A Corpus-Based Study on the Influence of Ethnic Language on Zhuang English Learners in China

Ming Yue[*], **Yingzhu Chen**

School of International Studies, Zhejiang University, Hangzhou, China

Email address:
yueming@zju.edu.cn (Ming Yue), chenyingzhu66@163.com (Yingzhu Chen)

Abstract: This study built a rare kind of interlanguage corpus of English compositions from middle school students of Zhuang ethnic group, exploring language transfer on English writing output of the students from their mother-tongue—Zhuang. Another interlanguage corpus is used as the reference corpus with English compositions by Chinese students of non-specific ethnic identity (but assumedly of Han majority) using Chinese as their mother tongue. Through comparison between the self-built Zhuang-Chinese-English learners' corpus and Chinese-English learners' one, this study found that not only are there distinctive differences between the spelling, negation, and auxiliary errors in English writing of Zhuang minority and Han majority students, but also mother-tongue transfer and training transfer have influence on Zhuang students' English output.

Keywords: Interlanguage Corpus, Spelling, Negation, Auxiliary, Transfer

1. Introduction

Language learners' learning environment and how they deal with several languages around them have still remained insufficiently clear to linguistics. The most difficult problems in education today involve issues of language and groups of children who are acquiring or using the societal language while at the same time interacting with family and community members who speak a heritage language (Valdés, 2005, pp. 410-426). Besides, engaging in code-switching for the children living in a bilingual context may differ from those who grow up in monolingual context. In order to explore transfer in the process of second language acquisition, linguists compare language varieties distinguished in terms of the medium (spoken vs. written), field (general vs. specialized), and geographical status (world Englishes) (Granger, 2003, pp. 392-396). Diversification of corpora provides massive and authentic data for empirical study on the linguistic output of different learner communities, leading to a new research methodology called CIA (Contrastive Interlanguage Analysis) (Granger, 1996, pp. 37-51), which involves both the comparison of native language (NL) to interlanguage and the comparisons of different ILs of the same language, assuming that different interlanguage corpora are probably characterized by their unique linguistic features

worthy of exploitation. At the same time, corpus linguistic has witnessed a trend of much more frequency-based and multifactorial statistical methods are on the rise (Gries, 2015, pp. 93-117). Gries aims to elevate corpus-based analysis in SLA/FLA to a new level of precision and predictive accuracy by using corpus-based multifactorial methods and regression modeling, considering interaction effect of variables, answering what and how one or more factors affect the choices of language learners (Gries and Deshors, 2014, pp. 109-136).

The current researches of multilingual interlanguage have investigated various learners with different L1s. However, there is no empirical study, especially employing quantitative approach, focusing on ethnic language learners. Only a few researchers have been undertaking qualitative interlanguage analyses with ethnic languages as mother-tongue. Trilingual groups like Tibetan-Chinese-English, Hmong-Chinese-English, Uygur language-Chinese-English trilingual group draw much attention, reaching a similar conclusion that the acquisition of English is more difficult for trilingual learners than bilingual learners, and that Chinese is kind of inference for their learning English (e.g. Zheng, 2010, pp.; He, 2014, pp. 140-144; Zhang, 2011, pp. 135-138). Fewer researchers have paid attention to negative transfer on learners' English pronunciation of Zhuang language (Liu and Li, 2011, pp. 37-38+54; Luo and Zeng,

2011, pp. 50-53). It can be seen that, the studies of interlanguage of ethnic learners is far legging behind the general interlanguage studies in the world.

Furthermore, specific language outputs, such as spelling, negation, and auxiliary errors of interlanguage corpora have not received adequate exploration. In term of learners' spelling, specific studies focus on the error analysis of this field. The performance of L2 learners in China also arouses interests. Based on the computer-aided error analysis, several findings have been shown in the study of He (He, 2001, pp. 199-205), more than half of the spelling errors of middle school students are related to phonetic errors; consonant errors contribute more than vowel to spelling errors, mainly in final consonants, doubled consonants, consonant cluster; silent letters and weak syllables may also result in spelling errors. This study first provides evidence of the effects of phonetic errors on spelling, and analyses the role of consonants. The spelling errors of regular and dyslexic bilingual Arabic-English students were investigated (Abu-Rabia and Sammour, 2013, pp. 58-68), demonstrating that phonetic errors were more prevalent in Arabic than in English, while semi-phonetic errors were more prevalent in English than in Arabic. In terms of negation and auxiliary, there was a study examining the variable use of the English verb *be* of Hong Kong primary school children, demonstrates that L1 does not play a strong role in transfer, in the use of English be but tends towards developmental aspects of omission and overgeneralization (Lee and Huang, 2004, pp. 211-228). All researches examine Chinese/Arabic as L1s and English as the FL while ignoring the complexity of learners with different mother-dialect/heritage language backgrounds, leaving the error analyses with the interlanguage corpora a bit coarse for their conclusions.

