界限意義。此外，B1 級學習者還會將「了₁」與「得」混淆。另一方面，關於「了₂」及「了₁₊₂」的偏誤，A2 和 B1 級學習者均出現「了₂」及「了₁₊₂」冗餘之偏誤句。

5. 結論

在本文研究中，藉臺師大學習者語料庫 TOCFL 語料中英語為母語者之 A2 和 B1 華語學習者為研究內容，探討各程度學習者在「了」字句用法的偏誤情況，以量化方式標示出 A2 及 B1 學習者針對「了」字句的偏誤句，並提出修正。研究結果得出 A2 級考生正確率為 89.7%，B1 級考生正確率為 82.2%，此結果接近鄧守信（1999）中初級華語學習者的正確使用率（82.7%）。

若將「了」分為三類，A2 級學習者和 B1 級學習者相同，最常使用的都是「了₁」，其次為「了₂」，最少使用的是「了₁₊₂」。若以偏誤率來看，A2 級學習者「了₁」的偏誤率最高，再來為「了₂」和「了₁₊₂」。B1 級學習者「了₁」的偏誤率也最高，其次為「了₁₊₂」和「了₂」。

本文分析偏誤後發現，英語為母語的學習者的確有將「了」泛化為過去時標記的情形。例如在動詞後有結果補語的句中，就算結果補語已表達動作完成，學習者仍加上「了」，標示在過去已完成的動作。

若要減少偏誤的發生，教師在教學上可強調「了₁」、「了₂」和「了₁₊₂」核心語義，避免學生誤用和混淆。此外，應介紹「了」和語義焦點的關係，並說明「了」與結果補語的語義，避免學生將「了」泛化為過去時的標誌。最後，針對中級學習者，也能加入關於「了」篇章功能和句法分布的教學，並與其他結構如「得」相比較以避免混淆。

以呂叔湘（1980）提到的六大「了」字句型來看，本文與鄧守信（1999）的發現相似，A2 級學習者第一類「動+了₁+賓」的偏誤數最多，排除僅有一筆偏誤數的第六類，第一類的偏誤率也最高，顯示「了₁」的學習難點。而第二「動+賓+了₂」、五「形+了₂」、六「名詞／數量詞+了₂」類的偏誤數則少、偏誤率較低，顯示「了₂」較容易習得。最後，第三類「動+了₁+賓+了₂」的使用率和錯誤率最低。B1 級學習者也有相仿的傾向，因此，建議在教學上可先介紹較容易理解的「了₂」和其相關句型，再介紹「了₁」的相關句型，而「了₁」、「了₂」同時存在的第三類句型則可留到最後。希望能由簡入繁，正確引導學習者並減少偏誤發生。

本文在研究上仍有許多限制，例如語料量上的不平均，A2 及 B1 的語料較多，但 B2 及 C1 的語料不足，因此未能更全面的針對各層級進行偏誤分析。另一方面，因篇幅關係本文僅探討母語為英語學習者在「了」字句上的偏誤情形，期望未來研究可多針對各語言背景學習者在「了」字句上的使用情況。

本文使用「華語學習者語料庫」，以大量的學習者語料為本，探討初級和中級華語學習者在學習「了」字句時，相似與相異的使用情形和偏誤類型。「華語學習者語料庫」為一共時性語料庫，語料來源為不同學習者在同一階段的語言使用；因此本文的分析對象不僅限於單一或少數學習者，所得結果較能接近多數學習者的普遍習得歷程。最後，期本文所歸納出教學時應釐清的重要觀念和建議之教學順序，能提供華語教材編寫的參考方向。

引用文獻

- Chang, W. V. (1986). *The particle LE in Chinese narrative discourse*. Gainesville, FL: University of Florida doctoral dissertation.
- Chu, C. C. (1999). *Discourse Grammar of Mandarin Chinese*. New York: Peter Lang Publishing.
- 孔令達 (1993)。兒童了字句的發展及相關問題的討論。第三屆全國現代語言學研討會論文集。北京：語文出版社。[Kong, L.-d. (1993). Discussion of Children's Development of "Le" Construction and Relevant Issues. *The Symposium of The Third National Conference on Linguistics*. Beijing: Language & Culture Press.]
- 王媚、張艷榮 (2007)。俄羅斯留學生“了”字句使用偏誤分析。雲南師範大學學報，5(1)，47-51。[Wang, M., & Zhang, Y.-r. (2007). An Analysis of the Errors in the Sentences with "Le" Made by Russian Students. *Journal of Yunnan Normal University*. 5(1), 47-51.]
- 李大忠 (1996)。外國人學漢語語法偏誤分析。北京：北京語言文化大學。[Li, D.-z. (1996). *Analysis of Mistakes of Foreign Learners in Chinese Grammar*. Beijing: Beijing Language and Culture University Press.]
- 呂叔湘 (1980)。現代漢語八百詞(增訂本)。北京：商務印書館。[Lu, S.-x. (1980). *Xiandai Hanyu Ba Bai Ci (Revised Version)*. Beijing: The Commercial Press.]
- 尚靜、王惠蓮 (2009)。以英語為母語的外國留學生“了”的偏誤分析——基于 HSK 動態作文語料庫。大眾文藝，第 1 期，138。[Xiao, J., & Wang, H.-l. (2009). The Error Analysis of "Le" Used by English Students—Based on HSK Dynamic Composition Corpus. *Art and Literature for the Masses*, 1, 138.]
- 屈承熹 (1999)。漢語認知功能語法。台北市：文鶴出版有限公司。[Qu, C.-x. (1999). *A Cognitive-Functional grammar of Mandarin Chinese*. Taipei: Crane Publishing Co., Ltd.]
- 屈承熹 (2003)。功能篇章語法及其在對外漢語教學上的應用。對外漢語教學語法探索。北京：中國社會科學出版社。[Qu, C.-x. (2003). Functional-Discourse Grammar and its Application on Chinese Teaching. *Exploration of Chinese Teaching Grammar*. Beijing: China Social Science Press.]
- 孫海平 (2010)。日本學生使用動態助詞“了”的偏誤分析。現代語文，第 11 期，136-139。[Sun, H.-p. (2010). The Error Analysis of the Particle "Le" Used by Japanese Students. *Modern Chinese*. 11, 136-139.]

- 陳俊光 (2008)。《對比分析與教學應用》。台北市：文鶴出版有限公司。[Chen, J.-g. (2008). *Contrastive analysis and its applications in language pedagogy*. Taipei: CranePublishingCo.,Ltd.]
- 張莉萍 (2013)。TOCFL 作文語料庫的建置與應用，載於崔希亮、張寶林（主編），《第二屆漢語中介語語料庫建設與應用國際學術討論會論文選集》（頁 141-152）。北京：北京語言大學出版社。
- 張寶林、崔希亮、任傑 (2004)。關於“HSK 動態作文語料庫”的建設構想。《第三屆全國語言文字應用學術研討會論文集》。[Zhang, B.-l., Cui, X.-l., & Ren, J. (2004). The Construction Concept of HSK Dynamic Composition Corpus. *The Symposium of The Third National Conference on Language Application*.]
- 湯廷池 (1999)。華語助詞「了」的意義與用法。《華文世界》，第 93 期，51-54。[Tang, T.-c. (1999). The Meaning and Usage of Chinese Particle “Le.” *The World of Chinese Language*. 93, 51-54.]
- 趙立江 (1997)。留學生“了”的習得過程考察與分析。《語言教學與研究》，第 2 期，112-124。[Zhao, L.-j. (1997). Observation and Analysis of Foreign Students' Acquisition of “Le.” *Language Teaching and Linguistic Studies*. 2, 112-124.]
- 鄧守信 (1999)。Acquisition of LE in L2 Chinese。《世界漢語教學》，第 1 期，56-64。[Teng, S.-h. (1999). Acquisition of LE in L2 Chinese. *Chinese Teaching In The World*. 1, 56-64.]

