Analysis on types of spelling errors in true Tibetan characters

Maocuo San[1,2,3,*], *Zhijie* Cai[1,2,3,4] , *Rangzhuoma* Cai[1,2,3,4], and *Jizhaxi* Dao[1,2,3]

[1]College of Computer Science and Technology, Qinghai Normal University, Qinghai Xining, China
[2]Key Laboratory of Tibetan Information Processing, Ministry of Education, Qinghai Xining, China
[3]Tibetan Information Processing and Machine Translation Key Laboratory of Qinghai Province, Qinghai Xining, China
[4]School of Computer Science and Technology，Southwest Minzu University, Sichuan Chengdu 610041, China

Abstract. Spelling error checking is a challenging research topic with a wide range of applications such as text editing, word processing, spell checking, teaching, etc. As an alphabetic language, spelling errors in Tibetan could be categorized into three types, namely, non-true type, true type, and punctuation misuse. In order to study true Tibetan syllable spelling error in much more depth, the article analyses the types of True Tibetan syllable spelling errors based on Tibetan word formation rules, grammar and semantic features laying a foundation for Tibetan spelling error checking research.

1 Introduction

With the rapid growth in the amount of information in Tibetan texts available online, Tibetan spelling error checking has become an urgent demand, raising huge interests in related research and application in the community. Given the fact that the more detailed and thorough the analysis of the types of spelling errors is, the more effective the design of spell-checking strategies will be, analyzing the types of errors in Tibetan texts, summarizing and categorizing the rules and commonalities of spelling errors are essential for developing in-depth and effective spelling checking methods. The spelling of Tibetan text includes three aspects: non-true characters, true characters, and punctuation. In recent years, researchers have conducted research on the spelling check of Tibetan non-true characters, and many valuable research results have been obtained [1-3]. Tibetan true-character spelling checking is also an important part of Tibetan text spelling checking, and scholars have also begun to pay attention to the research of true-character spelling checking. The analysis of the types of errors in the Tibetan true-character spelling check is the basic work of the true-character spelling check, but there are no related documents, which affects the development of the spell check technology of Tibetan text. This article takes Tibetan word formation rules, grammar and semantics as the starting point, analyzes the types of spelling errors in Tibetan true characters, and provides data support for the study of Tibetan true characters spelling checking technology.

* Corresponding author: 2627996852@qq.com

2 Research status

In 1967, British linguist Corder [4] proposed the concept of error analysis for the first time. He systematically analyzed the errors in the collected text corpus, and studied the nature and types of errors, which opened the era of text error analysis. Due to the complexity of the language itself, there are many types of text errors, and it is difficult to analyze the types of text errors. In order to analyze the types of spell check errors in depth, the Association for Computational Linguistics (ACL) has established a Natural Language Learning Special Interest Group (CoNLL) to discuss the analysis of spell check error types. The goal of CoNLL-2014 [5] is to automatically detect all types of grammatical errors in short English texts written by non-native English speakers and return the corrected text. Inspired by the shared task of analyzing the types of spell checking in English, a lot of researches on the analysis of types of spell-checking errors have been established in China, and this field has received extensive attention from researchers. The International Natural Language Processing and Chinese Computing Conference NLPCC has added a Chinese grammatical error correction task with the goal to detect and correct grammatical errors in Chinese sentences written by non-native Chinese speakers [6]. At the NLPCC2018 evaluation sessions, six teams from the Alibaba, Peking University and other institutions achieved good results. In 2018, Tan et al. analyzed five types of noun singular and plural errors, verb form errors, subject-predicate inconsistency errors, article errors, and preposition errors that ESL learners often make, and proposed a method based on LSTM and N-Grammatical error correction method [7]. In 2020, Liang et al. classified and analyzed the spelling errors of English learners, and designed an automatic spelling check system for the corresponding types [8].

Since the beginning of the 21st century, scholars have begun to analyze Tibetan spelling errors, mainly focusing on the analysis of non-truth spelling check types. In 2009, Dorje Dolma elaborated on the diversity of spelling errors in Tibetan texts, and used the n-gram model to solve the problem of checking Tibetan syllables [9]. In 2011, Guan Bai analyzed the types of errors in Tibetan characters and designed a method of proofreading the corresponding Tibetan syllable characters [10]. In 2013, Zhu Jie et al. discussed the spelling check of Tibetan syllables, the error check of Sanskrit transliteration, the check of continuous relations and the error check of Tibetan words based on the five defined types of Tibetan text errors, text proofreading system [2]. In 2017, Liu et al. calculated the types of spelling errors of non-true characters on the corpus containing more than 90 million syllables on Tibetan web pages according to predetermined rules, and analyzed the causes of the spelling errors [3]. The analysis of the types of errors in the Tibetan true-character spelling check is the basic work of the true-character spelling check, but there is no relevant literature yet. This article takes Tibetan word formation rules, grammar and semantics as the starting point, analyzes the types of spelling errors in Tibetan true characters, and provides data support for the study of Tibetan true characters spelling checking technology.

