Journal of Chinese Language and Computing 15 (4): (203-210)

The Research and Development of Computer Aided Contemporary Uighur Language Tagging System

Yusup Abaidula，Rezwangul，Abdiryim Sali
Department of Computer Science, School of Physi-Math Information
Xinjiang Normal University, Urumqi, Xinjiang, China 830054
yusup2002@sohu.com

Abstract

The research on the contemporary Uyghur information processing in the 20th century can be dated back to the beginning of 1980‘s. There are a lot of achievements up to now ， for example, fundamental theory and basic utility are established, various corpus are built， code standard is designed, Uyghur grammatical attribution research is carried out。Because the Uyghur language has its own nature，the Uyghur information processing has its own characteristics and methods。This paper briefly states the research work on Uyghur morphological analyzer and the computer-aided contemporary Uyghur corpus processing system[1]。

Keywords

Uighur language, Text processing, Tagging system, Corpus

1. Preface

Contemporary Uyghur information processing includes the input, output, recognition, and understanding of this language. In the course of analyzing the Uyghur characters, words, sentences and paragraphs, computer is used to process the Uyghur phonetics, morphology, semantics etc. Uyghur language has its own character set, so they have its own nature and needs different processing methods. From the 1980’s to now, a lot of work on the character level process of the language has been done. Now, more researches are focus on the processing of words and sentences. We also had some progress on this direction.

[1] This research is sponsored by the programs No. 60163002 and No. 60463005, National Natural Science Foundation of China (NSFC), and the Startup Fund of the Special Training Program for the Minority Talents by the Science Department of Xinjiang Uyghur Autonomous Region.

2. The subsystem for Uighur file format conversion

The first work of this research is to prepare the electronic text data for contemporary Uyghur tagging corpus. This work is done with the support from some publishers, such as “Xinjiang Daily”, Xinjiang Science and Technology publisher, Xinjiang Education Publisher, Xinjiang People’s Publisher. They provided us compiled electronic Uyghur language materials. Since these text materials are only suitable for DOS system, not for Windows. So we have to convert the format of the text. Our system has a module to for this conversion. By using this system, we carried out the format transformation of the Uyghur materials, and worked out a corpus of more than 8 million words.

These materials cover almost all the words in the text book of the elementary, junior, high, technical secondary schools and university as well as words used in the fields of law, literature, health, technology (especially the agriculture technology), history, and economics. This contemporary Uyghur language corpus is the basis for our further information processing. The information processing not only needs to summarize and define surface structure rules but also deep structure analysis. Next step, we are going to do statistics and processing for 4 million phrases. We divided the corpus into four parts, which are religious 20%, popular science 20%, political commentary 20%, and literature 40%.

3. Uighur word Tagging Standard

The most important work in setting up Uyghur tagging corpus is to create Uyghur tag set and standard. In word tagging, we started with a small tagging set. From the viewpoint of computer linguistics, and by reference to Contemporary Uyghur grammar, we added some tagging symbols. Finally we have 24 tagging types.

The tagging symbols of noun are up to 31 ,for example:

- Personal name Nk كىشى گىسمى
- Place name Ny يەر نامى
- Organ name Nt گورگان-تەشكىلاتلار نامى
- Time name Nw ۋاقىت گىسملىرى

According to the semantic and grammatical function, verbs are classified as personal, non-personal and auxiliary. According to the tense of the verb, it is divided into more than 80 types, for example:

- Verb V پېئىل
- Personal verb V1 شەخسلىك پېئىللار
- Imperative verb V101 بۇيرۇق تەلەپ رايى

In Uyghur language, non-personal verb differs from personal verb in their grammatical function. The grammatical function of the non-personal verb is similar to the nouns, the adjectives and the adverbs. If the grammatical function is similar to the nouns, it is called gerund. If the grammatical function is similar to the adjectives, it is called verbal. If the

grammatical function is similar to the adverbs, it is called adverbial. According to their basic function, gerund is tagged by Vn, verbal is tagged by Va, adverbial is tagged by Vd.

- Non-personal verb V2 شەخسسىز پېئىللار
- Verbal V21 سۈپەتداش
- Perfect participle V211 پۈتكەن ھاللىق سۈپەتداش

In Uyghur information processing, we created a small tagging set [9] and used it in the tagging of the Uyghur words.

4. Uighur Electronic Information Dictionaries

Our Uyghur Electronic Information Dictionary includes word-root dictionary, phrase dictionary and suffix dictionary. The vocabulary in the word-root dictionary phrase dictionary has about 60,000 items. Word dictionary includes 25,000 words. Every word at present has 16 attributes. In the word-root dictionary, the word-root structure and word-root format is as in Table 1.

