A study on syntactic function measurement of Tibetan nouns based on dependency tree

Zeren Zhuoma, kunyu Qi*

China Institution of Information Technology for Nationalities, Northwest Minzu University, Lanzhou730030, China

Email:3079530743@qq.com, *37795386@qq.com

Abstract. In this paper, the author makes a statistical study on the distribution of grammatical functions of Tibetan nouns according to the probabilistic valency model theory based on the dependency tree of Tibetan words established by himself, and summarizes the grammatical functions of Tibetan nouns by the probability of the existence of each word in each language interdependence From the perspective of research, the author conducts an in-depth study on the grammatical functions of Tibetan nouns, which can make it more clear to understand the use of Tibetan nouns in teaching and their functional distribution in the text.

Keywords: Tibetan; Noun; Dependency tree base; Probability valency.

1 Introduction

Treebank is a corpus with syntactic annotated information. In recent years, as an importan t resource for the study of lexology and syntax, semantics and machine translation, treebank has been paid more and more attention by researchers in the field of linguistics, and researchers have invested a lot of manpower and financial resources in treebank. At present, the representative dependency tree libraries in China include: Chinese Dependency Treebank (CDT) of Harbin Institute of Technology, Chinese Dependency Treebank (TCT) of Tsinghua University, and Chinese Multi-view Dependency Treebank of Peking University (PMT) Peking university Multi-view Chinese Treebank, Soochow university Chinese Treebank (SUCDT), etc[1]. However, in the case of Tibetan, there is no public and shared Tibetan dependency tree. The marked syntactic information in the treebank can provide data support for Tibetan nouns, verbs and other syntactic functions, so it is very important to build a set of Tibetan dependency syntactic treebank that accords with Tibetan grammar and has high accuracy.

Tibetan dependency tree database is a corpus that uses dependency grammar to mark dependency syntactic relations of Tibetan sentences. The dependency tree database contains three main contents: Dependency relationship, dependency direction, dependency distance, dependency relationship is the relationship between two words, including the dominant word, dependent word and dependency relationship has three parts,

each pair of words has a dominant word, there is a subordinate word, a word can dominate several words at the same time, but a word can not be dominated by two or more words; The dependency direction is positive and negative. When the dominant word is behind the subordinate word, the dependency relationship is the dependency relationship after the dominant word, and when the dominant word is in front of the subordinate word, the dependency relationship is the dependency relationship before the dominant word. The dependency distance is the distance difference between the dominant word and the subordinate word, so there are positive and negative. When the dependency is the dependency relationship after the dominant word, the dependency distance of the relationship is positive; when the dependency relationship is the dependency relationship before the dominant word, the dependency distance of the relationship is negative.

As a low-resource language, the study of Tibetan in all aspects is still in the development stage. Although good achievements have been made in Tibetan word segmentation, part of speech tagging, syntactic analysis and corpus construction, the research on syntactic analysis is far from enough, and there is no open source and shared syntactic tree library. Therefore, on the basis of constructing a dependency syntax treebank containing 5500 sentences, I made a statistical study of the syntactic functions of Tibetan nouns, and strive to provide some data support and materials for the syntactic research of Tibetan nouns. Table 1 for Tibetan sentences " བཙན་པོས་སློབ་དཔོན་གྱི་ཞབས་ལ་ཕྱག་ཕུལ། " interdependence syntactic structure with table.

Table 1. Dependency table for "བཙན་པོས་སློབ་དཔོན་གྱི་ཞབས་ལ་ཕྱག་ཕུལ།"

Words	Subordinate word	Characteristic of a certain word	Dominant word	Dependency relationship	Dependency distance
བཙན་པོ	1	NN	8	SUB	7
ས	2	PA	1	PK	-1
སློབ་དཔོན	3	NN	5	ATT1	2
གྱི	4	PG	3	PK	-1
ཞབས	5	NN	8	GOV	3
ལ	6	PL	5	PK	-1
ཕྱག	7	NN	8	OBV	1
ཕུལ	8	VT	0	ROOT	0
།	9	WW	8	WK	-1
…	…	…	…	…	…

As can be seen from Table 2, "བཙན་པོ" and "ཕུལ" are the dominant subject-predicate relations (SUB); "བཙན་པོ" and "ས" are the PK of dominant words; "སློབ་དཔོན" and "ཞབས" are the prepositional attributive relation (ATT1) of the dominant word; "སློབ་དཔོན" and "གྱི" are the PK of dominant words; "ཞབས" and "ཕུལ" are the dominant object relation (GOV); "ཞབས" and "ལ" are the PK of dominant words; "ཕ ྱ ག" and "ཕ ྱ ལ" is the dominant bing dynamic relationship (launched) after the word "ཕ ྱ ལ" and "།" is the dominant word

before the punctuation (WK) a; "ཚེས" is the core word, the core of the sentence, and is not dominated by other words.