Cross-Linguistic Error Types of Misused Chinese Based on Learners' Corpora¹

Keiko MOCHIZUKI*, Hiroshi SANO*, Ya-Ming SHEN* and

Chia-Hou WU⁺

Abstract

This paper presents an empirical study on the difficulties in learning Chinese as a second language based on learners' corpora written by native English speakers and native Japanese speakers at CEFR-based A2 and B1 levels. The first part of this paper will discuss the procedures for how to collect learners' corpora, proofread, establish an error tag system and annotate errors. Later it will focus on a significant difference in the production of “ — + Classifier” among the corpora of native English speakers and native Japanese speakers. The corpus of English native speakers displays an overuse of “ — + Classifier”, even in an atelic context like a negative construction or a conditional construction where a “ — + Classifier” should not occur. On the other hand, the corpus of Japanese native speakers displays a lack of “ — + Classifier”. This striking contrast is due to whether or not a determiner position exists in each language. Since English has a determiner position which accommodates an article, “a/an, the”, “this/that/my/your/~’s”, English-native learners tend to treat the “ — + Classifier” as an article although it does not appear in an atelic event structure. On the other hand, Japanese does not have any determiner position before a Noun Phrase, therefore it is assumed that

¹ This project, entitled "Construction of a Japanese-English-Chinese Online Error Corpus and development of English, Japanese and Chinese language pedagogy taking into account learners' native languages (2013-2015)", has been supported by the International Center for Japanese Studies ,Tokyo University of Foreign Studies(henceforth TUFS) and a Type B Research Grant, KAKEN 25284101 from the Japan Society for the Promotion of Science.

* Tokyo University of Foreign Studies

E-mail: mkeiko@tufs.ac.jp; sano@tufs.ac.jp; yamingshen@lingua-house.jp

⁺ National Taiwan Normal University

E-mail: clothoray@gmail.com

Japanese learners find it difficult to learn the conditions where a “ — + Classifier” is necessary.

Keywords: Learner’s Corpus, Annotation System, Error Analysis, Online Dictionary of Misused Chinese based on Learners’ Corpora, Interference of Mother Tongues.

1. Objectives of Constructing the Learners’ Error Corpus

The purposes of constructing the Learners’ Error Corpora can be divided into two categories. The first is to discover the errors made by advanced-level learners since we assume that these errors reflect grammatical difficulties, significant differences in conceptual representation between the target language and the native language, and a different focus of representation despite relatively easy sentence structures. We believe that lexical/syntactic areas that are difficult to learn are caused by cases where the natural language system itself is difficult and where translation is difficult due to negative transfer. Clarifying these differences will lead to improvements in language teaching materials.

The second purpose of the research is to obtain new findings for comparative linguistics. The error analysis of cross-linguistic learners’ corpora will enable us to distinguish language-specific error types based on the learners’ native language and universal error types which occur regardless of the learners’ native languages. Distinguishing these two features will also lead to the improvement of language teaching methodologies, especially those based on comparative perspectives between the learners’ native language and the target language.

2. Procedures

2.1 The ‘Full Moon’ Learner Corpus of Chinese at Tokyo University of Foreign Studies

The characteristics of the data set of the ‘Full Moon’ Learner Corpus of Chinese at Tokyo University of Foreign Studies (henceforth ‘Full Moon Corpus’) are as follows:

Table 1. Learner Corpus of Chinese ‘Full Moon Corpus’ at Tokyo University of Foreign Studies (TUFs), collected May 2013-August 2014.

Academic Year	Level Chinese Major Students	Number of essays	Approximate number of words	Number of students
2013	Advanced (4 th year)	95	45,500	35
	Intermediate (2 nd / 3 rd year)	132	51,200	58
2014	Advanced (4 th year)	21	12,500	23
	Intermediate (2 nd / 3 rd year)	34	25,100	69

These compositions are proofread by Chinese native speakers with an MA. or Ph.D in linguistics/language education and sufficient experience in teaching Chinese at university level. Proofread compositions clearly indicate errors and corrections so that the errors can be identified within the respective sentences.

The 'Full Moon Corpus' includes learner's information as shown in Table 2.

Table 2. Example of Learner's Profile

1	Learner's ID	Th_Ch_001
2	Name	Tokyo Taro
3	Major	Chinese
4	Year	3
5	Gender	male
6	Age	21
7	Nationality	Japan
8	Residential History	Canada 4-9 ; Japan 0-4,9-21
9	Native Language	Japanese
10	Language of Education	Japanese, English
11	Length of Chinese study	3 years and 2 months
12	Institution	Tokyo University of Foreign Studies
13	Study Abroad Experience Institution / Period	Mandarin Center, National Taiwan Normal University, August1-31. 2014
14	Speaking with my family	Japanese
15	Speaking with friends	Japanese
16	Language used in Elementary School	5-9 English, 9-12 Japanese
17	Language used in Junior High School	Japanese, English
18	Language used in Senior High School	Japanese, English
19	Test of Chinese as a Foreign Language (TOCFL)	Band B(2014)
20	HSK 汉语水平考试	5 級 (2012)
21	English TOEFL(iBT)	108 (2013)
22	TOEIC	955 (2012)
23	IELTS (academic)	8.0 (2013)

The 'Full Moon Corpus' has four key features : 1) compositions are written by experienced learners majoring in Chinese in Japan, 2) compositions go through an appropriate proofreading process conducted by university teachers, 3) errors and corresponding corrections are recorded, and 4) the detailed profiles of the learners are also recorded.

2.2 Error Tag Categories

There are two tag categories for misuse: Error and Modify. The Error tag indicates grammatical errors while the Modify tag indicates inappropriate use of expressions ('expression' tag), punctuation and Chinese characters as shown in Figure 1.

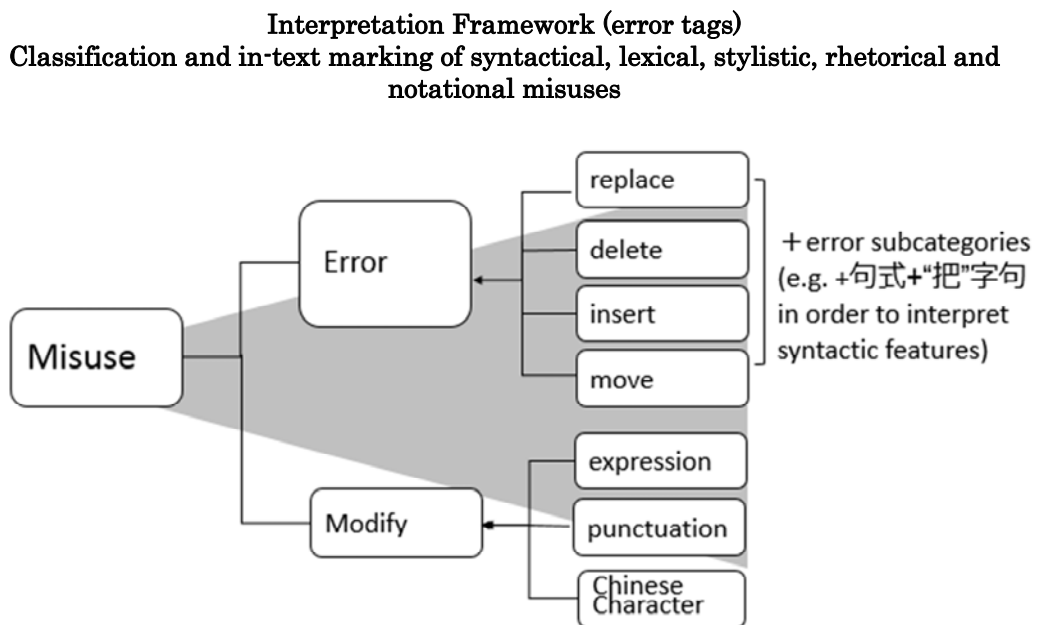


Figure 1. Misuse Tag System

The Error tag consists of the following four sub-categories: Replace, Delete, Insert and Move. The Replace tag indicates the need to replace an error with another correct expression. The Delete tag indicates that deleting an error will lead to a correct expression. The Insert tag indicates that inserting a new expressions will lead to a correct expression. The Move tag indicates a word order error.