3 Types of spelling errors in true Tibetan characters

3.1 Classification of spelling errors in Tibetan text

Tibetan is composed of letters as syllables, syllables as words, words as phrases, and phrases as sentences. Therefore, there are spelling errors at the letter-level, word-level, grammatical-level, semantic-level and punctuation. Non-true character errors refer to Tibetan typos that do not conform to the Tibetan grammar. For example, "ཀ" In "བསོད་ནམས" cannot be preceded by a word, such errors are non-true word errors. True characters refer to

words that comply with the Tibetan word formation rules but are wrong in the context. For example, in the sentence "ང་སློབ་མ་ཡིན།" (I am a student), each Tibetan character is correct individually. However, according to the meaning of the sentence, the word "སློབ" should be "སློབ་", which is a true character error. Through analysis, it is found that the types of punctuation errors in Tibetan texts are mainly in the use of syllable separators and single vertical characters, including two types of missing and redundant. For example, "དུས་ཚོད།" is missing the separator between the syllables "དུ" and "སོང", which belongs to the wrong type of punctuation absence; there are two types of errors in the sentence "ང་རྒྱ་སྐད་དུ་སོང།" (I go to the street). There are two types of errors, missing punctuation and punctuation. The syllable "རྒྱ"and "སྐད" appear between two syllables, Belongs to the type of redundant punctuation errors. The syllable "སོང" and the single vertical character "།" lack a syllable separator, which belongs to the type of missing punctuation errors.

Letter-level spelling errors are non-genuine errors. To judge whether Tibetan characters conform to the word formation principles in the grammar, the characters themselves are considered separately, and they have nothing to do with the context. Word level, grammatical level, and semantic level are the types of spelling mistakes of true characters, and it is judged whether the characters conforming to the principle of character formation are correct in the context.

3.2 Classification of spelling errors in True Tibetan characters

At present, spelling errors of true characters are the most concerned research content in the field of spell checking in Tibetan texts, and these research has great significance and value to Tibetan NLP in general. By analyzing the Tibetan grammar and Tibetan corpus, we analyzed the types of Tibetan true-type spelling errors, and concluded the types of Tibetan true-type errors, including word formation errors, grammatical errors, semantic errors, and joint errors. Class, see Table 1.

Table 1. Categories of spelling errors in True Tibetan characters.

Type 1 error	Word formation error	Grammatical errors	Semantic error	Concatenated error
Secondary error type	pre-added characters errors, upper-added characters errors, root characters errors, lower-added characters errors, vowel errors, post-added characters errors, Then add the character errors, component mixed errors	non-free function word added errors, Verb tense error	Semantic collocation errors, Abbreviation error, Predicate redundant error, Literal translation error	------

3.2.1 Word formation error

Tibetan word formation means that a single letter or a single syllable can be combined with other proper Tibetan characters or even with itself only to form a word. The spelling errors are caused by the similarity of the word formation or sound. This article divides Tibetan word formation errors into eight categories: pre-added characters errors, upper-added characters errors, root characters, lower-added characters errors, vowel errors, post-added characters errors, Then add the character errors, and component mixed errors. For example, in the sentence "མཚོའི་ནང་གི་གྲུ་རྒྱུད།" (A boat in the lake), the word "གྲུ་རྒྱུད" (boat) is wrongly

written "གྱི་ཚུང"(Knife).

3.2.2 Grammatical errors

Tibetan grammar consists of two parts: "*thirty ode*" and "*Character organization law*". There are 10 kinds of non-free function words in "*thirty ode*", each of which has its own adding rules. Its spelling errors are mainly reflected in the addition of non-free function words, that is, the addition of the current conjunction or function word is related to the addition of the preceding syllable. The spelling errors in the word organization mainly lie in the verb tense change, that is, the choice of the current verb tense depends on a key time word or some specific words in the sentence. According to these two parts, this paper divides grammatical errors into adding errors of non-free function words and verb tense errors, For example, "དེ" (should be "ཏེ") in the sentence "འཇོམ་དམུལ་དེ་བཤད།" (said with a smile) violates the rules of adding words to be described It is an error of adding a non-free function word; another example is the sentence "ཇོན་ཚད་ཨ་ཕས་ཡས་ཀ་དེ་སྒྲུབ་མྱོང་།" (the job that father did before) according to the time word "ཇོན་ཚད" (before) at the beginning of the sentence, which determines that the verb tense corresponding to the subject should be the past The tense "བསྒྲུབས", and the verb "སྒྲུབ" is the present tense, constitutes a verb tense error.