Attribution / Words	سۆز تۈركۈمى part of speech	ئىسىم ئائىلىسى noun system	سۆز مەنىسى semantic	ئېتىمولوگىيىسى etymology	كېلىش case	تەۋەلىك dependence/ ownership	سان number	دەرىجە degree	پېئىل دەرىجىلىرى verbal voice	بولۇشلۇق بولۇشسىزلىق Positive/ negative	ئىمكان ability/ non ability	شەخس personal individual	ھال زامان، رايى tense	گرامماتىكىلىق رولى grammatical function	چاستوتىسى frequency
سۇ water	n	N3	سۇيۇقلۇق	1	1	0	1	0	0	0	0	0	0	6	4396
چوڭ big	a	0	يوغان	1	1	0	1	6	0	0	0	0	0	6	7775
سەن you	r	0	سۆزلىگۈچى	1	1	0	1	0	0	0	0	1	0	6	4728
تېز harry	d	0	ئىلدام	1	0	0	0	0	0	0	0	0	0	5	2081
ئون ten	m	0	ساناق سانلىق بىرسى	1	1	0	1	0	0	0	0	0	0	6	2260

Table 1. Example of word-root dictionary

Attribution / Words	سۆز تۈركۈمى part of speech	[illegible] noun system	سۆز مەنىسى semantic	ئېتىمولوگىيىسى etymology	كېلىش case	تەۋەلىك dependence/ownership	سان number	دەرىجە degree	پېئىل دەرىجىلىرى verbal voice	بولۇشلۇق بولۇشسىزلىق Positive/negative	ئىمكان ability/nonability	شەخس personal individual	ھال زامان رايى tense	گرامماتىكىلىق رولى grammatical function	چاستوتىسى frequency
نى	N	0	0	1	4	0	0	0	0	0	0	0	0	4	220997
ىم	N	0	0	1	1	1	1	0	0	0	0	0	0	6	210007
دى	V	0	0	1	0	0	3	0	1	1	2	3	v111	2	111976-
ۋاتىدۇ	V	0	0	1	0	0	3	0	1	1	2	3	v161	2	963
ۋاتىمىز	V	0	0	1	0	0	2	0	1	1	2	1	v161	2	38
راق	V	0	0	1	1	0	0	8	0	0	0	0	0	3	1280
الايمەن	V	0	0	1	0	0	1	0	1	1	1	1	v144	2	125
دۇر	V	0	0	1	0	0	1	0	3	2	0	2	v111	2	34

Table 2 Example of the suffix dictionary

The structure of the machine electronic phrase dictionary and the format of the phrase is similar to those of word-root dictionary. Electronic suffix dictionary at present has 130,000 suffixes. Every suffix has 16 attributes. The electronic suffix dictionary and the format of suffix is as in Table 2:

5. The subsystem of the Uighur corpus management

The Uyghur corpus is not only a collection of text. It also contains a lot of original and structured texts as part of the corpus. Some structure information is saved in the corpus, for example, chapter name, paragraph name and its location. The function of this information is not only for searching text according to certain standard, but also a part of the corpus.

6. Uighur word tagging and segmentation rules

The words segmentation rules we used follow *Contemporary Uyghur theography dictionary* .At this point, the purpose of the Contemporary Uyghur word segmentation is to do word tagging. We produced a contemporary Uyghur electronic dictionary. This dictionary includes word-root dictionary and word -suffix dictionary, and is being used as a reference for segmenting word-root and suffix.

The segmented units are basically word-roots, word-affixes and word- suffixes. They are the basic units that are used in the information processing, and it has certain semantic and grammatical functions.

According to the statistics on the corpus, we found out the rules for forming words. IN the following examples: U stands for sentence word, A stands for word-root, B stands for word-affix, C stands for word- suffix . The rules of words forming are as in Figure 1.