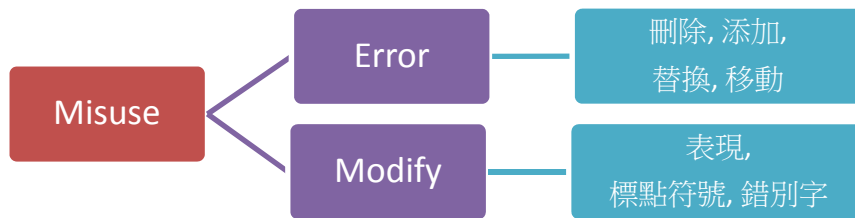
The Modify tag consists of the following three sub-categories: Expression, Punctuation and Chinese character. The Expression tag indicates that it is preferable to use another expression or that the misuse cannot be categorized as any one specific error. The Punctuation

tag indicates the need for correction in view of the style of writing. The Chinese character tag indicates the misuse of a Chinese character.

As subcategories of the error tag, we have designed the 74 tags as shown in (1) referring to the grammatical system in 『新編現代漢語』(張斌、齊滬揚等, 2002: 273-467).

(1) Tag List in Chinese

A. Subcategories of Misuse



B. Subcategories of Error

	大分類	小分類
1	名詞	時間名詞, 處所名詞, 方位詞
2	數詞	
3	量詞	
4	動詞	狀態動詞, 動作動詞, 存現動詞, 關係動詞, 能願動詞, 趨向動詞, 使令動詞
		及物動詞, 不及物動詞, 雙賓動詞
		重疊動詞
5	形容詞	
6	副詞	程度副詞, 範圍副詞, 時間副詞, 情態副詞, 否定副詞, 語氣副詞, 關聯副詞
7	代詞	人稱代詞, 指示代詞, 疑問代詞
8	連詞	
9	介詞	
10	助詞	結構助詞, 時態助詞, 時制助詞, 比況助詞, 表數助詞, 列舉助詞, 語氣助詞, 其他助詞
11	短語	量詞短語, 方位短語, 介詞短語, “的”字短語
12	主語	

13	賓語	雙賓語
14	補語	結果補語, 趨向補語, 可能補語, 程度補語, 情態補語, 數量補語, 介詞, 短語補語
15	疑問句	是非問句, 特指問句, 選擇問句, 正反問句
16	句式	主謂謂語句, “把”字句, “被”字句, 連動句, 強調句, 兼語句, 使役句, 存現句, 比較句, “連”字句
17	複句	並列複句: 承接複句, 遞進複句, 選擇複句, 注解複句
		偏正複句: 因果複句, 條件複句, 轉折複句, 讓步複句, 目的複句

2.3 Method of Proofreading and Annotation

We use the ‘TNR_Chinese Writing Correction2014’ and ‘TNR_Chinese Error Corpus Tagger2014’ (2014) tools developed by 于康(Yu Kang) and 田中良(Ryo Tanaka) for proofreading and annotation. The procedures are as follows. First, compositions written by learners in a WORD file are converted to text files. Next, errors and the corresponding corrections are added to the composition texts using the ‘TNR_Chinese Writing Correction 2014’ system. The following figure 2 is an example of proofreading using ‘TNR_Chinese Writing Correction 2014’.



Figure 2. Proofreading System

The 'TNR_Chinese Writing Correction 2014' system displayed in Figure 4 has two windows: the left window displays the composition text and the right window displays corrections. Each correction in the right window and its corresponding error expression in the left window are marked up in the same color for better visibility.

For annotation, 'TNR_Chinese Writing Correction2014' and 'TNR_Chinese Error Corpus Tagger2014' (2014) enable free creation of tags and the displaying of a tag list underneath the composition text as shown in Figure 3.

错误类型	删除	添加	替换	移动								
其他	标点符号	表现	错别字									
大分类	名词	数词	量词	动词	形容词	副词	代词	连词	介词	助词	短语	主语
	宾语	补语	疑问句	句式	复句							
名词	名词	时间名词	处所名词	方位词								
数词	数词											
量词	量词											
动词	动作动词	存在动词	关系动词	能愿动词	趋向动词	使令动词	状态动词					
动词	及物动词	不及物动词	双宾动词									
动词	重叠动词											
形容词	形容词											
副词	程度副词	范围副词	时间副词	情态副词	否定副词	语气副词	关联副词					
代词	人称代词	指示代词	疑问代词									
连词	连词											
介词	介词											
助词	结构助词	时态助词	时制助词	比况助词	表数助词	列举助词	语气助词	其他助词				
短语	量词短语	方位短语	介词短语	"的"字短语								
主语	主语											
宾语	宾语	双宾语										
补语	结果补语	趋向补语	可能补语	程度补语	情态补语	数量补语	介词短语补语					
疑问句	是非问句	特指问句	选择问句	正反问句								
句式	主谓谓语句	"把"字句	"被"字句	连动句	强调句	兼语句	使役句	存现句	比较句	"连"字句		
联合复句	并列复句	承接复句	递进复句	选择复句	注释复句							
偏正复句	因果复句	条件复句	转折复句	让步复句	目的复句							

Figure 3. Annotation System: Tag Buttons

The first step in annotating a composition is to designate the region of each misused expression in the composition text. The second step is to choose one of 'Replace 替换, Delete 删除, Insert 添加, Move 移动, Expression 表现, Punctuation 标点符号, Chinese Character 错别字' and click on the appropriate button. This procedure enables annotations to be made

automatically. The third step is to choose one of the error subcategories, e.g. ‘Resultative Complement 結果補語’. This click-annotation system greatly reduces the burden of annotation. ‘TNR_ Chinese Writing Correction 2014’ also has the function to convert annotated data into XML data.

