2012 International Conference on Image, Vision and Computing (ICIVC 2012)
IPCSIT vol. 50 (2012) © (2012) IACSIT Press, Singapore
DOI: 10.7763/IPCSIT.2012.V50.54

On the Investigation of the Using of Suffix of Modern Uyghur Written Language in Website

Yusup Abaydulla[1+], Xiangwei Qi[1] and Hasanjan Abliz[2]

[1]School of Computer sciences and technology, Xinjiang Normal University, Urumqi, China
[2]School of history and ethnology, Xinjiang Normal University, Urumqi, China

Abstract. In the paper, the definition and explanation of suffix of modern Uyghur written language explained first; the website resource of the corpus for the investigation, application area, the statistic method and time span of the corpus introduced then; third, the distribution of the suffix as by the first ten thousand are explained by the results and statistics of the suffix frequency in section.

Keywords: Modern Uyghur written language, website, suffix, statistics

1. Introduction

The Uyghur language belongs to the Turkic language family of the Altaic language family, it belongs to agglutinative language and its characteristics as vowel harmony, vowel reduction (or vowel rising), the rich of Word-formation and inflectional affixes, an inflection form of noun case, number and person, the various indication of verb, positional relation and constituent in order. The grammatical structure of modern Uyghur Language consists of roots, stem, affix and suffix. Affix has a function of changing word meaning and suffix also has a grammatical meaning.

2. Suffix Divided Method

2.1. The Resource of the Corpus

The corpus of Uyghur language of the investigation are collected from the words of 9 Uyghur websites which is closely related with the area of the daily life of people such as politics, economy, education, scientific research and health displayed in website from april 2006 to december 2009. The investigation of the object is the suffix of 9 Uyghur websites,and the research strictly accords with the standard of the syllable classification of orthography (1983) of Minority Language Committee of Xinjiang Uyghur Autonomous Region. It is used the method of the combination of auto-processing data with computer and Artificial supplementary revised text for statistical analyzing suffix of modern Uyghur written language.

2.2. Word-Formation Structure

The grammatical structure of modern Uyghur Language consists of roots, stem, affix and suffix. Affix has a function of changing word meaning and suffix also has a grammatical meaning. See below a model of word-formation of Modern Uyghur language, which can bring suffix divided as principle basis.

+ Corresponding author. Tel.: +15026087856.
E-mail address: 47266861@qq.com.

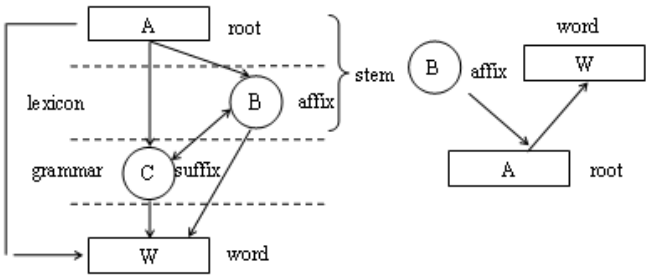

Fig. 1 A model of word-formation of Modern Uyghur language

2.3. The Method of Words Restore

It has finished suffix divided on the basis of the model of word-formation of Modern Uyghur language, the corpus of stem and suffix and the method of the combination of auto-processing data with computer and Artificial supplementary revised text for statistical analyzing suffix. The suffix divided progress and method as shown in following steps:

First, building the corpus which is closely related with the area of the daily life of people such as politics, economy, education, scientific research and health.

Second, entering whole suffix (not included affix) in modern Uyghur language into suffix corpus.

Third, it has bring some difficult to computer analysis of stem corpus in which shoud be select stem for analysing suffix on account of pronunciation reduction.

Fourth, in view of the foregoing, suffix divided method make use of computer analysis in consideration of computer ability as capability, speed and credibility.

Fifth, the above rules of the computer analysis is also applied for suffix divided which can be selected from suffix corpus in spite of occuring suffix original and its reduction.

Sixth, according to the requirement of the suffix statistics, build the suffix statistical system.

3. The Using of Suffix

The suffix of Uyghur words is various, especially the suffix of verb, which has a grammatical meaning. It is focused the suffix of noun and verb in the investigation, but it is not included frequent symbols such as Arabic numbers from 0 to 9, Percent(%) ,bracket (“ , ”)and units($,℃) and other symbols(invisible characters, and space characters). It is used the method of the combination of auto-processing data with computer and Artificial supplementary revised text for statistical analyzing suffixes of 197 649 words. On the statistics of the suffix, it can be occurred that word type as 197 649, suffix frequency as 118 848 and suffix type as 4448.

