# An empirical study on the distribution of Tibetan monosyllabic monomorphemic words based on corpus

Dongzhi Tsering [1,a*],Kunyu Qi[2,b], Cairang Yun[3,c]

[1]Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education Northwest Minzu UniversityLanzhou, China
[2]Gansu Provincial Key Laboratory of intelligent processing of national languages / Northwest Minzu UniversityLanzhou, China
[3]"One Belt One Road" Research Center for Multilingual Intelligent Processing and Humanities and Social Big Data / Northwest Minzu UniversityLanzhou, China

[a*]2730491560@qq.com, [b]37795386@qq.com,[c]773588480@qq.com

**Abstract.** As the most basic constituent unit of language, monomorphemic words are of great relevance to the fields of natural language processing, language ontology research, and language application. This paper takes Tibetan monomorphemic words as the research object and uses the Tibetan lexical annotation corpus that has been manually annotated and the Tibetan dictionary that contains a large number of Tibetan monosyllabic monomorphemic words and their paraphrase examples as the research materials. The extraction model of Tibetan monosyllabic monomorphemic words is designed by combining the dictionary corpus and the lexical annotation corpus. A comprehensive and accurate analysis of its distribution and concurrent class phenomenon is carried out through a statistical study based on the corpus. The paper obtains a graded word list based on relative generality, in which the total number of monosyllabic monomorphemic words such as nouns, verbs, adjectives, and adverbs is 1599. This is the first systematic and comprehensive statistical study of Tibetan monosyllabic monomorphemic words, which not only provides reliable data support for in-depth research in this field but also provides valuable references for natural language processing and computer applications in Tibetan.

**Keywords:** Tibetan monomorphemic words; Corpus; Statistical analysis;

## 1 Introduction

"Learning a language also means learning its CIs of lifestyles, institutional systems, and spiritual values as language is a carrier of culture". [1] "subsystems change at a different rate. A history of language is a history of changes in the language system. Order parameters are certain constants of the system that determine its state and behavior. It seems obvious that the vocabulary of a language is a dynamic subsystem that changes rapidly." [2] The total number of words is a variable and monomorphemic words are a constant. A monomorphemic word is an important unit of grammatical study and an


A. Rauf et al. (eds.), *Proceedings of the 3rd International Conference on Management Science and Software Engineering (ICMSSE 2023)*, Atlantis Highlights in Engineering 20,
https://doi.org/10.2991/978-94-6463-262-0_3

important word-forming unit in language, and the monomorphemic word can speculate and explain the pattern of production and evolution of a national script.

In the present time, the theory and technology of natural language processing are not yet perfect, and through the analysis of the current state of research on Tibetan information processing, then concludes that the theoretical system of intelligent processing of the Tibetan language needs further construction, and the technical methods still need to be improved. Metrological methods can help open up new research areas and help re-examine, review, and corroborate existing lexical theories and perspectives. The use of metrological research methods is a trend in contemporary lexical research, and the study of Tibetan monomorphemic words is a gap, so it is necessary to study Tibetan monomorphemic words by metrology.

"The metric study of Tibetan monosyllabic monomorphemic words is a corpus-based method which can use computer technology to count and analyze the distribution, frequency, and combination of Tibetan monosyllabic monomorphemic words to reveal the patterns and characteristics of Tibetan monosyllabic monomorphemic words."[3] (Zhaxi jia, Dondrub Tsering 2010)

## 2 A study of the cognition of monomorphemic words

The basic vocabulary is mainly composed of monomorphemic words with single morphemes, and the size of the basic vocabulary of a text can be known by determining monomorphemic words. "The internal structure of words is divided into monomorphemic words and synthetic words, and synthetic words include derived words and compound words".[4] Tibetan monomorphemic words include features such as commonality and temporal stability and strong word formation ability.