Digitization Framework (XML Data)

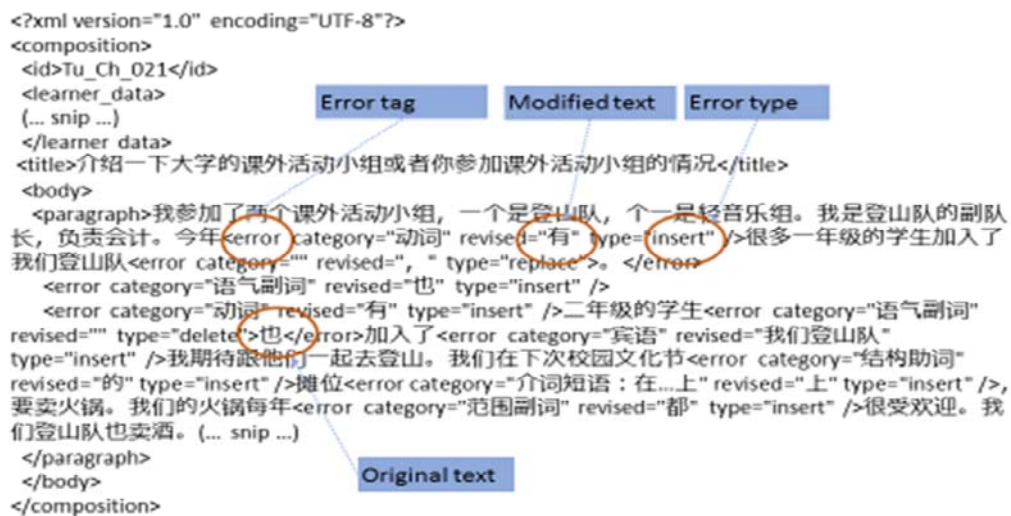


Figure 4. Digitization Framework (XML Data)

3. Cross-linguistic Analysis of Errors

We will discuss two significant error types in two learners’ corpora by comparing The Full Moon Corpus written by Japanese native speakers at TUFU with the TOCFL learners’ corpus of Chinese written by English native speakers (henceforth, TOCFL corpus)². (張莉萍 Chang Li-Ping:2013)

Table 3. the TOCFL English-Native Learners’ Corpus of Chinese

TOCFL (CEFR)	Number of Compositions	Number of Chinese characters	Number of Students
基礎(A2)	223	119,971	223
進階(B1)	344	31,852	344

² Special thanks are due to Professor Chang Li-Ping 張莉萍 and Professor Howard Hao-Jan Chen 陳浩然 at the Mandarin Training Center, National Taiwan Normal University for offering this learners’ corpus and guiding our work with their detailed comments.

3.1 Classifier Phrase(量詞短語) “一 + Classifier(量詞)”

One of the most significant error categories observable in The Full Moon Corpus is the lack of “一 + Classifier(量詞)” while the TOCFL Corpus displays an overuse of “一 + Classifier(量詞)”.張莉萍 Chang Li-Ping(2014:68) also indicates the same contrast between English-Native learners and Japanese-Native learners.

Table 4 compares the frequency of “一 + Classifier ‘-ge 個’ ” in The Full Moon Corpus and the TOCFL Corpus.

Table 4. the Frequency of “一 + Classifier ‘-ge 個’ ”

	CEFR Level	Number of Chinese characters	Occurrence of “一個”
The TOCFL English-Native Learners' Corpus	B1	119,971	586 tokens
	A2	31,852	159 tokens
	Total	151,823	745 tokens 1,490 Chinese characters
The Full Moon Japanese-Native Learners' Corpus	A2-B1	134,094	385 tokens 770 Chinese characters

Table 4 shows an interesting contrast in the frequency of “一個” between The TOCFL English-Native Learners' Corpus and The Full Moon Japanese-Native Learners' Corpus. The TOCFL English-Native Learners' Corpus displays a higher frequency than The Full Moon Japanese-Native Learners' Corpus. Upon conducting a chi squared test, a significant difference between the data sets was discovered (0.1%, $\chi^2=150.03$, $p=0.000$).

3.2 Lack of “一 + Classifier” : Japanese Learners

Let us examine the lack of “一 + Classifier(量詞)” in The Full Moon Japanese-Native Learners' Corpus. The following examples (2) to (18) show that each sentence lacks the bracketed “一 + Classifier” in The Full Moon Japanese-Native Learners' Corpus. There are almost no examples of overuse of “一 + Classifier” in The Full Moon Japanese-Native Learners' Corpus.

(2) Copula “是 Shi” Construction:

‘Topic(Old Information) + “是 Shi”+ Comment(New Information)’

- a. 我認爲這是(一種)有益的愛好。
- b. 但是，生孩子是(一件)不簡單的事。

- c. 從前我去過京都，京都是(一個)很美麗的地方。
- d. 但是，找到好工作並不是是(一件)好的事情。
- e. 小學生跟他們交流是(一個)好機會，但是，孩子們聽得懂他們的課嗎？
- f. 現在，環境問題是世界的(一個)很大的課題。

(3) Existential “You 有” Construction

- a. 東大和有(一個)很大的公園---東大和南公園，附近也有(一條)小河。
- b. 這台電視還有(一個)功能,那就是聽音樂。

(4) Perfective Construction with “-le 了”

這幾年，一位很有名的漫畫家畫了(一部)跟帶廣的挽曳賽馬有關的漫畫。

(5) “Give” Construction and “Become” Construction

- a. 比如，開始工作掙錢以後，我想送給父母(一份)禮物，例如，海外旅行。
- b. 我也在留心這些事情，希望能成為(一個)很好的領導!!

(6) Presentative Construction

最近他在車站附近開了(一家)中餐館。

(7) Resultative/Directional Verb Compound

其他同學也舉出(一些)有意思的食物，比如納豆，豆漿等等。

(8) “Modifier +的 DE+ Noun”

- a. 原來有很多溫泉的日本的(一個)特色就是飯店旅館業很發達。
- b. 在我的印象裡很深的(一件)事是小學 5 年級的時候媽媽幫助我練習跑步。
- c. 去年網絡上的(一篇)文章“中國女性和日本女性的一生”引人註目。

(9) ‘Source’ with New Information:

那個名字來源于(一條)從南到北延伸的坡道。

The reason why it is very difficult for Japanese learners of Chinese to learn the principle of “一 + Classifier” is because Japanese grammar is insensitive to ‘Boundedness’ (有界性) which controls the occurrence of “一 + Classifier”.

Shen(沈家煊) (1995) discusses the interaction between “一 + Classifier” and the concept of ‘bounded’ and ‘unbounded’ events. Shen (1995) indicates that a “一 + Classifier” is necessary before a ‘bounded’ Noun Phrase(NP) in ‘Telic’ events as follows:

(10) Indirect Object in a Move Construction:

- a. 盛碗裡兩條魚。
- b. *盛碗裡魚。

(11) Resultative Object (結果賓語)

- a. 蚊子叮了小王兩個大包。
- b. *蚊子叮了小王大包。

(12) Resultative Complement (結果補語)

- a. 打破兩塊玻璃。
- b. *打破玻璃。

(13) Directional Complement(趨向補語)

- a. 飛進來一個蒼蠅。
- b. *飛進來蒼蠅。

(14) Verb+ “-le 了” construction

- a. 吃了一個蘋果。
- b. *吃了蘋果。

Shen (1995)’s “bounded/unbounded” theory can explain why the following types, (4) Perfective Construction with “-le 了”, (5) GOAL in “Give” Construction and “Become” Construction, (6) Presentative Construction and (7) Resultative/Directional Verb Compound require “一 + Classifier” since all cases in (4)(5)(6)(7) have “telicity”, the subcategory of “bounded” concept in the temporal structure.

In (2) Copula “是 Shi” Judgement Construction and (3) Existential “You 有” Construction, “一 + Classifier” often appears after “是 Shi” / “You 有”. Both constructions have the following informational structure:

(15)

“是 Shi”/ “You 有”Construction	Topic	“是 Shi” / “You 有”	“一 + Classifier” NP
1) Informational Structure	Old Information		New Information
2) Boundedness			Bounded

It is supposed that the NP with new information is a bounded entity, because the NP with new information is a focus in terms of cognition.

3.3 Overuse of “一 + Classifier(量詞)” :English-Native Learners

We find the reverse phenomenon in The TOCFL English-Native Learners’ Corpus: the overuse of “一 + Classifier”. The following examples (16) to (23) show that the bracketed “一 + Classifier” should be deleted .