Therefore, this study will stress on one specific ethnic language context, Zhuang, which is the language of Zhuang people, the largest minority ethnicity of China. It is mainly used in Guangxi Zhuang Autonomous Region and southeast part of Yunan province, a member of Tai language of Sino-Tibetan family according to most Chinese linguists, being divided into Southern and Northern dialects (Li and Huang, 2004, pp. 239-256). The Zhuang people have their own language both in oral and written forms, unique culture, living habits, and the ways of thinking. The students grow up in Zhuang-speaking communities and pick up Zhuang as their mother-tongue, learning Chinese in kindergarten or primary school and then starting English courses in middle school, which is different from the most Han students who do not have a multilingual environment. But most Zhuang students have acquired proper Chinese proficiency by the time they enter high school, a situation which allows us to compare them against Han students in term of Chinese transfer on English learning.

By adopting the corpus-based approach, this study intended to explore two written corpora, ZLEC (Zhuang Learner English Corpus) and CLEC (Chinese Learner English Corpus) in two ways: qualitative comparison, observing negative transfer from mother-tongue; quantitative approach, statistical analysis on a variety of learners' corpora. Two research questions were put forward as follows:

- Are there any differences of spelling, negation, auxiliary errors between Zhuang-Chinese-English learners' corpus and Chinese-English learners' one?
- If there are differences, what are some of the probable reasons of above phenomena?

In the first section, this paper introduces the Zhuang language community, previous studies, and proposes the research questions. Section 2 is about the method, describing how the corpus data have been collected, tagged and statistically analyzed. The results and discussion are in Sections 3 and 4 Section 5 moves to the conclusion.

2. Method

2.1. Data

The self-built micro corpus ZLEC (Zhuang Learner English Corpus) collects English written compositions covers 12, 045 words of a Unit Test, which is regularly carried out after finishing teaching and learning of one unit of textbooks. Students participated in the unit test come from two senior high schools, Wuming High School, and Luobo High School, in Wuming County, Guangxi Zhuang Autonomous Region. The title and prompt of the composition is reported in the appendix. After finishing writing the composition, the students were invited to fill out a questionnaire about their language context and their mother-tongue. Then 102 compositions are selected which were written by students in Zhuang as their mother-tongue, and speaking Zhuang at home. Since Chinese is the medium of instruction in the process of learning English for these Zhuang students who usually speak Chinese at school, ZLEC is the interlanguage corpus of trilingual learners (see figure 1).

The reference corpus is a sub-corpus ST2 (senior high school students) of CLEC (Chinese Learner English Corpus) (Gui, 2004, pp. 129-139+216). CLEC is a professional corpus that covers over 1.19 million words of written compositions at different levels of English as Foreign Language learners in China. The ST2 covers 208, 088 words and involves a diversified and varied coverage of written essays from students whose mother-tongue is Chinese. Different from ZLEC, Chinese is the mother-tongue as well as the medium language in the process of learning by the learners in CLEC. That is, CLEC is the interlanguage corpus of bilingual learners (see Figure 1).
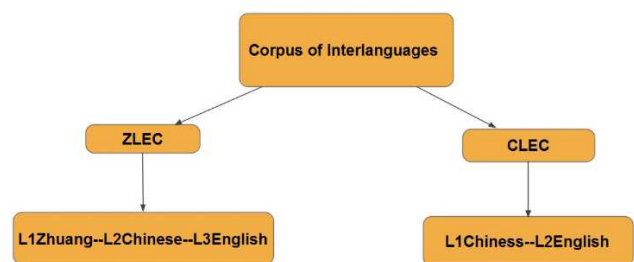

Figure 1. Interlanguage corpora used in this study

2.2. Annotation

In order to ensure accuracy of data retrieval, obvious spelling errors are tagged with the correct forms in clean texts. For instance, the original spelling is 'activteis', (activities) are tagged after the original word. ZLEC follows the tagging categories of CLEC.

