A Corpus Linguistic Perspective on the Appropriateness of Pop Songs for Teaching Chinese as a Second Language

Xiangyu Chi Gaoqi Rao
Beijing Language and Culture University
cxyblcu@163.com， raogaoqi@blcu.edu.cn

Abstract

Language and music are closely related. Regarding the linguistic feature richness, pop songs are probably suitable to be used as extracurricular materials in language teaching. In order to prove this point, this paper presents the Contemporary Chinese Pop Lyrics (CCPL) corpus. Based on that, we investigated and evaluated the appropriateness of pop songs for Teaching Chinese as a Second Language (TCSL) with the assistance of Natural Language Processing methods from the perspective of Chinese character coverage, lexical coverage and the addressed topic similarity. Some suggestions in Chinese teaching with the aid of pop lyrics are provided.

1 Introduction

The appropriateness of pop songs for language teaching has been widely discussed by scholars in recent years. At present, both quantitative and qualitative studies on the songs' application in the English teaching have developed maturely. Meanwhile, quantitative studies have not attracted enough attention in the Chinese teaching. Related studies prefer to explore the songs' potential in teaching assistance in the dimensions of lyrics' linguistic features, teachers' practical experience and students' knowledge. Therefore, this paper focuses on the similarity between lyrics and teaching material for TCSL. The rest of the paper is organized as follows: in section 2 the related work is surveyed, section 3 and 4 describe the CCPL corpus in detail, section 5 and 6 analyze the statistics of character coverage, lexical coverage and topic similarity of the CCPL corpus in TCSL. Conclusions and future works are finally drawn in Section 7.

2 Related Work

In the field of the English teaching, Plitsch (1997) raised the idea of using contemporary lyrics for more inspiring and close-to-reality language teaching. Werner (2012) compared the role of American English and British English in popular songs, and Werner (2018) outlined its didactic potential. Tegge (2017) examined the lexical coverage of pop lyrics in English teaching, using around 1,000 songs in two collections. Applications of Natural Language Processing (NLP) tools could also be found in related work: Penaranda (2006) uses text mining for empirically based genre assignments, involving linguistic anomalies. Napier/Shamir (2018) took a diachronic perspective and quantify emotional changes in lyrics since 1950.

In the field of the Chinese teaching, Shouhui Zhao and Qingsong Luo (1994) were the first to propose the introduction of songs into classes. They believed songs could create a good cultural atmosphere for Chinese as foreign language (CFL) learners, resulting in their high interest in studying. You Fu (2002) argued that songs could help solve the difficulties of traditional listening lessons from the perspective of psychology. Chenxu Yang (2019) affirms that music plays a positive role in enhancing CFL learners' sense of language, strengthening the ability of correcting errors and improving efficiency. Yanjing Wang (2011) puts forward that songs can help CFL learners to correct pronunciation, expand vocabulary, understand rhetoric, recognize proverbs, etc.

Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications, pages 129–137
Suzhou, China, December 4, 2020.

Based on the above, this paper found that previous studies in TCSL seldom used quantitative methods or NLP tools to extract and analyze language information contained in lyrics. Therefore, we attempt to improve the subordinate role which quantitative searches play in the appropriateness of pop songs for TCSL.

3 Corpus Building

The CCPL corpus contains lyrics in Chinese, even if a minority are in English. Given that this paper focus on the coverage of lyrics' content in TCSL, English words are excluded for noise reduction. Enabling further processing with NLP tools, the original lyrics need to be transferred into annotated contents. Therefore, word segmentation and part of speech (POS) tagging are needed. Words extracted by topic models are also presented for comparison.

3.1 Data Selection

A total of 1,110 pop songs were contained in the CCPL corpus, which were collected from authoritative competitions and lists such as CCTV Spring Festival Gala Evening, Top Chinese Music Awards, etc. The metrics for the selection are as follows: Firstly, pop songs are of high popularity among the mass media and reflect the social changes and the zeigeist. Secondly, the songs in the list tended to be the songs of the year. They were selected by the mass and experts voting among the songs of the year, and the organizer is officially authoritative and has certain credibility.

3.2 Data Annotation

We implemented *jieba*[1], a built-in module in python, to realize the segmentation and POS tagging of the lyrics. Fully automatic annotation of such units produces rather poor results, therefore we corrected the results with the assistant of bi-gram model.