(16) Conditional:

有什麼問題就跟我打(一通)電話吧！

(17) Plan:

我們游完泳我計畫我們去電影院看(一部)電影。

(18) Potential:

a.我們也可以去「西門町」看電影，打撞球，或去(一個)茶店談天說笑。

b.你看我已經可以用中文寫(一封)信，...

(19) Future Activity:

我記得你說過你喜歡丟飛盤，所以我會把(一張)飛盤帶來。

(20) Topic Noun in “是 Shi” construction:

我媽媽上上個週末來台灣看我。我們去的(一個)地方是花蓮。

(21) “When” Clause: Old Information

你開(一個)慶祝會的時候我不能參加是因為我在外國工作。

(22) Negation: “沒(有)”

a. 我在台北沒有發生(一個)大問題，……

b. 他們有一個農場，我去他們的家以前，還沒去(一個)農場…

(23) Missed Action:

今天他不但忘了帶手機，也忘了帶(一瓶)水。

It seems that the interlanguage of Chinese created by English native speakers displays the following incorrect overgeneralization:

(24) Overgeneralization by English-native learners of Chinese

a/an NP = “ — + Classifier” NP

Shen (1995)’s “bounded/unbounded” theory can also explain why “ — + Classifier” cannot appear in (16) to (23): all cases express atelic events and an entity in an atelic event should be unbounded. Shen (1995) indicates that a “ — + Classifier” cannot appear in the following atelic structures.

(25) Verb Reduplication(動詞重疊式):

a. (*) 今天要談談兩個問題。

b. *星期天在家洗洗一件衣服。

(26) Durative Aspect Marker “-Zhe 著”

a. Progressive Aspect: *他正吃著三碗飯。

b. Resultative State: *山上架著兩門炮。

(27) Negation:

a. *今天不談兩個問題。

b. *這個月不演三場電影。

3.4 Comparative Analysis of Error Types by Japanese Learners and English-Native Learners

The contrast between the lack of “ — + Classifier” in The Full Moon Japanese-Native Learners’ Corpus and the overuse of “ — + Classifier” in The TOCFL English-Native Learners’ Corpus suggests a difference in Noun Phrase Structures in Chinese, English, and Japanese.

Japanese syntax has no ‘functional category’, therefore there is no syntactic node (i.e. ‘determiner’) to accommodate a constituent like “a/an, the” while English has ‘determiner’ as Fukui (1995) proposes. This syntactic difference between English and Japanese causes the contrast between the lack and the overuse of “ — + Classifier” in Japanese-native learners and English-native learners.

In addition, Ikegami(池上)(1981), (2007) and Kageyama(影山)(1997), (2002) suggest that Japanese is an “unboundedness-oriented” “less-individualization” type language in terms of having no grammatical category of number, ellipsis of subject/object, and no determiner node. This “unboundedness-oriented”, “less-individualization” feature is reflected in second language acquisition of Chinese and English by Japanese learners. Since Japanese grammar has no syntactic strategy to individualize an entity/event, it is very difficult to acquire both the principle of “ — + Classifier” NP which appears in an bounded/individualized noun, and the usage of the articles “a/an, the” in English. According to “NTNU/TUFS Sunrise Learners’ Corpus of English”, the most frequent error category in the Japanese-native learners corpus is articles “a/an, the” as shown in “TUFS Online Dictionary of Misused English” :

<http://sano.tufs.ac.jp/lcshare/htdocs/?lang=english>

On the other hand, English is a “boundedness-oriented” “high-individualization” type language in terms of having an obligatory grammatical category of number, determiner node, and an obligatory subject/object. The reason why the English-native TOCFL corpus displays an overuse of “ — + Classifier” is because the principle of individualizing a noun is different between English and Chinese. Chinese cannot individualize a noun in an atelic unbounded event like a future event, a potential, a negation, a missed action or a conditional. On the other hand, in English, each noun is itself classified according to its property: countable or uncountable. The principle of individualization in English is not controlled by “Bounded/Unbounded” cognition.

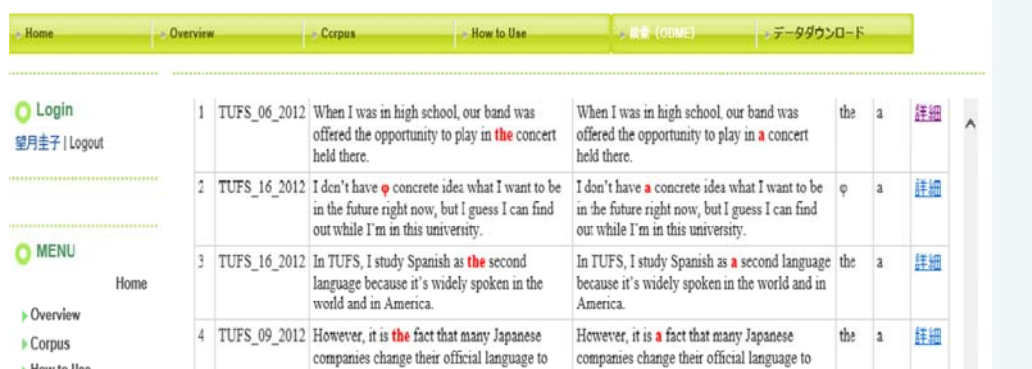
日本語 | English

国際日本研究センター (ICJS) 

オンライン英作文学習者コーパス・誤用辞典

Online Dictionary of Misused English

—— Based on a Learners' Corpus ——



1	TUFS_06_2012	When I was in high school, our band was offered the opportunity to play in the concert held there.	When I was in high school, our band was offered the opportunity to play in a concert held there.	the	a	詳細
2	TUFS_16_2012	I don't have a concrete idea what I want to be in the future right now, but I guess I can find out while I'm in this university.	I don't have a concrete idea what I want to be in the future right now, but I guess I can find out while I'm in this university.	φ	a	詳細
3	TUFS_16_2012	In TUFS, I study Spanish as the second language because it's widely spoken in the world and in America.	In TUFS, I study Spanish as a second language because it's widely spoken in the world and in America.	the	a	詳細
4	TUFS_09_2012	However, it is the fact that many Japanese companies change their official language to ...	However, it is a fact that many Japanese companies change their official language to ...	the	a	詳細

Figure 5. TUFS Online Dictionary of Misused English

4. Conclusion

This paper introduced an empirical study on the difficulties in learning “ — + Classifier(量詞)” in Chinese based on learners' corpora written by English-native learners and Japanese-native learners at CEFR-based A2 and B1 level. The interesting contrast between The TOCFL English-native learner's corpus and The Full Moon Japanese learners' corpus is the overuse and the lack of “ — + Classifier”.

The overuse of “ — + Classifier” in the English-native TOCFL corpus is due to the overgeneralization by English-native learners of Chinese that “a/an NP” is equivalent to “ — + Classifier” NP. On the other hand, the lack of “ — + Classifier” in The Full Moon Japanese learners' corpus is due to the lack of individualization in terms of cognition in Japanese. The different features of the three languages are summarized below:

(28) **Different Features in Number, Classifier and Degree of Individualization**

	1)Number(Singular/Plural)	2) Classifier	3) Degree of Individualization
English	Obligatory	No Classifier	High
Chinese	None except for 我們/這些	Rich system	Moderate “一 + Classifier” occurs in a “bounded” cognition
Japanese	None except for Watashi- <u>tachi</u> (we), kore- <u>ra</u> (these)	Not as rich a system as in Chinese	Low No article No determiner in syntax

This comparative research into cross-linguistic learners' corpora suggests that it is indispensable to explore the pedagogy of Chinese based on learners' native language to develop more efficient and advanced learning science.