CLEC-ST2 is error-tagged, with 11 general types of errors, more than 60 sub-types of errors are constructed by square brackets enclosed letters such as (fml), (vp6), (np4), which represent errors in spelling, tense, case in the whole corpora.

In order to better describe the unique error types in ZLEC, (Za) represents auxiliary errors, including misuse of auxiliary, lacking auxiliary and overuse of auxiliary; and (Zn) represents negation errors, including inversion of word order, misuse of negation auxiliary, etc..

A corpus analysis toolkit for concordancing and text analysis, ANTCONC 3.2.4 (Anthony, 2004, pp. 7 - 13) is applied for data retrieval.

2.3. Statistical Analysis

Error rates are calculated based on the number of the frequency of each type of error, and the occurrence of this kind of phenomenon. Since our study is based on the samples of 102 compositions which are relatively small, we could not for the moment get a variety of linguistic points for analysis. So we focus on three general kinds of errors: spelling, negation and auxiliaries, which were the key foci in the unit test from which we collected our corpus data. For spelling, the occurrence of "help" and "hope" divided by frequency of mixture between letters d-p is spelling error rates of this type. In addition, in negation errors and occurrence of auxiliary verbs in negative sentences divided by the number of wrong choice of auxiliary verbs is the error rates. Finally, in auxiliary errors, the total number of all forms of "be" divided by the number of misuse or lacking of "be" equals the percentage.

3. Results

3.1. Spelling Errors

The retrieval results show the differences in spelling errors between ZLEC and CLEC-ST2. The four typical types of spelling errors are as follows:

Type 1, mixture between letters d-p, for example, 'help' was written as 'held', 'hope' was written as 'hold'. As shown in Table 1, this type of spelling error exists twice in ZLEC, accounting for 8.2% error rates, on the contrary, 0 frequency in CLEC-ST2.

Type 2, mixture between letters k-d, the only error of this type is written 'forwark' as 'forward'. The results presenting error rates are 3.8% and 0%, respectively in ZLEC and CLEC-ST2.

Type 3, mixture between letters k-g is the main errors among the four types. The mixture of 'think' and 'thing' exist in the two corpora. Table 1 summarizes the error frequency 3

plus 6 equals 9, all together occupies 21.2% of error rate. There are 5 errors of this kind shown in CLEC-ST2, however, the error rate is 1.5%, distinctively lower than ZLEC.

Type 4, mixture between letters l-r, the example of this error is 'allow' was written as 'arrow'. Three 'arrow' are retrieved as wrong spelling in ZLEC. Conversely, no 'arrows' are used as 'allow' in CLEC-ST2, and the error rates in the two corpora are 2.6% and 0%.

Table 1. Spelling Errors

Error Type	Corpus	Frequency	Occurrence	Error Rate (%)
d-p	ZLEC	2	244	8.2
	CLEC-ST2	0	390	0
k-d	ZLEC	1	26	3.8
	CLEC-ST2	0	12	0
k-g	ZLEC	3	31	9.7
	CLEC-ST2	3	363	0.8
g-k	ZLEC	6	33	18
	CLEC-ST2	2	281	0.7
l-r	ZLEC	3	113	2.6
	CLEC-ST2	0	6	0

3.2. Negation Errors

After retrieving all the forms of negation, for instance, not, no, am not, is not, isn't, are not, aren't, do not, don't, mustn't, etc. It is found that the errors of negation of ZLEC are evident. Specifically, auxiliary choice and auxiliary verbs order in negative sentences are frequently employed improperly. As shown in the table, there are four types of errors in negative sentences with auxiliary verbs.

Type 1, Wrong choice of auxiliary verbs in negative sentences. The typical example is "they aren't give me personal time". Obviously, don't should be the proper choice rather than aren't in this context. The results show the wrong choice of auxiliary verbs accounting for 1.8% error rates in ZLEC, on the contrary, 0 frequency in CLEC-ST2.