In the process of manual processing, we took The Basic Processing of Contemporary Chinese Corpus at Peking University (PU) and *jieba* POS labeling specification as standard. Several fuzzy phenomena of in word segmentation are briefly introduced as follows.

Numerals and quantities. *Jieba* counts quantitative phrases as a segmentation unit, such as "一个". In this paper, they are divided into numerals and quantities, namely "一/m 个/q".

Separable words. Separable words are regarded as three segmentation units in the PU. This paper chooses to combine them, resulting in expressing concepts more completely and being directly labeled as verbs, such as "舍不得/v".

Overlapping words. The quantitative structure in the form of "ABB" is segmented, namely "一/m 颗/q 颗/q". This is to enable the vocabularies in CCPL to cover as many vocabularies in HSK as possible. Meanwhile, the adjective overlapping form "ABB" and the two-syllable verb and adjective overlapping form "AABB" are not segmented, such as "甜甜蜜蜜/a", "亮堂堂/a".

Morpheme + "们" or word + "们", which indicates the plurals of nouns should be segmented, such as "孩子/n 们/k", while words such as "我们" express a concept independently, we label them as "我们/r".

Monosyllabic verb +directional verb. *Jieba* processes this form into a segmentation unit such as "爱上". We process them into "爱/v 上/vf". This is also intended to cover as many vocabularies in HSK as possible

3.3 Topic Description

Each song in the CCPL corpus was provided with the top ten words extracted by two weighting schemes respectively, namely Textrank and TF-IDF. We took the words as reference to conclude the content of topics. For instance, the topic *Travel* contains words of both the field of *specialty* ("煎饼", "大葱") and the field of *sights* ("泰山", "黄河").The topic *Family* contains words of both the field of *family members* ("老爸", "老妈", "宝贝") as well as the words of *daily life* ("拼命", "风雨无阻", "跑调").

Figure1(a): Topic Family. Figure1(6): Topic Travel.

[1] https://github.com/fxsjy/jieba

Figure2(a): Topic Food.

Figure2(b): Topic Peking Opera.

4 Corpus Overview

The key statistics on the corpus are summarized in this section, such as word length and POS distribution, and its accessibility is also provided.

4.1 Word length

The CCPL corpus comprises 292,609 tokens and 3,320 types of characters together with 208,537 tokens and 12,231 types of words. Figure 3 shows the distribution of word length in detail. Among the 12,231 types of words, disyllabic words occupy the largest ratio, including 8,598 types such as “我们”, which appear 71473 times in total. Monosyllabic words such as “的” come next with 2077 types, which appear 131949 times in total. The longest word type is a six-syllable word, such as “一步一个脚印”.

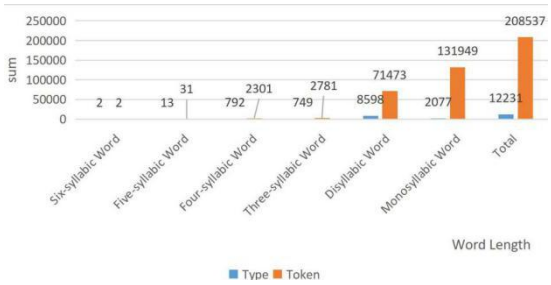

Figure 3: Distribution of word length.

4.2 Parts of speech

Table 1 illustrates the distribution of word length in detail. The CCPL corpus contains a total of 33 POSes. The verbs take vast majority (23.56%), followed by noun (18.09%). On the bottom, prefix amount to a rather low ratio, nearly zero. In addition, there are a considerable number of idioms, phrases, onomatopoeias and so on in the corpus, which reflect the feature of both spoken and written discourse in lyrics.