References

- Fukui, N. (1995). *Theory of projection in syntax*. Stanford: CSLI Publications, Tokyo: Kuroshio Publishers.
- Huang, C. T. J., Li, Y. H. A., & Simpson, A. (2014) *The Handbook of Chinese Linguistics*, John Wiley & Sons.
- Ikegami, Y. (1981). *Suru (DO)* vs “*Naru (BECOME)*” in the Typology in Language and Culture. Tokyo: Taishukan Publishers. [池上嘉彦 (1981)。「する」と「なる」の言語学：言語と文化のタイポロジーへの試論。東京：大修館書店]
- Ikegami, Y. (2007). *Japanese and Japanese Typology*. Tokyo: Chikuma Publishers. [池上嘉彦 (2007)。日本語と日本語論。東京：筑摩書房]
- Kageyama, T. (1996). *Verb Semantics: the Interface between Language and Cognition*. Tokyo: Kuroshio Publishers. [影山太郎 (1996)。動詞意味論一言語と認知の接点。東京：Kuroshio Publishers]
- Kageyama, T. (2002). *Japanese as a unbounded language*. Tokyo: Iwanami Publishers. [影山太郎 (2002)。けじめのない日本語。東京：岩波書店]
- Li, C. N. & Thompson. S. A. (1981). *Mandarin Chinese: A Functional Reference Grammar*(漢語語法). Taipei:Crane Publishing Co. Lid.
- Mochizuki, K. (2004). *Causative and Inchoative Alternation: Comparative Studies on Verbs in Chinese and Japanese*. Ph.D dissertation, National Tsing Hua University, Taiwan. <http://140.113.39.130/cgi-bin/gs32/hugsweb.cgi/ccd=lJolnl/record?r1=2&h1=0>

- Shen, Y.-M. (2009). *Resultative Compound Verbs in Chinese- From a Viewpoint of Comparative Analyses with Resultative Compound Verbs in Japanese and English Resultative Constructions*. Ph.D dissertation, Tokyo University of Foreign Studies.
<http://repository.tufs.ac.jp/handle/10108/56738>
- Tai, J. H.-Y. (1984). Verbs and Times in Chinese: Vendler's Four Categories. *Lexical Semantics*, 289-296, Chicago Linguistics Society.
- Vendler, Zeno. (1967). *Linguistics in Philosophy*. Cornell University Press.
- 沈家煊 (1995)。“有界”与“无界”。中国语文, 第5期, 367-380。 [Shen, J.-X. (1995). Boundedness and unboundedness. *Chinese Language and Writing*, 5, 367-380]
- 鄧守信 (2009)。對外漢語教學語法修訂二版。台北:文鶴出版有限公司。 [Deng, S.-X. (2009). *A Pedagogical Grammar of Chinese*. Taipei: Crane Publishing Co. Lid.]
- 張斌、齊滬揚等 (2002)。新編現代漢語。復旦大學出版社。
- 張莉萍 (2013)。TOCFL 作文語料庫的建置與應用。載於崔希亮、張寶林 (主編), 第二屆漢語中介語語料庫建設與應用國際學術討論會論文選集 (頁 141-152)。北京: 北京語言大學出版社。
- 張莉萍 (2014)。不同母語背景華語學習者的用詞特徵: 以語料庫為本的研究。中文計算語言學期刊 (*IJCLCLP*), 19(2), 53-72。

The Association for Computational Linguistics and Chinese Language Processing

(new members are welcomed)

Aims :

1. To conduct research in computational linguistics.
2. To promote the utilization and development of computational linguistics.
3. To encourage research in and development of the field of Chinese computational linguistics both domestically and internationally.
4. To maintain contact with international groups who have similar goals and to cultivate academic exchange.

Activities :

1. Holding the Republic of China Computational Linguistics Conference (ROCLING) annually.
2. Facilitating and promoting academic research, seminars, training, discussions, comparative evaluations and other activities related to computational linguistics.
3. Collecting information and materials on recent developments in the field of computational linguistics, domestically and internationally.
4. Publishing pertinent journals, proceedings and newsletters.
5. Setting of the Chinese-language technical terminology and symbols related to computational linguistics.
6. Maintaining contact with international computational linguistics academic organizations.
7. Dealing with various other matters related to the development of computational linguistics.

To Register :

Please send application to:

The Association for Computational Linguistics and Chinese Language Processing
Institute of Information Science, Academia Sinica
128, Sec. 2, Academy Rd., Nankang, Taipei 11529, Taiwan, R.O.C.

payment : Credit cards(please fill in the order form), cheque, or money orders.

Annual Fees :

regular/overseas member : NT\$ 1,000 (US\$50.-)
group membership : NT\$20,000 (US\$1,000.-)
life member : ten times the annual fee for regular/ group/ overseas members

Contact :

Address : The Association for Computational Linguistics and Chinese Language Processing
Institute of Information Science, Academia Sinica
128, Sec. 2, Academy Rd., Nankang, Taipei 11529, Taiwan, R.O.C.

Tel. : 886-2-2788-3799 ext. 1502 Fax : 886-2-2788-1638

E-mail: acclcp@hp.iis.sinica.edu.tw Web Site: <http://www.acclcp.org.tw>

Please address all correspondence to Miss Qi Huang, or Miss Abby Ho

The Association for Computational Linguistics and Chinese Language Processing

Membership Application Form

Member ID# : _____

Name : _____ Date of Birth : _____

Country of Residence : _____ Province/State : _____

Passport No. : _____ Sex: _____

Education(highest degree obtained) : _____

Work Experience : _____

Present Occupation : _____

Address : _____

Email Add : _____

Tel. No : _____ Fax No : _____

Membership Category : Regular Member Life Member

Date : ____/____/____ (Y-M-D)

Applicant's Signature :

Remarks : Please indicated clearly in which membership category you wish to register,
according to the following scale of annual membership dues :

Regular Member : US\$ 50.- (NT\$ 1,000)

Life Member : US\$500.- (NT\$10,000)

Please feel free to make copies of this application for others to use.

Committee Assessment :

中華民國計算語言學學會

宗旨：

- (一) 從事計算語言學之研究
- (二) 推行計算語言學之應用與發展
- (三) 促進國內外中文計算語言學之研究與發展
- (四) 聯繫國際有關組織並推動學術交流

活動項目：

- (一) 定期舉辦中華民國計算語言學學術會議 (Rocling)
- (二) 舉行有關計算語言學之學術研究講習、訓練、討論、觀摩等活動項目
- (三) 收集國內外有關計算語言學知識之圖書及最新發展之資料
- (四) 發行有關之學術刊物，論文集及通訊
- (五) 研定有關計算語言學專用名稱術語及符號
- (六) 與國際計算語言學學術機構聯繫交流
- (七) 其他有關計算語言發展事項

報名方式：

1. 入會申請書：請至本會網頁下載入會申請表，填妥後郵寄或E-mail至本會
2. 繳交會費：劃撥：帳號：19166251，戶名：中華民國計算語言學學會
信用卡：請至本會網頁下載信用卡付款單

年費：

- 終身會員： 10,000.- (US\$ 500.-)
- 個人會員： 1,000.- (US\$ 50.-)
- 學生會員： 500.- (限國內學生)
- 團體會員： 20,000.- (US\$ 1,000.-)

連絡處：

地址：台北市115南港區研究院路二段128號 中研院資訊所(轉)
電話：(02) 2788-3799 ext.1502 傳真：(02) 2788-1638
E-mail：aclclp@hp.iis.sinica.edu.tw 網址：<http://www.aclclp.org.tw>
連絡人：黃琪 小姐、何婉如 小姐