Type 2, Lacking of auxiliary verbs in negative sentences. This type of error also does not be retrieved in CLEC-ST2, but being found in ZLEC, for instance, "I not get along well will my parents". Auxiliary verb "am" is missing in the sentence. The frequency of lacking auxiliary verbs is 17 out of 329.

Type 3, Improper order of auxiliary verbs in negative sentences. In the sentence "I get don't along well with my parents", the order of "get" and "don't" need to reverse. This kind of error is categorized into type 3, improper order. Table 2 summarzing the error frequency is 3, which occupies 0.9% of error rate. There is no error of this kind shown in CLEC-ST2.

Type 4, Redundancy of auxiliary verbs in negative sentences. The example of this error is "they are doesn't allow me to join social activities". It can be seen clearly that "are" is unnecessary in this sentence. 5 error of this kind are retrieved in ZLEC. Conversely, zero frequency in CLEC-ST2.

Table 2. *Negation Errors in ZELC and CLEC*

Error Type	Corpus	Frequency	Occurrence	Error Rate (%)
Wrong choice of auxiliary verbs	ZLEC	6	329	1.8
	CLEC-ST2	0	1960	0
Lacking of auxiliary verbs	ZLEC	17	329	5.2
	CLEC-ST2	0	1960	0
Improper order of auxiliary verbs	ZLEC	3	329	0.9
	CLEC-ST2	0	1960	0
Redundancy of auxiliary verbs	ZLEC	5	329	1.5
	CLEC-ST2	0	1960	0

3.3. Auxiliary Errors

Auxiliary verbs in declarative sentences are retrieved based on the three categories, be, do and have, including misuse, overuse, and underuse of these auxiliary verbs. The auxiliary verbs in interrogative sentences and in negative sentences are not discussed due to few numbers of interrogative sentences, and auxiliary verbs have been calculated in 3.2.

The retrieve results show no "do" and "have" errors in declarative sentences. But "be" is misused in this type of error, for instance, "We are have different opinions…" and "It is let me upset very much…" can be taken as example of this type of error. It can be found in Table 3 that "be" errors appeared 10 times in compositions in ZLEC, occupying 3.7% error rates; however, only 4 out of 8125 in CLEC-ST2, occupying 0.049%.

Table 3. *Auxiliary Errors*

Error Type	Corpus	Frequency	Occurrence	Error Rate (%)
Be	ZLEC	10	270	3.7
	CLEC-ST2	4	8125	0.049

4. Discussion

The results of the present study directly answer our first research question "Are there any differences of spelling, negation, auxiliary errors between Zhuang-Chinese-English learners' corpus and Chinese-English learners's one?" The three types of errors and sub-type errors all show difference with various degrees in ZLEC and CLEC. Generally speaking, the error rates of spelling, negation, auxiliary of ZLEC are much higher than CLEC. It is worth noticing that some sub-types of errors appear zero frequency in CLEC, while appear 1 up to 50 frequency in ZLEC. The variable of errors from Zhuang-Chinese-English learners and Chinese-English ones learners in English writing could be accounted for in terms of L1 transfer, and training transfer. The reasons can respond to our second research question.

4.1. L1 Transfer

The interlanguage spelling, negation, auxiliary in our findings seem to indicate some influence from learners' L1, namely Zhuang and Chinese.

First of all, phonetic differences in Zhuang, Chinese and English play a key role in spelling errors. Certain types of spelling errors in ZLEC occurred as a result of the Zhang consonant system. The findings demonstrate that misspellings represent most of the consonant phonemes of the target words. One possible explanation for the distinctiveness between the frequency and error rates of ZLEC and CLEC-ST2 is that there is no voiceless plosive consonants /p/, /t/, /k/ in Zhuang language (Wei, 2011, pp. 166), especially in Northern Zhuang, whereas both Chinese and English have. It is a barrier for Zhuang students to pronounce /p/, /t/, /k/ correctly, and to distinguish the pairs of consonant /b/ from /p/, /d/ from /t/, and /g/ from /k/. The accuracy and rules will directly affect the spelling and memory of vocabulary, therefore, pronunciation and spelling have close relationship (Zhang, 2011, pp. 228-229). In addition, more misspellings in ZLEC occur as a result of the mixture of /r/ and /l. The omission of voiced consonant /r/ in Zhuang consonant system will contribute to the mixture of /l/ and /r/ (Peng and Huang, 2011, pp. 73-74), while Chinese and English do have /r/. The certain type of spelling errors may occur in those students with low mastery of the target language. A reflection of the negative effect of lacking certain consonants can be found in the result showing frequency of three types of misspellings and the gap of nearly 20 percentages in error rate of the substitution between letters k-d.