### 2.1 The concept of monomorphemic word

Monomorphemic words are a basic constituent unit in the language, and they are also the most basic lexical unit. They do not require any prefixes, suffixes or midfixes to be formed, and consist of only a root or stem. These monomorphemic words reflect the basic structure and rules of the language and are essential for learning the language and understanding its morphology and grammar. In addition to the language itself, monomorphemic words can also convey cultural values and ideas. For example, a word composed of a morpheme is often called a "monomorphemic word" in English, where "mono-" denotes a single morpheme, "morphemic" denotes a morpheme of a morpheme, and "word" denotes a word. Such words are usually the monomorphemicst lexical units, usually consisting of a basic root or stem without any prefixes or suffixes. For example, the English word "dog" consists of a single root word "dog" and is a typical monomorphemic word. In Chinese, a monomorphemic word is also a word consisting of a single morpheme, which is highly consistent with Tibetan and English. In traditional Tibetan grammar, the monomorphemic word is not described in terms of word structure, but most scholars now contrast it with the word "rkyng ming". The monomorphemic word contrasts with the compound word and provides an important

linguistic form for the morphological composition of the Tibetan vocabulary. This paper provides a Tibetan monomorphemic extraction method that can be used for other NLP tasks, which provides valuable resources for Tibetan language research The article focuses on the internal structure of Tibetan words in terms of monosyllabic words and monomorphemic word constructions and phenomena, and analyzes and summarizes the composition of Tibetan monomorphemic words with the help of literature research and corpus statistics based on the research results of Tibetan vocabulary. The rules of Tibetan monomorphemic words are verified, and the basic word list of monomorphemic words is summarized. h, promotes the development of Tibetan NLP applications, and is important for the construction of the Sino-Tibetan language resource base.

## 2.2 Tibetan monomorphemic words

Tibetan is largely an adherent language and also has the characteristics of an isolated language. Words are composed of several morphemes, each representing a grammatical or semantic element. These morphemes are usually attached to roots or stems by affixes to form complex words. There is a growing number of scholars who study Tibetan from modern linguistic theory, but a complete system of modern language theory has not yet been developed.

There is a rich theory of the concept and classification of words in Tibetan, and it is very important to explain the Tibetan word (ming) before studying the Tibetan monomorphemic words. There is a book on lexis by Jiang Di of the Chinese Academy of Social Sciences, "About Lexis and Morphology."[5] and "Introduction to Tibetan Lexicography" by Professor Deng Ge of Tibetan University, in which Professor Deng Ge names monomorphemic words (ming rkyng) and studies the composition and morphology of Tibetan monomorphemic words.

"Tibetan Morphology and Lexicon"(by Nathan W. Hill) [6] gives some examples of Tibetan monomorphemic words, such as "gshegs", meaning "good death"), etc. The book also introduces the morphological changes of some monomorphemic words and the ways of forming derived words.

In conclusion, a monomorphemic word in Tibetan is a word that consists of only one root or stem without any prefixes, suffixes, or medial affixes. It contrasts with compound words and provides an important linguistic form for the morphological composition of the Tibetan vocabulary. There are also some related studies in the field of computers for the processing of Tibetan monomorphemic words. For example, Cho et al. (2013), "We proposed a method to automatically extend Tibetan monosyllabic monomorphemic words using the pictoriality score of the MRC psycholinguistic database. We used a morphological analyzer based on roots and affixes to distinguish Tibetan monosyllabic monomorphemic words from compound words, and then used the imageability scores of English monosyllabic monomorphemic words from the MRC database as a reference to assign corresponding imageability scores to Tibetan monosyllabic monomorphemic words." [7]

# 3 Construction of a monomorphemic word extraction system for Tibetan

The construction of the Tibetan language corpus is a huge linguistic research project, especially the large-scale and high-quality annotated Tibetan language corpus, which is becoming more and more important in the field of natural language processing of Tibetan, not only as a corpus for language models, but also as test data for various language models and practical tools, and at the same time, this corpus is also an important resource for theoretical linguists to mine linguistic laws from it.