POS	Token	Ratio	POS	Token	Ratio	POS	Token	Ratio
(verb) v	49133	23.56%	(conjunction)c	2790	1.34%	(phrase)l	255	0.12%
(noun) n	37724	18.09%	(directional verb) vf	2648	1.27%	(other proper noun) nz	165	0.08%
(pronoun) r	23845	11.43%	(exclamation)e	2591	1.24%	(space)s	132	0.06%
(auxiliary) u	20501	9.83%	(modal partical) y	2396	1.15%	(descriptive word) z	129	0.06%
(adjective) a	16870	8.09%	(idiom)i	1789	0.86%	(abbreviation)j	57	0.03%
(adverb) d	16787	8.05%	(noun of place) ns	1489	0.71%	(verb & noun) vn	55	0.03%
(prepositional)p	7829	3.75%	(substantival morpheme) ng	856	0.41%	(adjective & adverb) ad	30	0.01%
(numeral) m	5414	2.60%	(onomatopoeia)o	813	0.39%	(adjective & noun) an	30	0.01%
(time) t	4971	2.38%	(suffix)k	700	0.34%	(verbal morpheme) vg	9	0.00%
(quantity) q	4004	1.92%	(noun of personal name) nr	441	0.21%	(noun of time) nt	7	0.00%
(noun of locality) f	3797	1.82%	(distinguishing words) b	278	0.13%	(prefix)h	2	0.00%

Table 1: Distribution of POS.

4.3 Time and Source Description

Figure 4 illustrates the number of songs collected in the CCPL corpus by the year of publication. It can be inferred that the songs in corpus spans the period from 1990 to 2020.

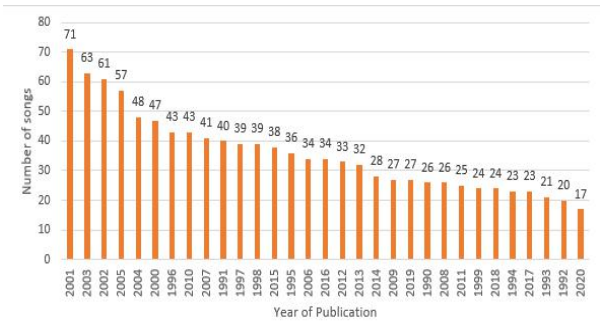

Figure 4: Year of publication distribution.

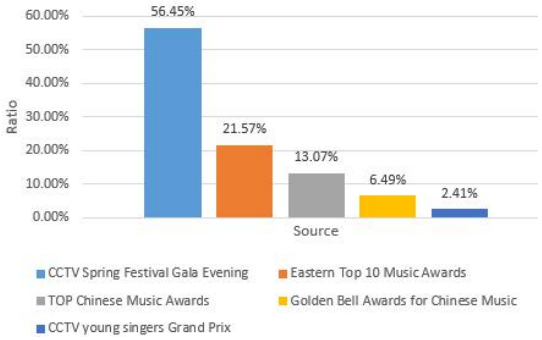

Figure 5: Source distribution.

Figure 5 shows the number of songs published in our corpus by source. We find that over 50% of all songs in the CCPL corpus are from the CCTV Spring Festival Gala Evening and only around 2 are from CCTV young singers Grand Prix.

The CCPL corpus could be downloaded online. We provide the corpus in eXtensible Markup (xml) file formats.[2]

5 Character and Word Analysis

HSK is a standard international Chinese language proficiency test for CFL learners. As an international and authoritative test, it is suitable to be used as a golden standard to measure whether lyrics are appropriate for assisting TCSL. There are 6 levels in the new Syllabus of HSK. CFL learners who reach HSK6 should be competent to master 2,500 Chinese characters and 5,000 words in total.

In this section, repetitive types of characters and words were removed. Character frequency and word frequency were computed to respectively generate the character list and word list for the comparison between the CCPL and HSK in character and word.

5.1 Characters in HSK

Table 2 indicates the character type coverage of the CCPL corpus in HSK. The ratio decreases step by step. The coverage rate at HSK1 is the highest, reaching 100.0% meanwhile that at HSK6 level is the lowest, reaching 84.3%. However, the shared types cover 91.9% of the total 2500 Chinese characters in HSK, indicating that the overall situation of character type coverage is optimistic. Even so, there are 1022 new characters beyond the scope of HSK, among which contains rare characters such as “魁”, “曹”, “桀”, “鸄” that are hard to recognize.