Secondly, negation pattern reflects the grammar feature of the mother-tongue. The errors in negative sentence in ZLEC and CLEC present huge difference, indicating far lower proficiency of negation pattern among Zhuang students. In Chinese and English, the position of negation adverbs is in front of predicate verbs. However, there are three sequences of negation adverbs in Zhuang, V+Neg, Neg1+V+Neg2, and Neg+V (Li and Wu, 2008, pp. 37-39), which represent three stages of the development of negation pattern of this ethnic language. That is to say, negation adverbs can be put in advance, or after predicate verbs, or double negation adverbs with one in advance and the other after predicate verbs. This structure of Zhuang, different from Chinese and English, confuses Zhuang-Chinese-English learners when they try to express negative patterns.

Thirdly, the statistics of table 3 demonstrate the usage of auxiliary that might be a difficult point for Zhuang-Chinese-English learners rather than Chinese-English ones. Differ from Zhuang and Chinese, the grammatical relationship among words in English depends on morphological changes, particularly, conveying verbs tense, voice and aspect with auxiliary and modal verbs. Nevertheless, the grammatical relationship among words in Zhuang and Chinese depends not on morphological changes, but on notional words, functional words or particle, and syntactic methods (Ruan, 2010, pp. 446-447). The point is, both of Zhuang and Chinese using similar way of expressing grammatical changes but produce different output of

auxiliary error rates, which would be explained in next reason.

4.2. Training Transfer

In considering how these errors may be accounted for, input factors in students' learning environment related to students' daily access to the written English language and teachers' instruction probably affect their output. Since ZLEC is produced by Unit Test of Zhuang middle school students, they follow textbooks and teachers' requirements in this study with no exception. A further search into the children's school textbooks led us to discover that the children's learning environment and conditions for learning (textbook input and transfer of training) could play an equally significant role. Another research specialized in teachers' knowledge and decision of three verbs with the results that students' expression in English writing appear to be a connection with teachers' knowledge structure (Yang, 2006, pp. 110-115). Besides, students tend to follow task prompts without creating, and even simply translating the main idea of the prompts in this type of time-limited practice. Therefore, some Chinese structures will directly influence students' English writing. Furthermore, unbalanced education level, particularly faculty resource in different areas in China has long been criticized, leading to great teachers rush into first and second-tier cities, and lack of qualified teachers in remote rural areas. Wuming County is this kind of remote area where students start learning English from middle school, while most students in other parts of China start their English course from primary school and ever some of them from kindergarten. The fact that having less learning time and learning later might explain the data of high error rates of ZLEC than of CLEC.

5. Conclusion

Based on the comparative study between the learners' corpora of minority and majority ethnic groups, we find that the acquisition/ learning of English spelling, negation and auxiliary is different between Zhuang-Chinese-English learners and Chinese-English ones with a gap in error rates of the three types. L1 transfer, including phonetic rules and grammatical rules of three languages, might be the main factor contributed to higher error rates form ZLEC compared with CLEC, and the training transfer should not be neglected due to students' exposure to textbook, teacher's guidance and task prompts.

Although our interpretation on the sample of 102 compositions by Zhuang learners and on their self-reported mother-tongue proficiency should be shown with caution, this is to the best of our knowledge, the first serious study using learners' corpus to explore how children are acquiring the English language while at the same time interacting with family speaking the Zhuang heritage language and with community members speaking the Chinese as their second language.

For future work, more data needed to be collected from the Zhuang minority learners. And the samples of Han students with Chinese in the same class with the Zhuang students would be gathered to build a closer reference corpus with little background differences other than their mother tongues. A significance test would also be applied to the data analysis and also to the questionnaire of mother-tongue proficiency for sounder statistical conclusions.