中華民國計算語言學學會

個人會員入會申請書

會員類別	<input type="checkbox"/> 終身 <input type="checkbox"/> 個人 <input type="checkbox"/> 學生	會員編號	(由本會填寫)	
姓名		性別	出生日期	年 月 日
			身分證號碼	
現職		學歷		
通訊地址	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>			
戶籍地址	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>			
電話		E-Mail		
申請人：			(簽章)	
中華民國 年 月 日				

審查結果：

1. 年費：

- 終身會員： 10,000.-
- 個人會員： 1,000.-
- 學生會員： 500.- (限國內學生)
- 團體會員： 20,000.-

2. 連絡處：

地址：台北市南港區研究院路二段128號 中研院資訊所(轉)
 電話：(02) 2788-3799 ext.1502 傳真：(02) 2788-1638
 E-mail：acclp@hp.iis.sinica.edu.tw 網址：<http://www.acclp.org.tw>
 連絡人：黃琪 小姐、何婉如 小姐

3. 本表可自行影印

The Association for Computational Linguistics and Chinese Language Processing (ACLCLP) PAYMENT FORM

Name: _____(Please print) Date: _____

Please debit my credit card as follows: US\$ _____

VISA CARD MASTER CARD JCB CARD Issue Bank: _____

Card No.: _____ - _____ - _____ - _____ Exp. Date: _____(M/Y)

3-digit code: _____ (on the back card, inside the signature area, the last three digits)

CARD HOLDER SIGNATURE: _____

Phone No.: _____ E-mail: _____

Address: _____

PAYMENT FOR

US\$ _____ Computational Linguistics & Chinese Languages Processing (IJCLCLP)

Quantity Wanted: _____

US\$ _____ Journal of Information Science and Engineering (JISE)

Quantity Wanted: _____

US\$ _____ Publications: _____

US\$ _____ Text Corpora: _____

US\$ _____ Speech Corpora: _____

US\$ _____ Others: _____

US\$ _____ Membership Fees Life Membership New Membership Renew

US\$ _____ = Total

Fax 886-2-2788-1638 or Mail this form to:

ACLCLP

% IIS, Academia Sinica

Rm502, No.128, Sec.2, Academia Rd., Nankang, Taipei 115, Taiwan

E-mail: aclclp@hp.iis.sinica.edu.tw

Website: <http://www.aclclp.org.tw>

中華民國計算語言學學會 信用卡付款單

姓名：_____ (請以正楷書寫) 日期：_____

卡別： VISA CARD MASTER CARD JCB CARD 發卡銀行：_____

信用卡號：_____ - _____ - _____ - _____ 有效日期：_____ (m/y)

卡片後三碼：_____ (卡片背面簽名欄上數字後三碼)

持卡人簽名：_____ (簽名方式請與信用卡背面相同)

通訊地址：_____

聯絡電話：_____ E-mail：_____

備註：為順利取得信用卡授權，請提供與發卡銀行相同之聯絡資料。

付款內容及金額：

NT\$ _____ 中文計算語言學期刊(IJCLCLP) _____

NT\$ _____ Journal of Information Science and Engineering (JISE)

NT\$ _____ 中研院詞庫小組技術報告 _____

NT\$ _____ 文字語料庫 _____

NT\$ _____ 語音資料庫 _____

NT\$ _____ 光華雜誌語料庫1976~2010

NT\$ _____ 中文資訊檢索標竿測試集/文件集

NT\$ _____ 會員年費： 續會 新會員 終身會員

NT\$ _____ 其他：_____

NT\$ _____ = 合計

填妥後請傳真至 02-27881638 或郵寄至：

11529台北市南港區研究院路2段128號中研院資訊所(轉)中華民國計算語言學學會 收

E-mail: aclclp@hp.iis.sinica.edu.tw

Website: <http://www.aclclp.org.tw>

Publications of the Association for Computational Linguistics and Chinese Language Processing

	<u>Surface</u>	<u>AIR</u> <u>(US&EURP)</u>	<u>AIR</u> <u>(ASIA)</u>	<u>VOLUME</u>	<u>AMOUNT</u>
1. no.92-01, no. 92-04(合訂本) ICG 中的論旨角色與 A Conceptual Structure for Parsing Mandarin -- Its Frame and General Applications--	US\$ 9	US\$ 19	US\$15	_____	_____
2. no.92-02 V-N 複合名詞討論篇 & 92-03 V-R 複合動詞討論篇	12	21	17	_____	_____
3. no.93-01 新聞語料庫字頻統計表	8	13	11	_____	_____
4. no.93-02 新聞語料庫詞頻統計表	18	30	24	_____	_____
5. no.93-03 新聞常用動詞詞頻與分類	10	15	13	_____	_____
6. no.93-05 中文詞類分析	10	15	13	_____	_____
7. no.93-06 現代漢語中的法相詞	5	10	8	_____	_____
8. no.94-01 中文書面語頻率詞典 (新聞語料詞頻統計)	18	30	24	_____	_____
9. no.94-02 古漢語字頻表	11	16	14	_____	_____
10. no.95-01 注音檢索現代漢語字頻表	8	13	10	_____	_____
11. no.95-02/98-04 中央研究院平衡語料庫的內容與說明	3	8	6	_____	_____
12. no.95-03 訊息為本的格位語法與其剖析方法	3	8	6	_____	_____
13. no.96-01 「搜」文解字—中文詞界研究與資訊用分詞標準	8	13	11	_____	_____
14. no.97-01 古漢語詞頻表 (甲)	19	31	25	_____	_____
15. no.97-02 論語詞頻表	9	14	12	_____	_____
16. no.98-01 詞頻詞典	18	30	26	_____	_____
17. no.98-02 Accumulated Word Frequency in CKIP Corpus	15	25	21	_____	_____
18. no.98-03 自然語言處理及計算語言學相關術語中英對譯表	4	9	7	_____	_____
19. no.02-01 現代漢語口語對話語料庫標註系統說明	8	13	11	_____	_____
20. Computational Linguistics & Chinese Languages Processing (One year) (Back issues of <i>IJCLCLP</i> : US\$ 20 per copy)	---	100	100	_____	_____
21. Readings in Chinese Language Processing	25	25	21	_____	_____
TOTAL				_____	_____

10% member discount: _____ **Total Due:** _____

• **OVERSEAS USE ONLY**

- PAYMENT : Credit Card (Preferred)
 Money Order or Check payable to "The Association for Computation Linguistics and Chinese Language Processing " or “中華民國計算語言學學會”