## 3.1 Introduction to the Corpus

The lexical annotated corpus is an annotated data set formed by annotating each word in a natural language text with its lexical identity. The lexical annotated corpus of Tibetan language text comes from the Key Laboratory of the Ministry of Education of Northwest University for Nationalities, with a size of 50 million words. The larger the size of the corpus the richer the linguistic data it contains, providing more examples of linguistic phenomena and language changes. At the same time, a balanced corpus covers different fields, different language types, different genres, etc. The corpus of this dissertation covers to the greatest extent the vocabulary in Tibetan culture and the traditional Tibetan disciplinary classification of ten ming culture, in addition to philosophy, religion, grammar, fiction, prose, poetry, aphorisms, epic poetry, folk songs, doxology, history, sect history, monastic history, biographies, news, Internet commentaries, periodicals The corpus covers various fields of Tibetanculturese, such as politics, law, textbooks, and dictionaries. The corpus covers all areas of Tibetan culture, thus reflecting the diversity and complexity of the Tibetan language more comprehensively. A large and balanced corpus can improve the reliability and accuracy of the data. By collecting and organizing a large amount of linguistic data, data bias, and errors can be avoided as much as possible, thus reflecting the phenomena and patterns of language more accurately.

"The dictionary has rich linguistic content. The dictionary is a word-based collection. Although there are linguistic components that are not "words" that serve as lexical items, it is still possible to say that words and lexical items are roughly equal to each other. "The lexical items have word forms, pronunciations, interpretations, examples, and labels, which reflect the forms, sounds, and meanings of the words, while the examples provide a small context and the labels reveal the semantic or usage characteristics of the words." [8] Therefore, the main purpose of this thesis is to filter the lexical items in the dictionary. The dictionary corpus is obtained from the corpus database of the "One Belt, One Road" Multilingual Intelligent Processing and Humanities and Social Data Research Center of Northwest University for Nationalities. The dictionary is divided into six corpora, which are listed in detail in Table 1.

**Table 1.** Six Common Tibetan Dictionaries.

| Dictionary Name | Publisher | Author | Date of publication | Number of Entries | Evaluation |
|---|---|---|---|---|---|
| Geshe Chuza's Tibetan Dictionary | Beijing Ethnic PP | Geshe Quji Zhaba | 1957 | 25363 | With the academic attainments and rich practical experience of Tibetan scholar Geshe Quzha, it is one of the dictionaries that have both authority and practicality. |
| The New Book of Orthography | China Tibetology PP | Specialist in Tibetan Studies at the Chinese Academy of Social Sciences | 1983 | 6954 | It is the first essential tool for Tibetan studies and Tibetan language teaching. |
| Tibetan-Chinese Dictionary(Zhang) | Beijing Ethnic PP | Zhang Yisun, Editor-in-Chief | 1985 | 53503 | It is the first comprehensive Tibetan-Chinese dictionary and encyclopedia of Tibetan studies in Chihing and Research Gna. |
| Tibetan-English Dictionary(Northwest) | Gansu Nationalities Press | Tibetan Language Teaching and Research Group, Northwest University for Nationalities | 1996 | 25193 | An important tool for Tibetan language teaching and Tibetan-Chinese translations |
| Tibetan-English Dictionary | Oxford University Press, UK | Sarat Chandra Das, Editor-in-Chief | 1902 | 57153 | The dictionary can also be used as a reference dictionary for English learners to understand the Tibetan language and Tibetan culture. |
| Monlam Tibetan Dictionary | TBRC | MonlamIT | 2019 | 107065 | Highly evaluated and recognized by the general Tibetan public and academia |

### 3.2 Pre-processing of the corpus

"Tibetan character statistics are the basic research method of corpus linguistics, which is empirical in nature." [9] because Tibetan character statistics can not only provide reliable data support for Tibetan information processing but also have important reference value for linguistics as well as cryptology and other disciplines of research, so

statistics of Tibetan monomorphemic words need to construct a statistical system of Tibetan monosyllables first.

**3.2.1 An algorithm for extracting Tibetan monosyllables from a lexical annotation corpus.**The lexical annotation corpus adopts the Tibetan Unicode international encoding system and mainly counts the lexicality of each word and the frequency of each lexicality in the text. The software is designed in Python language. See Table 2 for details.

