A standardized Thesaurus of Tibetan and Chinese cognates and non Cognates

Man Zeng,Shan Li,Danhui Wang,Zibei Niu
Tibet University, Lhasa, Tibet, 851000

Abstract: To evaluate Tibetan Chinese cognate words and non cognate words, and construct the corresponding thesaurus, so as to lay the foundation for the study of Tibetan Chinese bilingual cognate words. Systematically evaluate the phonetic similarity of Tibetan and Chinese cognate and non cognate words, and collect basic information such as subjects and vocabulary. Construct a thesaurus containing 60 pairs of Tibetan Chinese cognates and 66 pairs of Tibetan Chinese non cognates. The thesaurus can be used for Tibetan Chinese bilingual vocabulary recognition and Tibetan Chinese bilingual education.

Key words: Tibetan Chinese bilingual;Cognate words;Non cognate words

1 Introduction

One of the core issues in bilingual research is how to construct bilingual mental lexicon. In this regard, researchers mainly use words with similar forms and / or semantics between languages (such as cognates) to investigate. Cognate words refer to words with similar forms (orthography and pronunciation) and semantics between languages (e.g., "_____" = "China"). In contrast, non cognate words refer to translation equivalents that are only semantically similar between languages (for example, "_____" = "chalk"). Before formal research, researchers first need to establish a thesaurus of cognate and non cognate words. At present, researchers have established a vocabulary database of Dutch English, German English and Japanese English cognate and non cognate words. However, there is no Thesaurus of Tibetan and Chinese cognates and non cognates. In order to fill this gap, we evaluated the phonological and semantic similarity and familiarity of Tibetan and Chinese cognate words and non cognate words, and established the corresponding thesaurus.

2 Method

2.1 Research materials

Referring to the previous literature, we invite Tibetan postgraduates majoring in Chinese language and literature at Tibet University to select Tibetan and Chinese words from the Tibetan Chinese dictionary, the Tibetan Chinese Lhasa spoken language dictionary and the new Tibetan English Dictionary of modern Tibet. It is required that all words are nouns and disyllabic words. A total of 180 pairs of cognate words and 180 pairs of non cognate words are selected. Because Tibetan prefers free translation to loanwords, 108 pairs of cognates with both free translation and transliteration are deleted in this study, and only 72 pairs of cognates with Tibetan Chinese transliteration are retained. At the same time, polysemy in non cognate words is deleted, and only 80 pairs of Tibetan Chinese non cognate words translated one-to-one are retained. Finally, in order to ensure that the subjects fill in carefully, an additional 20 pairs of non translated equivalents are added as filler words.

2.2 Research procedure

Referring to the previous literature, phonological similarity, semantic similarity and familiarity were measured using the Likert 5-point scale (1 = very low, 5 = very high). In order to help the subjects understand the contents of the questionnaire, the instructions and topics of the questionnaire are in Tibetan. Before each survey, the researcher will give a brief explanation to the subjects. Encourage the subjects to read aloud. For the phonological and semantic similarity of vocabulary, the subjects were required to evaluate it based on their own intuition; As for the familiarity of vocabulary, the subjects were required to assess the frequency of using these words in oral, writing, reading and listening.

The above assessments of lexical similarity and familiarity were conducted in the form of questionnaires. Before the formal questionnaire survey, all subjects filled in the informed consent form. In order to ensure that the subjects understand the contents of the survey, the main examiner will explain the instructions to the subjects.

2.3 Subjects

A total of 60 Tibetan undergraduate students from Tibet University were recruited in this study and divided into three groups with 20 students in each group. The phonological similarity, translation equivalence and lexical familiarity of vocabulary were evaluated respectively. Three subjects rated the phonetic similarity of most words as 4 or 5, so they were excluded; Four subjects rated the semantic similarity of most words as 1 or 2, so they were deleted. In addition, 7 subjects were rejected because their answers were incomplete. The remaining 46 subjects (21 males) were between 18 and 25 years old, with an average age of 21.23 (SD = 1.96). All subjects' mother tongue is Tibetan, and they live in Tibetan areas most of the time, without language barriers. Before the formal investigation, the subjects need to score their Chinese second language level on a 5-point scale from listening, speaking, reading and writing. The scoring standard is 1-5. 1 represents very low ability, and 5 represents very high ability. On average, the subjects started to speak Chinese at the age of 7.65 (sd=1.79),

and their overall second language level was 3.80 (sd=0.70). See Table 1 for the specific information of the subjects. All subjects voluntarily participated in the survey, and each subject received a certain amount of remuneration after the survey. Before the formal investigation, all subjects signed the informed consent form.