3.2.3 Semantic error

The Tibetan grammar system is very rich. Unclear understanding of the Tibetan grammar or the meaning of words will cause logical errors as we refer to semantic errors, which can be divided into semantic collocation errors, abbreviation errors, predicate redundancy errors, and literal translation errors. There are four types of errors.

Semantic collocation errors refer to grammatical rules but unreasonable semantic collocations. Semantic collocation errors in Tibetan generally occur in the semantic collocation of verbs, the use of transitive and intransitive verbs, the semantic collocation of quantifiers, and the usage of honorifics. For example, in the phrases "འདོམ་ད་ལ་ཚད" (measurement), "ཐྲུ་ལ་ཚད" (measurement ruler), and "སོར་ལ་ཚད" (measurement), although there is no problem in language structure and grammar, improper verb collocation is an error in verb meaning collocation (the correct collocation is "འདོམ་ད་ལ་འཇལ", "ཐྲུ་ལ་ཚད", "སོར་ལ་བསྟར"); in the sentence "བཀྲ་ཤིས་ཀྱིས་ཇོང་ཤོ་ཚད" (Tashi cuts down the tree), "ཚད" is an intransitive verb, meaning that the tree is automatically broken without any external force, but in this example sentence The behavior of "cut down the tree" is closely related to the subject "བཀྲ་ཤིས", so the transitive verb "བཅད" (correctly "བཀྲ་ཤིས་ཀྱིས་ཇོང་ཤོ་བཅད") should be used in this sentence, which belongs to the misuse of transitive and intransitive verbs; The sentence "ང་ལ་སྐྱང་གཞི་ཟུང་།" (I'm sick) is an honorific usage error (the correct one should be "ང་ལ་ན་ཚ་ཟུང་།" or "བདག་ལ་སྐྱང་གཞི་ཟུང་།"). This is because the subject "ང" is a first-person pronoun. In Tibetan, honorifics are generally not used for oneself, but for the listener, the elders, distinguished persons, and those who have passed. In addition to the above types of errors, the use of affirmative words "ཡིན" and "རེད" often occurs in semantic collocation errors. "ཡིན" (རང་དང་རང་གི་ཐྱོགས་སྐོན་ར) is used when referring to the person or being close to the person, and is generally used with "ང", "ང་ཚོ", "ཁྱ་ཅག", etc. "རེད" (གཞན་དང་གཞན་གྱི་ཐྱོགས་སྐོན་ར) is used when referring to or related to others, generally used in conjunction with "ཁོ", "ཁོ་ཚོ", "ཏེ", "ཏེ་ཚོ", etc. For example, in the sentence "ཁོ་བློ་བཟང་ཡིན", "ཁོ" should be matched with "རེད" or "ང" should be matched with "ཡིན". Therefore, the correct sentence should be

"ཐ་བྷོ་བཟང་རེད" or "ང་བྷོ་བཟང་ཡིན".

There is a method called abbreviation in Tibetan. This method shortens long words into shorter syllables or into one syllable. The purpose of this method is to stay true to the original text without changing the theme or central idea of the original text. For example, "བཀྲ་ཤིས" (Auspicious or Tashi) is abbreviated as "བཀྱིས", and the original meaning of "བཀྱིས" (the future tense of "སྐིད") is to lead, guide, and quote. This abbreviation law violates the purpose of abbreviating loyalty to the original text, and its original meaning there is an ambiguity between them, which is a type of abbreviation error.

The error that occurs when the predicate is added after the object and predicate form a phrase is called predicate redundancy error. For example, the "ཐག་གཅོད" in "ཐག་གཅོད་བྱས" is composed of the object (ཐག) and the predicate (གཅོད), "ཐག" means able, and "གཅོད" is an action verb related to the object. There is no need to add a predicate "བྱས" (the correct usage is "ཐག་བཅད") after it. Predicate redundant error. Predicate redundancy errors occur frequently in Tibetan texts, such as "བེད་སྤྱོད་བཏང", "རྒྱ་བསྐྱེད་བྱས", etc., are all predicate redundancy errors.

A type of literal translation error often appears in the translated text, which is an error that violates the requirement of maintaining the original content and the original form during literal translation. For example, the literal translation of “草原上鲜花盛开” (flowers on the grassland) is "རྩྭ་ཐང་སྔེར་དུ་མེ་ཏོག་བཞད". Although there are no grammatical errors in this sentence, there are semantic problems (should be translated as "རྩྭ་ཐང་དུ་མེ་ཏོག་བཞད།") due to the influence of Chinese, which is a literal translation error.