**Table 2.** The final sentence of a caption must end with a period.

| Algorithm: The number of occurrences of words and the number of occurrences of different lexical properties are counted in the Tibetan lexical annotation corpus. |
|---|
| def pudict(word, tagg, dic):<br>  if word in dic:<br>  dic[word][tagg] = dic[word].get(tagg, 0) + 1<br>  else:<br>  dic[word] = {tagg: 1}<br>def dic_sort_output_ex(odic, dic, fo):<br>  print(odic)<br>  word_list = sorted(odic.items(), key=lambda d:d[1], reverse=True)<br>  with codecs.open(fo, "w", "utf-16") as fw:<br>  for word, freq in word_list:<br>  chunk = "%s\t%d\t" % (word, freq)<br>  if the word in dic:<br>  tmp = "\t".join("%s:%d" % (t,f) for t,f in dic[word].items())<br>  chunk += tmp + "\r\n"<br>  fw.write(chunk) |

This algorithm counts the frequency and lexicality of words. If the word already exists in the dictionary, the number of occurrences of the word's corresponding lexical property is increased; otherwise, a new record is created and the number of occurrences of the lexical property is initialized to 1. The words are sorted by frequency and the results are written to a file. First, sort the words by the number of occurrences from largest to smallest, then record the number of occurrences of each word and the corresponding lexical property into a string, and finally write this string to the output file. If the word has corresponding lexical records in the dictionary, those lexical properties are output along with the corresponding occurrences.

**3.2.2 An algorithm for extracting Tibetan monosyllables from dictionaries.** All the corpora involved in this thesis use the Tibetan Unicode international encoding system. The Tibetan dummy word set is preset, and the frequency of collocations in Tibetan vocabulary is counted using Tibetan symbols and syllable points as cut marks. The software was designed in Python language. The detailed algorithm is shown in Table 3.

**Table 3.** The dictionary monosyllabic word frequency statistics calculation method.

| Algorithm: The dictionary monosyllabic word frequency statistics calculation method. |
|---|
| def readfile():<br>for line in fr:<br>loc = line.find('=')<br>if loc != -1:<br>w = line[:loc]<br>if is_wordone(w):<br>dic[w] = dic.get(w, 0) + 1<br>def is_wordone(w):<br>wa = w.strip('\u0f0d').strip('\u0f0b')<br>loc = wa.find('\u0f0b')<br>return loc == -1 |

This algorithm is used to read a text file and count the frequency of monosyllabic words that appear in it. A monosyllabic word is a word that has only one syllable. The text file is read line by line, the word in each line is extracted, and then it is determined whether the word is a monosyllabic word or not. If it is a monosyllabic word, its occurrences are added by 1, if not, it is ignored. The result of the statistics is saved in a dictionary. A helper function is used to determine whether a word is a monosyllabic word or not. It will remove some special characters from the word and then see if the remaining part has only one syllable. The results obtained by the above algorithm are as follows:

To give the reader a visual effect, the light-colored bars in Figure 1 represent the word order that is already present in the dictionary, while the dark-colored ones represent the monosyllabic word order.

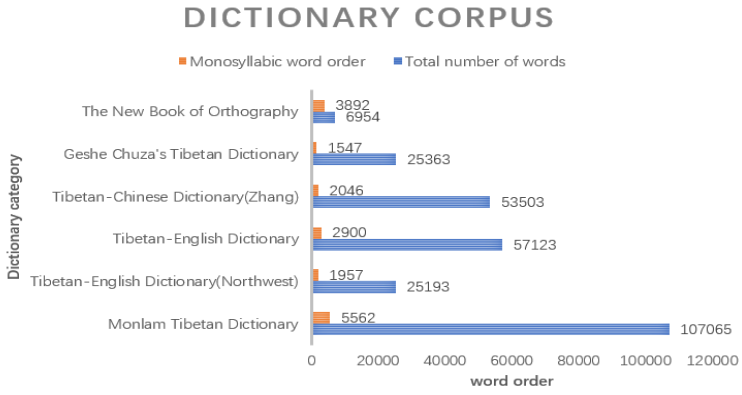


**Fig. 1.** Histogram of the lexical corpus.

In linguistics, the so-called word generality is a composite indicator of how commonly words are used in various areas of language application. Genericity takes into account both the distribution rate and frequency of words. In this paper, we use the six common Tibetan dictionaries to get the relative generality as shown in Figure 2. If a

word is included in six dictionaries, it means high generality in this word, and if the word is only in one dictionary, it means low generality.