Table 1Number of subjects, age and time of second language learning, and second language level (m ± d)

	Speech similarity assessment	Semantic similarity assessment	Chinese familiarity assessment
Number of subjects	15	15	16
Age of subjects	20.87 (2.06)	21.67 (2.03)	21.20 (1.80)
Time to start learning a second language	7.93 (1.98)	7.73 (1.94)	7.27 (1.54)
Subjects' second language learning time	12.80 (2.85)	13.73 (3.24)	13.93 (2.00)
Subjects' second language listening level	3.73 (0.80)	3.53 (0.52)	3.80 (0.70)
The subjects' oral level of second language	4.07 (0.80)	3.67 (0.61)	3.67 (0.70)
Subjects' second language reading level	4.20 (0.77)	4.27 (0.46)	4.20 (0.80)
Subjects' L2 writing level	3.33 (0.62)	3.53 (0.52)	3.60 (0.51)
The overall level of the subjects' second language	3.83 (0.81)	3.75 (0.60)	3.81 (0.70)

3 Research results

The evaluation results of phonological similarity, semantic similarity and familiarity between Tibetan and Chinese cognates and non cognates are shown in Table 2. Among the 72 pairs of cognates, 7 pairs of cognates were deleted because their phonetic similarity scores were lower than 3; The familiarity score of 5 pairs of cognates was also lower than 3, so they were also removed, and finally 60 pairs of cognates remained. Among the 80 pairs of non cognate words, 9 pairs of non cognate words scored 3, so they were deleted; The semantic similarity of 3 pairs of non cognate words is less than or equal to 3, and they will also be deleted. The familiarity of 2 pairs of non cognate words is less than or equal to 3, so they will also be removed. Finally, 66 pairs of non cognate words remain. See Appendix for 60 pairs of cognates and 66 pairs of cognates.

Because the number of cognates and non cognates in this study is not equal, and the two groups of data are found to be non normal distribution during data analysis, Wilcoxon nonparametric test is used to analyze the data. Data analysis is conducted in the R language environment (r core team, 2022). Referring to previous literature, we first analyze the phonological and semantic similarities of Tibetan and Chinese cognates and non cognates, and then further analyze the familiarity of vocabulary, the number of Chinese strokes and the number of Tibetan characters.

Table 2The descriptive statistics results of phonetic similarity, semantic similarity, familiarity and word length of cognates and non Cognates

	Tibetan characters	Chinese stroke	Phonetic similarity	Semantic similarity	Familiarity
Cognate words	5.82 (1.48)	15.3 (4.11)	3.79 (0.27)	4.37 (0.30)	3.85 (0.42)
Non cognate words	5.92 (1.46)	14.3 (4.79)	1.76 (0.20)	4.43 (0.24)	3.95 (0.32)

The results show that the phonological similarity of Tibetan and Chinese cognates is significantly higher than that of non cognates ($w = 3960$, $P < 0.05$). Cognate words and non cognate words match on semantic similarity ($w = 1699.5$, $P = 0.17$), lexical familiarity ($w = 1651.5$, $P = 0.11$), Chinese strokes ($w = 2173.5$, $P = 0.34$) and Tibetan characters ($w = 1852.5$, $P = 0.53$). The statistical results of phonetic similarity, semantic similarity, familiarity and word length of cognates and non cognates are shown in Table 2.

In addition, referring to the previous literature, we also analyzed the correlation between the phonological similarity, semantic similarity, lexical familiarity, the number of Chinese strokes and the number of Tibetan characters of Tibetan and Chinese cognates and non cognates. The analysis results show that the absolute value of the correlation between the above attributes is less than 0.3, so it can be considered that there is no correlation between the above attributes. See Table 3 for the analysis results.