Appendix

Title: A Letter to Psychologist
Pretend you were Lihua. You are upset due to the relationship with your parents. You are required to write a letter for help.
1. Express your intention;
2. Argue with your parents very often;
3. No personal space, no free choice
4. Want to get help
(The above title and prompt are given to the students in Chinese)

References

[1] Valdés, G. Bilingualism, heritage language learners, and SLA research: Opportunities lost or seized? *The Modern Language Journal*, vol. 89.3, 2005. pp. 410-426.

[2] Granger, S. The International Corpus of Learner English: A New Resource for Foreign Language Learning and Teaching and Second Language Acquisition Research. *TESOL Quarterly*, vol. 24.2, 2003. pp. 392-396.

[3] Granger, S.. "From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. 37-51, 1996.

[4] Gries, S. T. Some Current Quantitative Problems in Corpus Linguistics and a Sketch of Some Solutions. *Language and Linguistics*, vol. 16.1, 2015. pp. 93-117.

[5] Gries, S. T. & Deshors, S. C. Using regressions to explore deviations between corpus data and a standard/target: two suggestions. *Corpora*, vol. 9.1, 2014. pp. 109-136.

[6] Zheng, L. The Development of Metalinguistic Awareness of Hmong Learners in Third Language Acqusition. Doctoral Dissertation, Southwest University, 2010.

[7] He, X. On the Transfer in Third L anguage Acquisition: A Case Study of Tibetan Undergraduates' English Vocabulary Learning in Sichuan Province. *Foreign Language and Literature*, vol.01, 2014. pp. 140-144.

[8] Zhang, C. Third Language Acquisiton and the Implication for English Teaching of Mongolian Middle School Students. *Chinese Journal (Foreign Lauguage Teaching and Education)*, vol.11, 2011. pp. 135-138.

[9] Liu, M. & Li, Y. A Study of the Influence of Mother-tongue for English Pronunciation of Zhuang College Students. *Guangxi Education*, vol.21, 2011. pp. 37-38+54.

[10] Luo, H. & Zeng, J. An Analysis of the Negative Transfer of Mother Tongue on English Pronunciation of Chuang College Students. *English Square*, vol.Z4, 2011. pp. 50-53.

[11] He, A. Spelling Error Analysis on Students' Writing. *Foreign Language Teaching and Research*, vol.03, 2001. pp. 199-205.

[12] Abu-Rabia, S. & Sammour, R. Spelling Errors' Analysis of Regular and Dyslexic Bilingual Arabic-English Students. *Open Journal of Modern Linguistics*, vol. 03.01, 2013. pp. 58-68.

[13] Lee, N. & Huang, Y. Y. To Be or Not To Be—The Variable Use of The Verb Be in The Interlanguage of Hong Kong Chinese Children. *Regional Language Centre Journal*, vol. 2.35, 2004. pp. 211-228.

[14] Li, X. & Huang, Q.. The introduction and development of the Zhuang writing system. *Language Policy in the People' s Republic of China.*: Springer. 239-256, 2004.

[15] Gui, S. A Cognitive Model of Corpus-based Analysis of Chinese Learners' Errors of English. *Modern Foreign Languages*, vol.02, 2004. pp. 129-139+216.

[16] Anthony, L. AntConc: A Learner and Classroom Friendly, Multi-Platform Corpus Analysis Toolkit. *An Interactive Workshop on Language e-Learning*. (2004). pp. 7 - 13.

[17] Wei, M. Analysis on Negative Transfer from Mother-tongue Zhuang to English Learning. *New Western*, vol.10, 2011. pp. 166.

[18] Zhang, B. The Effects of Phonological Awareness in Word Spelling. *Overseas English*, vol.02, 2011. pp. 228-229.

[19] Peng, Q. & Huang, Q. Contractive Analysis on the Zhuang Dialect and English Pronunciation and the Countermeasures. *Journal of Jixi University*, vol.12, 2011. pp. 73-74.

[20] Li, J. & Wu, Y. A Study of Words Order in Negation in Kam-Tai Languages. *Minority Languages of China*, vol.02, 2008. pp. 37-39.

[21] Ruan, W. A Comparison of Passive Voice in English, Chinese and Zhuang. *Scientific and Technological Information*, vol.34, 2010. pp. 446-447.

[22] Yang, C. Non- native Uses of Verb Constructions in the Interlanguage of College English Learners in China. *Journal of Hunan University (Social Sciences)*, vol.03, 2006. pp. 110-115.