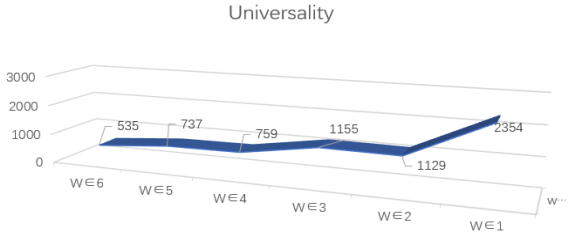


**Fig. 2.** Relative generality of monomorphemic words

Conducting statistics and analysis can tell the relative generality, W refers to the word, ∈ refers to the inclusion, and the number after it indicates the number of dictionaries. w ∈ 6 means that this inclusion is in six dictionaries. w ∈ 6, w ∈ 5, w ∈ 4 The purity of monomorphemic components is high, and the purity of monomorphemic words of w ∈ 3, w ∈ 2, w ∈ 1 is low. In particular, the monosyllabic monomorphemic words in the dictionaries are largely consistent with the basic Tibetan vocabulary, with a very high degree of overlap.

The number of monosyllables in W ∈ 6 is 535, the number of monosyllables extracted from W ∈ 5 is 737, and the number of monosyllables in W ∈ 4 is 759, and the combination of 4, 5, and 6 yields 2031 monosyllabic words, which are then combined with the lexical annotation to obtain 1633. The number of deactivated words (dummy words) is 1599, so I have 1599 monosyllabic monomorphemic words in the lexical corpus.

### 3.3 Technical route of the experiment

This is the general flow diagram of extracting monosyllables from a lexically annotated corpus and a lexical corpus. The lexically annotated corpus is mainly based on a large number of texts, using word separation and annotation. The lexical preponderance of six dictionaries is selected from multiple dictionaries, and the table obtained from the frequency count and the lexical annotation corpus with the lexical properties and the table are neutralized, and finally, a dummy word set is created and the dummy words are deleted, leaving the Tibetan monosyllabic monomorphemic words.

**3.3.1 Two types of monosyllabic word lists merged.**4363 words were obtained from the lexical annotation and 2031 words from the dictionary, and 1599 words were merged.

*I. read the contents of the three input files (lexical annotation monosyllabic, dictionary monosyllabic, and combined monosyllabic monomorphemic words) and merge them into one output file.*

*II. read the contents of the output file and delete the words in the list of dummy words (deactivated words).*

*III. Matchatch the remaining words with the contents of another file and save the result to a new file. 4. read the contents of the new file again and delete the words in the list of dummy words, and save the final result to a new file.*

*IV. Read the contents of the new file again and delete the words that belong to the list of dummy words, and save the final result to a new file.*

**3.3.2 Deactivation of words.**The dictionary corpus obtained 2031 monosyllabic words, but it should be stated here that the words here also include the imaginary part, the imaginary words in Tibetan include grammatical, free imaginary words, and unfree imaginary words, some imaginary words have real meaning, and some imaginary words are half real and half imaginary, so removing the pure imaginary words will greatly improve the purity of the word list.

In Tibetan, purely imaginary words occur very frequently, but they have very little actual meaning in the text, so they can be regarded as deactivated words. Of the 80 dummy words in Tibetan, 30 have real meanings.

The following is a detailed list of words and their meanings for each of them. In the preposition zhing, etc., "zhing shing shig" means field, wood, and louse. In the clitics, "no bo ro lo" means when, corpse, flag, and road, respectively. The final words "ngo do mo vo ro lo so to" mean hear, look, double, divination, oh, corpse, year, and tooth. The "ru ra la du" in the digraph means wing, goat, mountain, and several. The suffix "ba bo mo" means cow, male, female, or divination. The negative "mi min mid" means person, non, none. The ornamental set of "jng" means light. The source auxiliary "ns ls" means barley, karma, or do. The tone auxiliary "na la" means age and mountain. The words "na" and "la" are not only dummy words of the locative auxiliary but also dummy words of the tone, and "mo" is a dummy word of the person suffix and a dummy word of the final word. "su" is not only an interrogative pronoun but also a modal auxiliary, and 30 words have a real meaning, so it is necessary to keep these 30 dummy words in the monosyllabic word list.