Table 3The phonetic similarity, semantic similarity, lexical familiarity, the number of Chinese strokes and the correlation between the number of Tibetan characters of Tibetan and Chinese Cognates

	1	2	3	4	5
1. voice similarity	_It is necessary to	-0.06	-0.10	0.03	-0.04
2. semantic similarity		_It is necessary to	0.23	0.05	-0.17
3. vocabulary familiarity			_It is necessary to	-0.03	-0.12
4. number of Chinese strokes				_It is necessary to	-0.05
5. number of Tibetan strokes					_It is necessary to

4 conclusion

By evaluating the phonetic and semantic similarities between Tibetan and Chinese cognates and non cognates, this study aims to construct a standardized Thesaurus of Tibetan and Chinese cognates and non cognates, which includes 60 pairs of Tibetan and Chinese cognates and 66 pairs of Tibetan and Chinese non cognates. In the future, researchers can select words that meet the requirements from the vocabulary database according to the experimental needs to conduct bilingual cognates related research in the paradigms of picture naming, masking translation initiation, word judgment and eye tracking. In addition, it is also helpful for the research on Tibetan Chinese bilingual education.

In addition, however, this study also has some limitations. First, the number of subjects in this study is small (n = 46). The validity of this thesaurus needs to be further verified and improved through the accumulation of sample size in future studies; Secondly, the number of this thesaurus is also small, and future research should further enrich the standardized Thesaurus of cognate words and non cognate words.

References:

[1] Heredia R, cielicka a a B, falandays B, et al. bilingual toxic ambiguity resolution[m].2019

[2] Poort EVA D, Rodd Jennifer M. a database of Dutch English cognates, interlingual graphs and translation equivalents[J] Journal of cognition, 2019,2 (1)

[3] David B Allen, Kathy conklinCross linguistic similarity and task demands in Japanese English bilingual processing[J] PLoS One, 2017,8 (8)

[4] Ton Dijkstra, koji Miwa, Bianca brummelhuis, Maya sappelli, Harald baayenHow cross language similarity and task demands affect associate recognition[j]Journal of memory and language, 2009,62 (3)

[5] Allen, D., & Conklin, K. cross lingual similarity norms for Japanese – English translation assets[j]Behavior research methods, 2013, 46 (2), 540 – 563

[6] Friel Brian M., Kennison Shelia MIdentifying German – English cognates, false cognates, and non cognates: methodological issues and descriptive norms[j]Bilingualism: language and cognition, 2001,4 (3)

[7] Chen Shifa, Fu Tingting, Zhao Minghui, Zhang Yuqing, Peng Yule, Yang Lianrui, Gu xiaolanMasked translation priming with confidence of cross script cognates in visual word recognition by Chinese learners of english: an ERP study [j]Frontiers in psychology, 2022,12

[8] Yisun Zhang Tibetan Chinese dictionary [m]Beijing: Ethnic Publishing House, 1985

[9] Daojia Yu Chinese Lhasa spoken dictionary [m]Beijing: Ethnic Publishing House, 1988

[10] Goldstein m CThe new Tibetan English Dictionary of modern tibetan[m]University of California Press, 2001

[11] R core team r: a language and environment for statistical computingR foundation for statistical computing, viennahttps://www.R-project.org/, November 28, 2022

[12] Jiaxin Chen,Yang Liu,Suxia Wen Masked translation priming effect of cognates and non cognates in unbalanced bilinguals [j]Psychological research, 2019,39 (04): 332-336

[13] Higashitani n. cross script associate priming effects on visual word recognition: effects of Japanese loanword associates in L2 Japanese learners[D] University of kansastwo thousand and fifteen

[14] Hoshino Noriko, Kroll Judith F. coordinate effects in picture naming: does cross language activation survive a change of script?[J] Cognition, 2008106 (1)

[15] Allen, D. the impact of coordinate strategy training on guessing accuracy for unknown visually presented words: the case of Japanese learners of englishLanguage teaching research, 2022, 0 (0)

This paper is supported by the “high level talent training project” (2020-gsp-s155) for Postgraduates of Tibet University