Table 3 describes the character type coverage of HSK in CCPL. It worth to note that types in HSK1 and HSK2 contributes the lowest coverage rate (4.5%) while HSK6 shows the most optimistic situation with the coverage rate of 25.4%. That reveals preliminarily that the presence of new characters is difficult for CFL learners of elementary level (HSK1 and HSK2) in understanding the lyrics content, and when CFL learners reach intermediate level (HSK3 and HSK4) even advanced level (HSK5 and HSK6), the situation will get improved with the superior ability of mastering new characters. It can be inferred that the present of new characters for learners in intermediate level or advanced level may be a good approach to improve the reservation of Chinese characters.

Table 4 describes the character token coverage of HSK in the CCPL corpus. Although the character type coverage of HSK in corpus performs unsatisfying, it can be seen in Table 4 that the 2,298 shared types constitute 96.9% of the content in corpus, in other words, the 1,022 new characters beyond the scope of HSK account for only 3.1% in the corpus. It can also be seen that tokens of HSK 1 and HSK 2 contribute the largest coverage rate, which shows simple characters tend to appear in elementary level. CFL learners at that stage can review the characters they have grasped.

[2] URL: https://pan.baidu.com/s/1txjD9hx2ZlnGYgjDaGxslQ
Password: ykhu

	HSK1	HSK2	HSK3	HSK4	HSK5	HSK6	HSK 1-6
Shared Types	150	150	298	391	466	843	2298
Types in HSK	150	150	300	400	500	1000	2500
Coverage Rate (Shared Types/HSK)	100.0%	100.0%	99.3%	97.8%	93.2%	84.3%	91.9%

Table2: Shared types in HSK.

	HSK1	HSK2	HSK3	HSK4	HSK5	HSK6	HSK1-6
Shared Types	150	150	298	391	466	843	2298
Types in CCPL	3320	3320	3320	3320	3320	3320	3320
Coverage Rate (Shared Types/ the CCPL corpus)	4.5%	4.5%	9.0%	11.8%	14.0%	25.4%	69.2%

Table 3: Shared types in CCPL.

	HSK1	HSK2	HSK3	HSK4	HSK5	HSK6	HSK1-6
Shared Tokens	118080	47009	48958	31910	21024	16489	283470
Tokens in in CCPL	292609	292609	292609	292609	292609	292609	292609
Coverage Rate (Shared Tokens/ the CCPL corpus)	40.4%	16.1%	16.7%	10.9%	7.2%	5.6%	96.9%

Table 4: Shared tokens in CCPL.

5.2 Words in HSK

Table 5 illustrates that the word type coverage in the CCPL corpus in HSK is decreasing step by step, among which the coverage rate in HSK 1 is the highest (93.3%) while that in HSK 6 is the lowest (51.6%). However, the lexical coverage of the CCPL corpus in HSK is not optimistic with the coverage rate of 51.6%. Besides, according to Table 6, the word types of HSK only cover 20.6% in the CCPL corpus, which infers that there is a great amount of words beyond the scope of HSK. Indeed, idioms, proverbs and other kind of phrases can be found in the CCPL corpus, such as “种瓜得瓜, 种豆得豆”, “说闲话”, “鞠躬尽瘁”, etc. Beyond doubt, it is a challenger for CFL learners to master those obscure words without cultural background and language environment.

In table 7, the 2,514 shared word types constitute 67.1% of the content in CCPL in total. It can be inferred that the words beyond the scope of HSK occupies a great part ratio in CCPL, resulting in obstacles in language teaching and learning.

	HSK1	HSK2	HSK3	HSK4	HSK5	HSK6	HSK 1-6
Shared Types	140	137	248	417	707	938	2580
Types in HSK	150	150	300	600	1300	2500	5000
Coverage Rate (Shared Types/HSK)	93.3%	91.3%	82.7%	69.5%	54.4%	37.5%	51.6%

Table 5: Shared types in HSK.

	HSK1	HSK2	HSK3	HSK4	HSK5	HSK6	HSK 1-6
Shared Types	136	137	246	406	692	904	2514
Shared Types in CCPL	12231	12231	12231	12231	12231	12231	12231

Coverage Rate (Shared Types/ the CCPL corpus)	1.1%	1.1%	2.0%	3.3%	5.7%	7.4%	20.6%

Table 6: Shared types in CCPL.