Regardless of the language, nouns are the preferred object of study,[2] because "the things that nouns represent are generalized, including the names of objects and persons, of substances, of animals and living things, of facts, phenomena, events, and of non-procedural procedural features (properties, actions, states) as independent things."[3] The universality of nouns in Tibetan corpus can be seen.

2 Theoretical Basis and Method

2.1 Theoretial Basis

Dependency grammar, also known as valency grammar, was initially used most in the study of the German language as a grammatical theory proposed by Tesnier in his Fundamentals of Structure Syntax in 1959. However, the dependency grammar proposed by Tesnier has been widely quoted in the linguistic circles all over the world, and it is very popular in the fields of lexicology, syntax and semantic studies.

"In Tesniere's description of valences, the verb is likened to an atom with hooks, which the verb uses to attract the same number of action elements as it as its subordinate components. The number of hooks that a verb has, that is, the number of action elements that that verb can control, constitutes what is called the valency of the verb.[4] The valences of verbs are divided into two parts: centrifugal force and centripetal force. Centrifugal force refers to the number of dependents controlled by the verb. Centripetal force refers to the number of dominant words that govern the verb when it is a subordinate." According to the magnitude of centrifugal and centripetal forces, we can see the magnitude of a word's force. Valency is the potential ability of a word to combine with a word. It is a static description of a word. Traditional linguists study the force of a word through qualitative methods. In this paper, based on statistical corpus, the dominance of each word in dependency tree and the dependency relationship of subordinate words are quantitatively counted, so as to study the potential combination ability of each word, that is, the collocation ability of a word or the valency ability of a word. The theory of valency is not only the valency of verbs, but also of nouns.[5] We can get the probabilistic valency model of Tibetan nouns by the dependency relationship between dominant and subordinate words in the dependency syntax tree. This quantitative study is a supplement to the study of the syntactic function of Tibetan nouns, and the statistical results are helpful to obtain a clearer understanding of the syntactic function of Tibetan nouns.

2.2 Corpus

The corpus of this paper is selected from textbooks, poems, essays, stories, biographies, papers, medical works and other texts from primary school to eighth grade in Tibetan areas of five provinces, with a total of 22 articles, 5454 sentences, 41,950 dependency

relationships, and 47,404 words. Table 2 is the Tibetan dependency marker set formulated on the basis of the corpus.

Table 2. Sets of Tibetan dependency markers

Dependency relationship	tag	Dependency relationship	tag
Case subject predicate	SPV	Verb complement relation	VCM
Pass case subject-predicate	SVI	Preposition attributive	ATT1
Object relation	OBV	Postpositive attributive	ATT2
Object relation	GOV	Coordinating relation	COO
Apposition relation	APP	Position relation	NNF
Reference relation	QUO	Number association	MKM
Modal relation	VAU	Paracomotive relation	ADV
Figurative relation	FIG	Place relation	LOV
Interrogative relation	QUV	Source relation	SOV
Topic relation	TOV	Exclusion relation	EXV
Temporal relation	VTE	Comparison relation	COV
Consequent relation	LNK	Moment-time relation	TIV
Parallel relation	AKA	Negative relation	NEG
Symbol relation	WK	Manner relation	FSV
Successive relation	PK/CK/UK	Causal relation	CAV
head	ROOT	Goal relation	BEV

2.3   Research Methods

In the dependency tree database corpus of this paper, there are 41,950 dependency relationships. This paper uses Python language to extract the dependency relationship between "dominant words" and "subordinate words" from the Tibetan dependency tree database, and uses Excel software to make a statistical analysis of the dependency relationships through which Tibetan nouns dominate subordinate words, and the dependency relationships by which nouns are dominated by dominant words.

3   Result Analysis

In the dependency syntax tree, a word can dominate either one word or several words at the same time. This paper studies the main functions of Tibetan nouns in syntactic components based on the statistics of the number of times that Tibetan nouns are dominated by dominant words.

3.1   Syntactic Function Distribution of Tibetan Nouns

In the Tibetan dependency tree database constructed by the text, this paper counts the frequency of dependencies between Tibetan nouns and dominant components when they serve as subordinate and dominant words. Due to the convenience of statistical research, this paper only lists dependencies with frequencies of more than 100. Table 3 shows the frequency, proportion and example sentences of dependencies between nouns and dominant words:

Table 3. shows the frequency, proportion and example sentences of the dependency between subordinate and dominant nouns