The remaining words in the word list after deleting purely imaginary words are monosyllabic monomorphemic words. The number of purely imaginary words is 46, of which "zhe na" "she na" and "ce na" are two-syllable words. 43 purely imaginary words remain. All the above Tibetan characters are in Latin transcription.

## 4 Analysis of experimental results

"Kulkarni proposes a new computational approach for tracking and detecting statistically significant linguistic shifts in the meaning and usage of words. " [10] Tibetan monosyllabic monomorphemic words refer to words that contain only one syllable and only one morpheme, specifically including nouns, verbs, adjectives, adverbs, counters, etc. In this paper, monomorphemic words do not include purely imaginary words (stop words), bisyllabic and multisyllabic compound words, and derivatives.

Tibetan monosyllabic monomorphemic words mainly include some basic words, names of people, places, animals, plants, and so on. In Tibetan texts, monosyllabic monomorphemic words appear more frequently, but their semantic meanings are

relatively monomorphemic and often need to be used in combination with other words to express richer meanings. Therefore, the treatment and use of monosyllabic monomorphemic words are of great importance in Tibetan text processing and language research.

### 4.1 Monomorphemic word master word list analysis

Firstly, a total of 2031 monosyllabic words were obtained from the lexical corpus. Secondly, 4363 monosyllabic words were obtained from the lexical annotated corpus, and finally, the 2031 monosyllabic words in the lexical corpus and the 4363 monosyllabic words obtained from the lexical annotation statistics were combined to obtain 1633 monosyllabic words, and 1599 monosyllabic monomorphemic words were obtained after removing the discontinued words.

The number of monomorphemic words in the lexical corpus changed from 2031 to 1599, which is 432 fewer monomorphemic words, mainly because these are uncommon words, technical terms, and dialect words, as well as the fact that the scope of the annotated corpus is not broad enough and the sample size of the corpus is not large enough, thus leading to the missing words. "Monaghan used a similarity measure based on contextual co-occurrence vectors. The semantic feature representation was not available for the monomorphemic polysyllabic words because it was derived only for monosyllabic words. There were 5138 monosyllabic words with both co-occurrence- and feature-based semantic representations."[11]

Dictionary corpus and lexical annotation corpus are two different kinds of corpora and serve different purposes. The lexical corpus mainly collects and organizes basic information about words. The lexical annotation corpus, on the other hand, is used to assign lexical properties to words, i.e., to identify the roles that words play in sentences, such as nouns, verbs, adjectives, etc. While the lexical corpus is central in the extraction of monosyllabic monomorphemic words, the lexical annotated corpus is not less useful in the text to help us understand and analyze the words in the text more accurately, to better obtain the lexical occupancy of each word in a large corpus. It is of great significance for later training models for text classification, sentiment analysis, and machine translation.

The core goal of this thesis research is to obtain a relatively complete monosyllabic monomorphemic word list in Tibetan through a statistical experimental method combining lexical annotated corpus and lexical corpus. The word list contains five categories of nouns, verbs (transitive and intransitive verbs), adjectives, and adverbs, which are all basic Tibetan words according to the observed analysis.

### 4.2 Monomorphemic word master word list analysis

By analyzing the monomorphemic words under each lexical category, we can understand the basic vocabulary and grammar rules under different lexical categories and further deepen our understanding and mastery of the language. Among the nouns, we can learn more about the use of noun monomorphemicxes, and we can find that there are some differences in the frequency of use of different noun monomorphemicxes. For

example, monomorphemic words such as "ng", "ming", and "chos" are used more frequently, while "ske " and "mchn" and "kld" are used less frequently. This information can help learners master the basics of language more effectively, and can also provide valuable references for research in natural language processing and machine translation. In addition to the usage of monomorphemic words, this study can further analyze the usage and meaning of different words in different contexts. For example, in verbs, "song" can indicate the action of walking or leaving; in nouns, "mi" can indicate the concept of a human or a specific person, and can also be a Negative prefix. These analyses can help learners better understand and use language, and also provide more accurate semantic information for research in the fields of natural language processing and machine translation.