	HSK1	HSK2	HSK3	HSK4	HSK5	HSK6	HSK1-6
Shared Tokens	71182	21138	15763	15451	11279	8242	139986
Tokens in CCPL	208537	208537	208537	208537	208537	208537	208537
Coverage Rate (Shared Tokens/ the CCPL corpus)	34.1%	10.1%	7.6%	7.4%	5.4%	4.0%	67.1%

Table 7: Shared tokens in CCPL.

6 Topic Analysis

In this section, for the comparison between lyrics and HSK in topic level, we take the *New General Syllabus for Teaching Chinese* as reference to summarize types of topics discussed in teaching. Meanwhile, we will introduce two weighting schemes with the function of extracting main topics conveyed in lyrics, and provide a more refined classification for the types of lyrics. On the bottom, we conduct a comparative analysis.

6.1 Topics in Textbooks

The teaching topics discussed in practice are diverse and complex, which are difficult to be collected. Zhang Hangli (2019) introduced the topics required in the *New General Syllabus for Teaching Chinese*. Based on that, Zhang classified the topics in *HSK Standard Course*, a textbook with high authority on the market, from the perspectives of topic type and topic content. We took Zhang's study as a standard and divided the passages in 9 copies[3] of *HSK Standard Course* into 16 types of topics, such as *Personal Information*, *Family Life*, *Values*, etc. For instance, *Sunshine always comes after storm* in *HSK Standard Course 4 (I)* and *Reading and Thinking* in *HSK Standard Course 5 (II)* respectively convey the view about success and the attitude towards learning, both of which talk about people's subjective views on specific events. They are classified into *Values* due to the contents we summarize.

6.2 Topics in Lyrics

Removed stop words, we used TF-IDF and Textrank algorithms to uncover the topic addressed in lyrics and divided them into types by the same standard of topics classification in HSK.

6.2.1 TF-IDF algorithm

TF is the abbreviation of Term Frequency, and IDF is the abbreviation of Inverse Document Frequency. The more times a word appears in a particular document and the less it appears in other documents, the more it can reflect the content of the document and the more likely it is to become a key word. Its TF-IDF value is also higher. The formula is shown as follow:

$$TF - IDF = TF * IDF = TFij * log(\frac{N}{Ni})$$

TF_{ij} is the frequency of the term i in document j, N is the total number of documents, and Ni is the number of documents that contain the term i. For instance, top 10 keywords of the song *XiaoFang* uncovered by TF-IDF are "小河", "善良", "谢谢", etc.

6.2.2 Textrank algorithm

The idea of Textrank algorithm is derived from the Pagerank algorithm of Google. Textrank is used to extract keywords, which can be explained by PageRank idea: If a word appears after many words, it means that the word is relatively important. Meanwhile, If a word with a high Textrank value is followed by a word, the

[3] Each level of HSK1-3 is one copy, and each level of HSK4-6 is divided into two copies.

Textrank value of that word is raised accordingly. The formula is shown as follow:

$$S(Vi) = (1-d) + d * \sum(j,i) \frac{Wji}{\sum vk \in out(Vj) Wjk} S(Vj)$$

It can be described as: the weight of a word I in Textrank depends on the weight of the edge (j, i), which is composed of each point j before i, and the sum of the weights of the point j to the other edges. Top 10 keywords of the song *Welcome to Beijing* extracted by Textrank are "欢迎", "北京", "打开", etc.

6.2.3 The Shared Topics in Lyrics and HSK

According to keywords outputted by TF-IDF and Textrank, we removed the repeated units to get a cluster of keywords for each song. Words in clusters have a certain semantic correlation, which could be regarded as the semantic representation of topics. For instance, the word cluster of the song *Ice-sugar gourd* contains 14 words. Several words describe the appearance, such as "竹签", "好看". Some words tell the meaning of this snack, such as "幸福", "团圆". Other words indicate its function, such as "治病". Given that the passage *Radish Cake in hometown* in *HSK Standard Course 5 (II)* discusses about a kind of delicious food and is classified into the type of *Language and Culture*, the topic of *Ice-sugar gourd* is summarized as *Language and Culture* as well. Figure 6 shows the topic distribution in the CCPL corpus.