Subordinate word	Frequency	proportion	Example	Dependency structure	Characteristic of a certain word
OBV	3452	8.23%	བཀྲ་ཤིས་ཀྱིས་ལམ་བུ་ཕྱིན།	ལམ་བུ（w）ཕྱིན(g)	n(w)+v(g)
TIV	1005	2.4%	ཚོགས་པ་སྐོབ་གྲུར་འགྲོ།	ཚོགས་པ(w)འགྲོ(g)	n(w)+v(g)
GOV	972	2.32%	ཁོར་དྲིས་པ།	ཁོ(w)ར་དྲིས་པ(g)	n(w)+v(g)
SPV	857	2.04%	སྨུག་གིས་ལུ་གུ་ཞིག་ཆྱེར།	སྨུག(w)ཆྱེར(g)	n(w)+v(g)
ATT2	822	1.96%	རྒྱལ་པོ་ཡང་སྐྱབ་གཉྱག་ལས་ཟིན།	སྐྱབ(g)གཉྱག་པ(w)	n(g)+n(w)
COO	721	1.72%	ར་ཁང་དང་ཁྱི་ཁང་ཟན་ད་སྐྱེལ།	ར་ཁང(w)ཁྱི་ཁང(g)	n(w)+n(g)
ADV	250	0.6%	མཚོན་ཆ་རྣམས་དུས་གུ་མང་པོར་ཚགས།	དུས་གུ(w)ཚགས(g)	n(w)+v(g)
TOV	185	O.44%	གཤོག་ལག་ཁང་ནི་དུས་རབས་བདུན་པར་བཞེངས་པ་ཡིན།	གཤོག་ལག་ཁང(w)བཞེངས(g)	n(w)+v(g)
SOV	135	0.32%	གནམ་ནས་ཆར་འབབ།	གནམ(w)འབབ(g)	n(w)+v(g)

Note: In Table 3, the symbol g represents the dominant word and w represents the subordinate word.

As can be seen from Table 3, when nouns are subordinate words, except for the core words of ATT2 (postattributive) and COO (parallel relation), which are nouns, the dominant words of other dependency relations are all formed by verbs, which shows that verbs have the most powerful valency ability in Tibetan. Firstly, the most syntactic component of nouns in the dependency tree is object, OBV (object relation) and GOV (object relation) account for 10.55% of the total tree. The second is attributive, ATT2 (postattributive) and COO (parallel relation) account for 3.68% in the total tree library; Thirdly, adverbial components, TIV (temporal relation), ADV (auxiliary dynamic relation) and SOV (source relation) accounted for 3.32% of the total tree library. Finally, subject component and topic component accounted for 2.04% and 0.44% of the total tree library respectively.

In addition to serving as the main syntactic components such as subject, object and attribute, nouns can also serve as the components of metaphor, question and complement. Although there is no relevant literature on the syntactic function of Tibetan nouns, the function of Tibetan nouns as syntactic components is the same as that of Chinese nouns, which mainly serve as subject, object and attributive components. It can be seen that nouns have these major syntactic functions in any language.

3.2   Probabilistic Valency Model of Tibetan Nouns

According to the dominant power of Tibetan nouns and the number dominated by other dominant words, this paper calculates the binding force of Tibetan nouns, and finally gives the probability valence mode diagram of Tibetan nouns in the form of a graph. Figure 1 shows the probability valence mode diagram of Tibetan nouns:
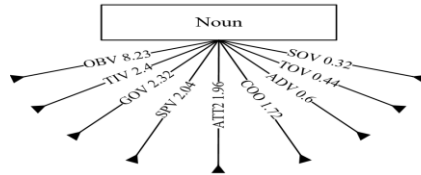

Fig. 1. Probability valency pattern of Tibetan nouns

It can be seen from Figure 1 that the valency ability of Tibetan nouns is very strong, and Tibetan nouns can be either dominant or subordinate words. In the figure, the inward arrow indicates the dependency ratio (%) between the noun as a subordinate word and the dominant word, and the binding force of the noun is centripetal force. As in ལས་བྱ་བྱེད།, the noun ལས་བྱ is a subordinate of the verb བྱེད. According to the above data, we can divide nouns into the main syntactic function and the secondary syntactic function. The main syntactic function of nouns is to act as subject, attribute and object. The secondary function is adverbial and other syntactic components.

4   Concluding Remarks

This paper makes a quantitative analysis of the syntactic functions of Tibetan nouns on the basis of the dependency syntax tree of Tibetan, and draws the conclusion from the dependency data of nouns as dominant and subordinate words that in the dependency relations of Tibetan nouns as dominant words, the premodification and postmodification are the majority, while when nouns as dominant words dominate other words, the nouns are mostly gerunds.

References

1. LIU Dingjia, ZHANG Ziyan. Development and Applications of Syntactically Annotated Corpora: Retrospect and Prospect[J]. Technology Enhanced Foreign Languages, 2022 (06): 82-88+113.
2. Quirk, R,. et al, 1985, A Comprehensive Grammar of the English Language [M],London: longman.
3. Wang Yong, Liu Haitao. Quantitative Properties of Rusian Nouns[J]. Journal of Zhejiang University (Humanities and Social Sciences) ,2013,43(06):174-186.
4. GAOSong. A Quantitative Study on Syntactic Functions of Nouns in Mandar in Chinese : Based on Chinese Dependency Treebank[J].TCSOL Studies,2010(02):54-60.
5. Haliday, M, A,K,1994, An Introduction to Functional Grammar [M],London :Edward, Arnold.