3.1. The Suffix Frequency More Than Ten Thousand

The 32 suffixes which is frequently occurred more than ten thousand as shown in table I

Table 1 32 kinds of suffix occurred in the frequency of 10,000 times

Suffix	Frequency	Suffix	Frequency	Suffix	Frequency
نى	194 595	دۇ	55 657	چە	25 168
نىڭ	172 251	ۋې	44 938	تىن	25 217
كى	96 199	دىكى	48 311	دىغان	24 938
دە	91 243	لار	44 330	تە	23 903
دا	84 646	ۋش	41 600	قان	21 979
غان	81 867	قا	38 215	ۋش	21 229
دى	78 132	لىرى	31 363	گەن	18 517
سى	76 036	لەر	28 206	ىدىغان	18 144
تى	71 463	گەن	27 441	مۇ	12 898
غا	65 773	كە	26 884	تىكى	10 541
گە	65 128	تا	25 487		

It can be seen from the table above that the suffix displayed three functions as:

First, the double functions of the suffix. for example, the suffix of "-چە" has double functions in Uyghur language as suffix and affix. When "-چە" adds to the words of سىلەر [you] as suffix, When it adds to the words of ئۇيغۇر [Uyghur] as affix.

Second, the same suffix occurs multi-grammatical functions. For example, the "-مىز" has two grammatical functions in Uyghur language as مىز + بالا [child] of noun person, and as مىز + باشلاي [I start] = باشلايمىز [we will start] of verb person.

Third, it is displayed that the suffix of noun occurred higher more than the suffix of verb based on the data of corpus.

3.2. The Analysis of the Suffix Frequency in Section

The analysis of suffix frequency, word type and average length as shown in table II.

Table 2 Word type frquency in section

frequency	word type	frequency in all	average length	frequency	word type	frequency in all	average length
1	1405	1405	8.3218	6-10	550	4185	6.9164
2	711	1422	7.8284	11-20	340	4942	6.1709
3	443	1329	7.4650	21-100	346	14413	5.3092
4	317	1268	7.3028	>100	117	88789	3.9573
5	219	1095	7.0868				

In the table above, we can see that low frequency of words in corpus are mass, and it is gliding from one to five in frequency but is going up from six to ten. It shows that quantity of low frequency words are mass and the corpus content involved is wide in website words.

3.3. The Analysis of the Suffix Length in Section

The analysis of the suffix length in section as shown in figure 2.

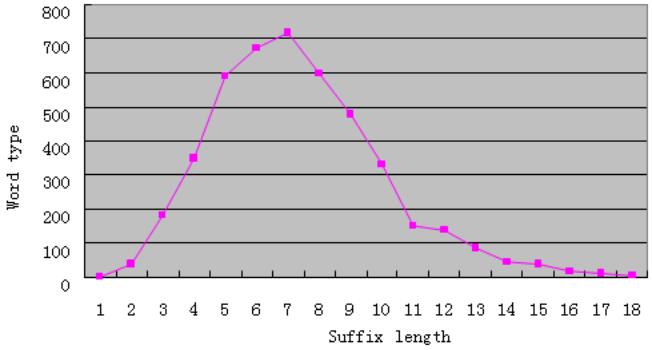

Fig. 2 The analysis of the suffix length in section

In the table above, we can see that the suffix length from one to three and from 13 to 18 are frequently used in website on the basis of the analysis of the one to three points, climax at seven point and four suffix only at 18 point.

The analysis of the relation of the suffix length and its frequency as shown in figure 3.

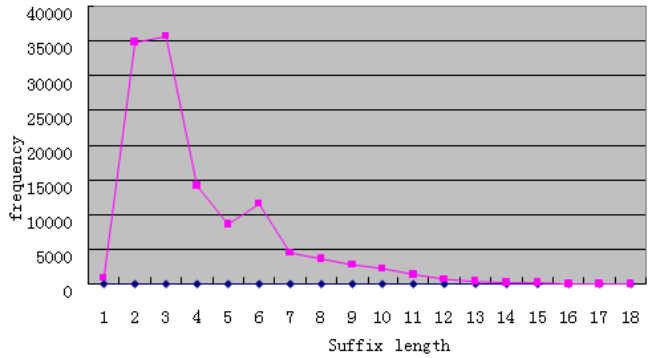

Fig. 3 The analysis of the relation of the suffix length and its frequency

In the table above, we can see that the suffix length at one point with its frequency as 941, at two point with its frequency as 37489, at five point with its frequency as 593, and at seven point its frequency as only 4.the table shows that more than the suffixes at 13 point are lower in website words in the corpus.

4. Acknowledgment

This paper is funded by the Ministry of Science and Technology project "the key technology research and demonstration application of Tibetan / Uyghur language resource monitoring"

5. References

[1] Hamit Tomur. Modern Uyghur grammer. National press, (6)1987.

[2] Abliz Yakup. Modern Uyghur explanatory dictionary.national press,[11]1991.

[3] Yusup Abaydulla.on the relevant dispose of central word driving syntax. Computer application and software, [6]1999.

[4] Wang tie Kun, Zhang pu.Language Situation in China: 2005(2).The Commercial Press.

[5] Wang tie Kun, Zhang pu.Language Situation in China: 2006(2).The Commercial Press.

[6] Wang tie Kun, Zhang pu.Language Situation in China: 2007(2).The Commercial Press.

[7] Wang tie Kun, Zhang pu.Language Situation in China: 2008(2).The Commercial Press.

[8] Yi kun xiiu, Gao shi jie.Uyghur grammer.Minzu University of china press.(2) 1998.

[9] Cheng shi Liang. Modern Uyghur grammer.Xinjiang people press.(9)1996.