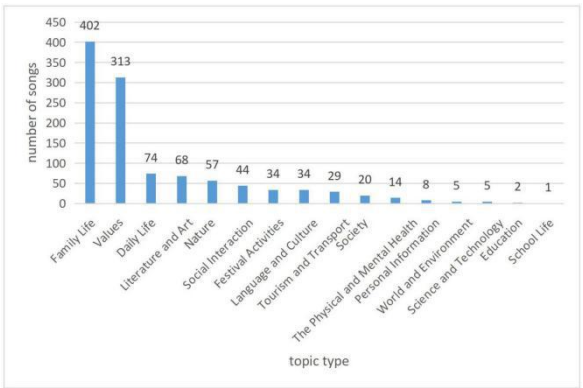

Figure 6: Topic types in the CCPL corpus.

According to Figure 6, *Family Life* takes the vast majority (402), followed by Values (313) and by 10 types of topics (range from 8 to 74). Love songs account for the largest proportion in the CCPL corpus. That is the reason why the topic *Family Life* contributes the largest number. It should be noted that the topic distribution in the corpus is sparse. *Personal Information*, *Science and Technology*, *Education* and *School life* contributes under 5 songs each.

Figure 7, Figure 8, Figure 9 suggests the topic distribution in HSK.

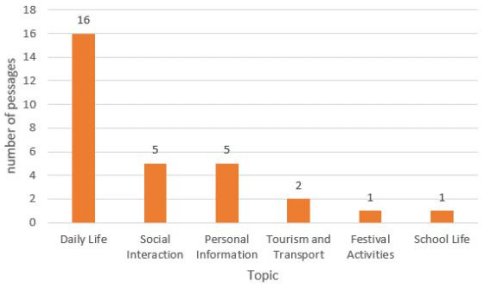

Figure 7: Topic distribution in *HSK Standard Course 1&2.*

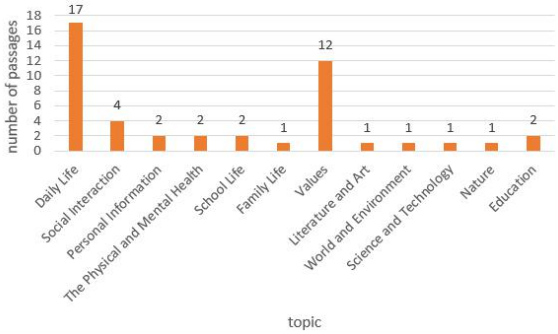

Figure 8: Topic distribution in *HSK Standard Course 3&4.*

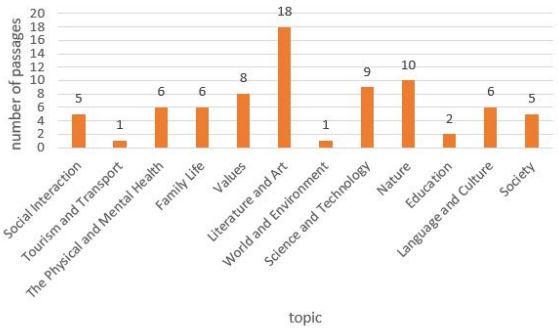

Figure 9: Topic distribution in *HSK Standard Course 5&6.*

As is shown in Figure 7, topics in elementary level tend to be simple and practical, such as *Daily Life*, *Social information* and *Personal information*. They are related to almost all aspects of daily life and include specific interaction environment. And songs classified into these

topics account for a certain number in the CCPL corpus, except *School Life*.

It can be noted in Figure 8 that topics in intermediate level tend to be diverse and difficult, among which begin to involve the outlook and attitude towards health, culture and philosophy, such as *The Physical and Mental Health*, *Values* , *Literature and Art*. Even so, the proportion of *Daily Life* still ranks first. As we can observe with the combination of Figure 6 and Figure 8, teachers have multiple choices when selecting topic-related songs. Meanwhile, they still get limited if the passages in HSK pay attention to *School Life* and *Education*.

According to Figure 9, topics discussed above get involved in advanced level. Different from the levels above, difficult topics account for furtherly more proportion in HSK 5 and HSK6, among which *Literature and Art* ranks first. This also confirms that relevant songs in the CCPL corpus could help broaden students' horizon and cultivate cultural atmosphere.