**Table 4.** Major lexical distribution of monosyllabic monomorphemic words

| Words | Word Nature | Word order | Frequency/total monomorphemic words |
|---|---|---|---|
| Nouns | nn | 1249 | 0.781113 |
| transitive verbs | vt | 565 | 0.353346 |
| intransitive verbs | vi | 401 | 0.250782 |
| Adverbs | dd | 152 | 0.095059 |
| Adjectives | aa | 284 | 0.177611 |

Tibetan monosyllabic monomorphemic words have a variety of lexical properties, but in this dissertation, only the five most prominent ones are elaborated(see Table 4): nouns, verbs (transitive and intransitive verbs), adverbs, and adjectives. These lexical properties are the most basic and commonly used lexical properties in the language, and therefore they are more important compared to other lexical properties. Among them, nouns and verbs are the most basic lexical properties in the language, and they can form the backbone of an utterance and express the most basic information. Adverbs and adjectives can further expand and describe the characteristics and states of nouns and verbs, making the language richer and more precise in expressing information.

### 4.3 Monomorphemic word-cum-class analysis

From the lexical distribution table of monomorphemic words, we can see that noun, verbs (transitive verbs, intransitive verbs), adverbs, and adjectives total 2651 times, but the total number of monosyllabic monomorphemic in this table is 1599, a total of 1052 more, the result of this reason is because of the high degree of compatibility in Tibetan, he phenomenon of monosyllabic monomorphemic words that are both nouns and verbs or adverbs or adjectives is extremely common. Since the grammatical structure of Tibetan is more flexible, unlike some other languages which have a clearer division of lexical and grammatical usage. This provides room and possibility for the emergence of partitive classes. "For important words in a corpus, such as polysemous words, the form of a single vector leads to ambiguity in the word's meaning. Clearly, ambiguous or incorrect semantic information will ultimately affect the quality of text representations."[12]

In Table 5, we can know that the high parthenogenesis of Tibetan words and the severity of parthenogenesis in Tibetan are related to the grammatical characteristics and cultural background of the Tibetan language.

**Table 5.** Tibetan high-frequency monosyllabic monomorphemic words word list.

| Words | Total number of words | Main word type | Other words |
|---|---|---|---|
| La | 29945 | pl:26620 | cc:3040 nn:98 pe:10 vi:6 rr:8 gg:5 nt:2 ub:4 cu:46 pg:43 um:9 pn:4 pa:39 aa:1 pl:1 uc:1 nf:1 uo:1 pc:1 |
| Ra | 28171 | pl:27414 | cc:570 uo:1 nn:64 ww:3 gg:1 dd:6 nf:2 pe:3 cu:19 pg:52 ub:2 pa:20 pn:1 ue:1 vi:1 nt:9 nv:1 qq:1 |
| Ns | 19164 | cu:15303 | pn:3071 pl:80 cc:62 nn:30 uv:1 pg:9 pa:5 ub:2 dd:8 um:3 vi:4 ca:2 pe:1 rr:2 un:1 gg:1 uf:3 pc:2 nf:1 uo:2 us:6 |
| Du | 13519 | pl:13451 | nn:7 mm:3 ww:1uo:1 dd:3 pg:21 pa:10 pn:3 rr:3 cc:6 nt:9 nf:1 |
| Bys | 7809 | vt:7684 | nn:5 xx:1 pc:1 pa:2 um:6 ut:110 |
| zhig | 7469 | ub:6423 | nn:8 ue:2 mm:796 um:212 vi:3 rr:8 cu:1 qq:6 cc:1 vt:1dd:8 |

First of all, the grammatical structure of Tibetan is more flexible, unlike some other languages which have more clear divisions in lexical and grammatical usage. This provides room and possibility for the emergence of partitive categories.