3.2.4 Concatenated error

There is a type of error in the Tibetan text. When this error is changed, an error occurs in other positions. This type of error is called concatenated error. For example, in the sentences "དབུལ་ཕོངས་ཁྱིམ་ཚངས་ཀྱི་དཔལ་འབྱོར་ནུས་པ་མ་གཏེ་ནས་ཞན" and "སྐྱན་འཛེབས་ཕྲན་པའི་རོལ་དབྱང་གིས་སེམས་ཀྱི་སྣང་བ་སྐྱིད", the nouns "ཁྱིམ་ཚངས" and "རོལ་དབྱང" are spelled incorrectly, the usage of the non-free function words "ཀྱི" and "གིས" is correct, and when the wrong "ཁྱིམ་ཚངས" and "རོལ་དབྱང" are corrected to the correct words "ཁྱིམ་ཚང" and "རོལ་དབྱངས", The subsequent non-free function words "ཀྱི" and "གིས" make mistakes and are joint errors (the correct sentence should be " དབུལ་ཕོངས་ཁྱིམ་ཚང་གི་དཔལ་འབྱོར་ནུས་པ་མ་གཏེ་ནས་ཞན" and "སྐྱན་འཛེབས་ཕྲན་པའི་རོལ་དབྱངས་ཀྱིས་སེམས་ཀྱི་སྣང་བ་སྐྱིད").

4 Conclusion

There will be spelling errors in the process of using any language. This article uses Tibetan word formation rules, grammar and semantics as the starting point, analyzes the types of Tibetan true-type spelling errors, and divides Tibetan text into non-true characters errors, true characters errors and punctuation errors, and then further divide the true characters errors into the first level error types, such as word formation errors, grammatical errors, semantic errors and concatenated errors. The second level classification of word formation error types, grammatical error types and semantic error types is made. The results of this research lay the foundation for the downstream task of Tibetan spelling checking technology. On the basis of this achievement, we will study its spell check methods for different types of errors to improve the performance of automatic spell checking of Tibetan text.

This research was financially supported by the National Natural Science Foundation of China (61866032, 61966031), Qinghai Provincial Department of Science and Technology (2019-SF-129), "Changjiang Scholars and Innovation Team Development Program" Innovation Team Funding Project (IRT1068), Qinghai Provincial Key Laboratory Project (2013-Z-Y17, 2014-Z-Y32, 2015-Z-Y03), Key Laboratory of Tibetan Information Processing and Machine Translation (2013-Y-17).Qinghai Normal University 2020-2021 Innovative Training Project.

References

1. Cai Zhijie, Sun Maosong, Cairang Zhuoma. A method for spell checking Tibetan characters based on vector model[J].Journal of Chinese Information Processing,2018,32(09):47-55.
2. Zhu Jie, Li Tianrui, Liu Shengjiu. Tibetan text automatic proofreading method and system design [J]. Journal of Peking University (Natural Science Edition), 2014, 50(01):142-148.
3. Liu Huidan, Hong Jinling, Nuo Minghua, Wu Jian. Statistics and analysis of spelling errors in Tibetan syllables based on large-scale internet corpus[J].Journal of Chinese Information Processing,2017,31(02):61-70.
4. S.P.Corder.1967.the Significance of Learner's Errors [M]. International Review of Applied Linguistics,(5):161-170.
5. Ng H T, Wu S M, Briscoe T, et al. The CoNLL-2014 Shared Task on Grammatical Error Correction[C]// Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task. 2014.
6. Zhao Y, Jiang N, Sun W, et al. Overview of the NLPCC 2018 Shared Task: Grammatical Error Correction[C]// CCF International Conference on Natural Language Processing and Chinese Computing. Springer, Cham, 2018:439-445.
7. Tan Yongmei, Yang Yixiao, Yang Lin, Liu Shuwen. Automatic correction method for grammatical errors of ESL articles based on LSTM and N-gram [J].Journal of Chinese Information Processing, 2018, 32(06):19-27.
8. Liang Maocheng, Deng Hailong. Research on automatic spell check for large-scale English learner corpus construction [J].Foreign Language Audio-visual Teaching, 2020(01):31-37+5.
9. Duojie Zhuoma. Research on the application of N-element model in local error checking of Tibetan text [J].Computer Engineering and Science, 2009, 31(4): 117-119+123.
10. Guan Bai. Research on modern Tibetan syllable characters in automatic proofreading [J].Journal of Tibet University (Natural Science Edition), 2011, 26(01):69-75.
11. S. P. Corder.1971.Idiosyncratic Dialects and Error Analysis [J].IRAL, (8):95-110.
12. S. P. Corder.1981.analysis and interlanguage [M]. Oxford: Oxford University Press.