7 Conclusion and Future Work

In this paper we introduced the CCPL, a homogeneous corpus comprising of 1,110 pop songs. The paper offers the experimental results demonstrating the extent to which pop songs could be appropriate to TCSL teaching. According to the characters coverage, lyrics are suitable for CFL students in elementary level to review the characters they have grasped. As for students in intermediate level and advanced level, they can not only consolidate what they have learned, but also increase their reservation of characters. Lexical coverage is not optimistic. Words beyond the scope of HSK bring barriers in both teaching and learning. Besides, topic coverage confirms the strength of lyrics in introducing cultural background and create multiple language environment, even if teachers may encounter small obstacles in songs selection.

For future work, we intend to give a more refine comparation between lyrics and HSK in character and lexical coverage. The shared and rare characters and words in each song will be calculated. In addition, we will score the appropriateness of each song for language teaching with designing weight formula which includes the factors such as definition, figure of speech, sentence repetition, etc. With relevant information, we can carry out the songs' recommendation for CFL learners at all levels.

8 Acknowledgements

We thank Gaoqi Rao for contributing a lot in consultation and bidding and Mengyao Suo in data selection.

This study was supported by the projects from National Language Committee Project (YB135-90) and BLCU supported project for young researchers program (supported by the Fundamental Research Funds for the Central Universities) (20YCX150).

References

Chenxu Yang. 2019. The Implement of Chinese Music in Teaching Chinese as a Foreign Language. *Today's Massmedia*, 27(03), pages 148-151.

Hangli Zhang. 2019. *A Study on the Distribution of Topic and Content of International Curriculum for Chinese Language Education in HSK Standard Course*, Master's Themes. Sichuan International Studies University, Chongqing, China.

Liping Jiang. 2014. *HSK Standard Course 1*. Beijing Language and Culture Press, Beijing, China.

Liping Jiang. 2014. *HSK Standard Course 2*. Beijing Language and Culture Press, Beijing, China.

Liping Jiang. 2014. *HSK Standard Course 3*. Beijing Language and Culture Press, Beijing, China.

Liping Jiang. 2014. *HSK Standard Course 4(I)(II)*. Beijing Language and Culture Press, Beijing, China.

Liping Jiang. 2015. *HSK Standard Course 5(I)(II)*. Beijing Language and Culture Press, Beijing, China.

Liping Jiang. 2015. *HSK Standard Course 6(I)*. Beijing Language and Culture Press, Beijing, China.

Liping Jiang. 2016. *HSK Standard Course 6 (II)*. Beijing Language and Culture Press, Beijing, China.

Napier, K. and Shamir, L., 2018. Quantitative Sentiment Analysis of Lyrics in Popular Music. *Journal of Popular Music Studies*, 30(4), pages 161-176.

Plitsch, A. 1997. Music + Song = Authentic Listening in the Language Classroom. *In Der Fremdsprachliche Unterricht Englisch*. 31 (1), pages 4–13.

Roman Schneider. 2020. A Corpus Linguistic Perspective on Contemporary German Pop Lyrics with The Multi-Layer Annotated «Songkorpus». In proceedings of LREC 2020, *the 12th Conference on Language Resources and Evaluation*. Marseille, French, 11-16 May, pages 842-848. https://www.aclweb.org/anthology/2020.lrec-1.105/

Shouhui Zhao, Qingsong Luo. 1994. Singing is introduced into Chinese class. *Chinese Language Learning*,1994(4), pages 47-51.

Shiwen Yu, Huiming Duan, Xuefeng Zhu, Bin Sun. 2002.The Basic Processing of Contemporary Chinese Corpus at Peking University. *Journal of Chinese Information Processing*, 2002(5), pages 49-64.

Tegge, F., 2017. The lexical coverage of popular songs in English language teaching. *System*, 2017(67), pages 87-98.

Werner, V., 2012. Love is all around: a corpus-based study of pop lyrics. *Corpora*, 7(1), pages 19-50.

Werner, V. (Ed.), 2018. The language of pop culture. *Routledge Studies in Linguistics 17*. New York: Routledge.

You Fu. 2002. The Introduction of Chinese Songs in Chinese Listening Class. *Journal of Beijing Institute of Education*, 2002(3), pages 64-66.

Yanjing Wang. 2011. The Prevalence of "Sinicism" Songs and Its Application in TCSL. Journal of Sichuan University of Science & Engineering (Social Sciences Edition), 2011(5), pages 86-89.