Secondly, Tibetan is a minority language with a relatively small vocabulary. To make full use of the limited vocabulary, people often use words in a partitive way, so that the same word can play different roles in different contexts.

Again, there are some special expressions and usages in Tibetan culture and contexts that need to be expressed using parthenogenesis.

Finally, parthenogenesis is a widespread linguistic phenomenon in Tibetan, which reflects both the flexibility of Tibetan grammar and the limited vocabulary and cultural background of Tibetan. Therefore, when learning and using Tibetan, one needs to have a certain understanding and mastery of partitive categories to better understand the structure and operation of the language.

## 5 Conclusion

The formulation of Tibetan monomorphemic words is of great help to the field of Tibetan grammar and Tibetan information processing. The theoretical guidance of lexicography, supported by a large and representative database, determines monomorphemic words in Tibetan as more scientific, accurate, and reasonable. Combining lexical annotated and lexical corpus allows for a more comprehensive and accurate understanding and analysis of the most basic words in a language. This is very helpful for learning and mastering the basics of a language, improving language proficiency, and deepening understanding of the language. At the same time, monomorphemic word statistics can also provide more accurate corpus and semantic information for research

in fields such as natural language processing and machine translation, so that more accurate and efficient algorithms and models can be designed. However, the conclusion of determining word lists by the frequency of word occurrences in this paper is basically based on the results of statistical data, which has a strong empirical basis, but is a preliminary discussion of related research, and therefore needs further in-depth study by related researchers.

## ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (62266038).

## References

1. Tong, Ho Kin, and Lin Hong Cheung. "Cultural identity and language: A proposed framework for cultural globalization and glocalization." *Journal of Multilingual and Multicultural Development* 32.1 (2011): 55-69.)
2. Dombrovan, Tetiana. An introduction to linguistic synergetics. Cambridge Scholars Publishing, 2018.
3. Zhaxijia and Dunzhu Ciren, "The grammatical information of Tibetan case particles for natural language processing," Journal of Chinese Information Processing, vol. 24, no. 5, pp. 18-23, 2010. (in Chinese)
4. Lisheng Wang, "A comparative study of English and Chinese 'words' based on lexical morphology," Zhongzhou Academic Journal, no. 2, pp. 248-251, 2011. (in Chinese)
5. Di Jiang, Tibetan Lexicon and Morphology. Beijing: Peking University Press, Oct. 2022, pp. 109-112. (in Chinese)
6. Hill, Nathan W., and Lauren Gawne. "The contribution of Tibetan languages to the study of evidentiality." Evidential systems of Tibetan languages 302 (2017): 1-38.K.W.Cho et al., "Automatic expansion of the MRC psycholinguistic database imageability ratings," in Proc. ACL 2013, Sofia, Bulgaria, Aug. 2013, pp. 1455-1464.
7. Cho, K.W., Webb, N., Feldman, L.B., Strzalkowski, T., Shaikh, S., Broadwell, G.A., Liu, T., Boz, U., Taylor, S. (2013). Automatic expansion of the MRC psycholinguistic database imageability ratings. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013), Sofia, Bulgaria. pp. 1455-1464.
8. Xinchun Su, Dictionary and Lexical Quantitative Research. Shanghai: Shanghai Lexicographical Publishing House, Dec. 2013. (in Chinese)
9. Kaibao Hu and F. Yang, "Corpus-based literary research: Connotation and significance," Journal of Zhejiang University (Humanities and Social Sciences), vol. 49, no. 5, pp. 143-156, Sep. 2019. (in Chinese)
10. Kulkarni, Vivek, et al. "Statistically significant detection of linguistic change." Proceedings of the 24th international conference on the world wide web. 2015.
11. Monaghan, Padraic, et al. "How arbitrary is language?." Philosophical Transactions of the Royal Society B: Biological Sciences 369.1651 (2014): 20130299.
12. Guo, Shun, and Nianmin Yao. "Polyseme-aware vector representation for text classification." IEEE Access 8 (2020): 135686-135699.