• E-mail : acclcp@hp.iis.sinica.edu.tw

Name (please print): _____ Signature: _____

Fax: _____ E-mail: _____

Address : _____

中華民國計算語言學學會 相關出版品價格表及訂購單

編號	書目	會員	非會員	冊數	金額
1.	no.92-01, no. 92-04 (合訂本) ICG 中的論旨角色 與 A conceptual Structure for Parsing Mandarin--its Frame and General Applications--	NT\$ 80	NT\$ 100	_____	_____
2.	no.92-02, no. 92-03 (合訂本) V-N 複合名詞討論篇 與 V-R 複合動詞討論篇	120	150	_____	_____
3.	no.93-01 新聞語料庫字頻統計表	120	130	_____	_____
4.	no.93-02 新聞語料庫詞頻統計表	360	400	_____	_____
5.	no.93-03 新聞常用動詞詞頻與分類	180	200	_____	_____
6.	no.93-05 中文詞類分析	185	205	_____	_____
7.	no.93-06 現代漢語中的法相詞	40	50	_____	_____
8.	no.94-01 中文書面語頻率詞典 (新聞語料詞頻統計)	380	450	_____	_____
9.	no.94-02 古漢語字頻表	180	200	_____	_____
10.	no.95-01 注音檢索現代漢語字頻表	75	85	_____	_____
11.	no.95-02/98-04 中央研究院平衡語料庫的內容與說明	75	85	_____	_____
12.	no.95-03 訊息為本的格位語法與其剖析方法	75	80	_____	_____
13.	no.96-01 「搜」文解字—中文詞界研究與資訊用分詞標準	110	120	_____	_____
14.	no.97-01 古漢語詞頻表 (甲)	400	450	_____	_____
15.	no.97-02 論語詞頻表	90	100	_____	_____
16.	no.98-01 詞頻詞典	395	440	_____	_____
17.	no.98-02 Accumulated Word Frequency in CKIP Corpus	340	380	_____	_____
18.	no.98-03 自然語言處理及計算語言學相關術語中英對譯表	90	100	_____	_____
19.	no.02-01 現代漢語口語對話語料庫標註系統說明	75	85	_____	_____
20.	論文集 COLING 2002 紙本	100	200	_____	_____
21.	論文集 COLING 2002 光碟片	300	400	_____	_____
22.	論文集 COLING 2002 Workshop 光碟片	300	400	_____	_____
23.	論文集 ISCSLP 2002 光碟片	300	400	_____	_____
24.	交談系統暨語境分析研討會講義 (中華民國計算語言學學會1997第四季學術活動)	130	150	_____	_____
25.	中文計算語言學期刊 (一年四期) 年份: _____ (過期期刊每本售價500元)	---	2,500	_____	_____
26.	Readings of Chinese Language Processing	675	675	_____	_____
27.	剖析策略與機器翻譯 1990	150	165	_____	_____
		合 計		_____	_____

※ 此價格表僅限國內 (台灣地區) 使用

劃撥帳戶：中華民國計算語言學學會 劃撥帳號：19166251

聯絡電話：(02) 2788-3799 轉1502

聯絡人：黃琪 小姐、何婉如 小姐 E-mail: acclcp@hp.iis.sinica.edu.tw

訂購者：_____ 收據抬頭：_____

地 址：_____

電 話：_____ E-mail: _____

Information for Authors

International Journal of Computational Linguistics and Chinese Language Processing (IJCLCLP) invites submission of original research papers in the area of computational linguistics and speech/text processing of natural language. All papers must be written in English or Chinese. Manuscripts submitted must be previously unpublished and cannot be under consideration elsewhere. Submissions should report significant new research results in computational linguistics, speech and language processing or new system implementation involving significant theoretical and/or technological innovation. The submitted papers are divided into the categories of regular papers, short paper, and survey papers. Regular papers are expected to explore a research topic in full details. Short papers can focus on a smaller research issue. And survey papers should cover emerging research trends and have a tutorial or review nature of sufficiently large interest to the Journal audience. There is no strict length limitation on the regular and survey papers. But it is suggested that the manuscript should not exceed 40 double-spaced A4 pages. In contrast, short papers are restricted to no more than 20 double-spaced A4 pages. All contributions will be anonymously reviewed by at least two reviewers.

Copyright : It is the author's responsibility to obtain written permission from both author and publisher to reproduce material which has appeared in another publication. Copies of this permission must also be enclosed with the manuscript. It is the policy of the CLCLP society to own the copyright to all its publications in order to facilitate the appropriate reuse and sharing of their academic content. A signed copy of the IJCLCLP copyright form, which transfers copyright from the authors (or their employers, if they hold the copyright) to the CLCLP society, will be required before the manuscript can be accepted for publication. The papers published by IJCLCLP will be also accessed online via the IJCLCLP official website and the contracted electronic database services.

Style for Manuscripts: The paper should conform to the following instructions.

- 1. Typescript:** Manuscript should be typed double-spaced on standard A4 (or letter-size) white paper using size of 11 points or larger.
- 2. Title and Author:** The first page of the manuscript should consist of the title, the authors' names and institutional affiliations, the abstract, and the corresponding author's address, telephone and fax numbers, and e-mail address. The title of the paper should use normal capitalization. Capitalize only the first words and such other words as the orthography of the language requires beginning with a capital letter. The author's name should appear below the title.
- 3. Abstracts and keywords:** An informative abstract of not more than 250 words, together with 4 to 6 keywords is required. The abstract should not only indicate the scope of the paper but should also summarize the author's conclusions.
- 4. Headings:** Headings for sections should be numbered in Arabic numerals (i.e. 1.,2,...) and start from the left-hand margin. Headings for subsections should also be numbered in Arabic numerals (i.e. 1.1. 1.2...).
- 5. Footnotes:** The footnote reference number should be kept to a minimum and indicated in the text with superscript numbers. Footnotes may appear at the end of manuscript
- 6. Equations and Mathematical Formulas:** All equations and mathematical formulas should be typewritten or written clearly in ink. Equations should be numbered serially on the right-hand side by Arabic numerals in parentheses.
- 7. References:** All the citations and references should follow the APA format. The basic form for a reference looks like
Authora, A. A., Authorb, B. B., & Authorc, C. C. (Year). Title of article. *Title of Periodical*, volume number(issue number), pages.

Here shows an example.

Scruton, R. (1996). The eclipse of listening. *The New Criterion*, 15(30), 5-13.

The basic form for a citation looks like (Authora, Authorb, and Authorc, Year). Here shows an example. (Scruton, 1996).

Please visit the following websites for details.

- (1) APA Formatting and Style Guide (<http://owl.english.purdue.edu/owl/resource/560/01/>)
- (2) APA Style (<http://www.apastyle.org/>)

No page charges are levied on authors or their institutions.

Final Manuscripts Submission: If a manuscript is accepted for publication, the author will be asked to supply final manuscript in MS Word or PDF files to clp@hp.iis.sinica.edu.tw

Online Submission: <http://www.acclp.org.tw/journal/submit.php>

Please visit the IJCLCLP Web page at <http://www.acclp.org.tw/journal/index.php>

C Contents

Special Issue Articles:

Chinese as a Foreign Language

Guest Editorial : Special Issue on Chinese as a Foreign Language i
Lung-Hao Lee, Liang-Chih Yu, and Li-Ping Chang,
Guest Editors

Papers

HANSpeller: A Unified Framework for Chinese Spelling
Correction..... 1
Jinhua Xiong, Qiao Zhang, Shuiyuan Zhang, Jianpeng Hou, and
Xueqi Cheng

A Study on Chinese Spelling Check Using Confusion Sets and
N-gram Statistics..... 23
Chuan-Jie Lin, and Wei-Cheng Chu

Automatically Detecting Syntactic Errors in Sentences Written by
Learners of Chinese as a Foreign Language..... 49
Tao-Hsing Chang, Yao-Ting Sung, and Jia-Fei Hong

Automatic Classification of the “De” Word Usage for Chinese as a
Foreign Language..... 65
Jui-Feng Yeh, and Chan-Kun Yeh

以「華語學習者語料庫」為本的「了」字句偏誤分析 [The Error
Analysis of “Le” Based on “Chinese Learner Written Corpus”]. 79
董子昀(Tzu-Yun Tung), 陳浩然(Howard Hao-Jan Chen),
楊惠媚(Hui-Mei Yang)

Cross-Linguistic Error Types of Misused Chinese Based on
Learners' Corpora..... 97
Keiko Mochizuki, Hiroshi Sano, Ya-Ming Shen, and Chia-Hou Wu