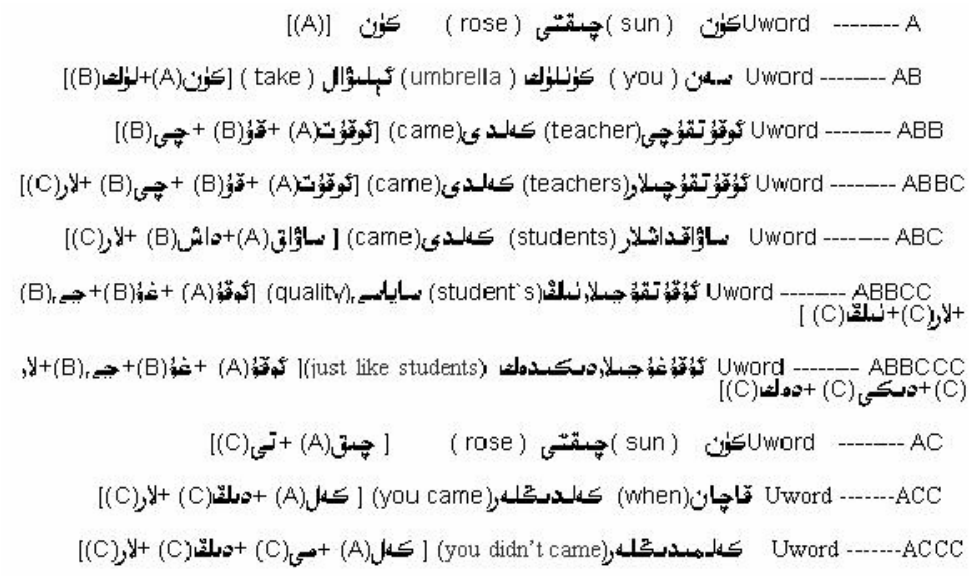

Figure 1. The rules for forming words

On the basis of above-mentioned concepts, we have the following four rules for sentence elements formation, they are:

Uword -------- A كۈن(sun) is word class(word root)

Uword -------- AB كۈنلۈك (umbrella) is word root + word-formation affix

Uword ----- ABC ساۋاقداشلار (students) is word root + word-formation affix+ word-inflection suffix

Uword -------- AC چىقتى(rose) is word root + word-inflection suffix

Figure 2. The rules for forming sentence.

7. The sub system for the character statistics

Character frequency statistics subsystem is composed of three modules: text reading module, statistic computing module, and statistic data management module. The language of our program is Borland C++ Builder, the background database is Microsoft access 2000.When we analyzed the statistic results, we found that statistics effectively reflected the frequency of Uyghur letters in the Uyghur text. Through statistics of 4 million words, we obtained the frequency of the contemporary Uyghur characters. The frequency of contemporary Uyghur character is as shown in Figure 3.

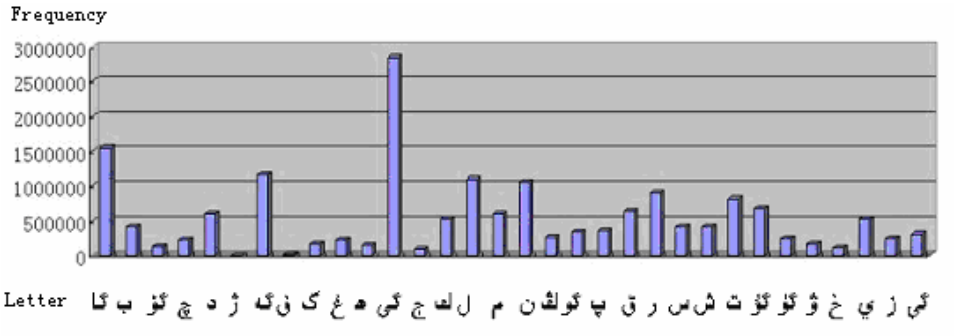

Figure 3. Frequency of contemporary Uighur characters

We can see, the most frequent one appeared 2842094 times, which is about 17% of all the occurrences. The second one appeared 1547472 times, which 9% of the whole occureances. The third one is about 7%. The fourth one about 6%.The least used one appeared 2223 times, which is 0.0129%. The second least appeared 11079 times, or 0.064%. The third least used one appeared 84583, or 0.49%.

8. The subsystem of the syllable statistics

Syllable frequency statistic subsystem is composed of three modules: text reading module, syllable statistic compute module, and syllable statistic data management module. The program language we used is Borland C++ Builder, the background database is Microsoft access2000. Trough statistics of the 4 million words' real corpus , we found 7000 syllables in Uyghur. By analyzing the statistic results, we found the frequency of the 7000 syllables in Uyghur. Among them, 6065 syllables appeared more than one time. 3730 syllables appeared over three times. 1682 syllables appeared 100 times. 138 syllables appeared 1000 times, and 9 syllables appeared over 110000 times. The 9 most frequent syllables and their frequencies are as shown in the Figure 4..

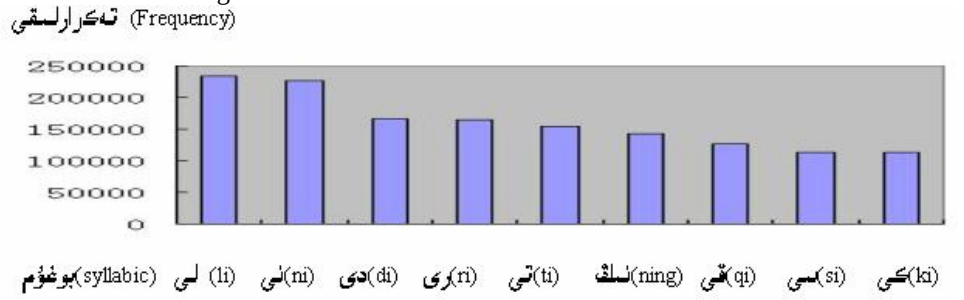

Figure 4. Frequency of contemporary Uighur syllables

9. The subsystem of the word statistics

In this real corpus of 4 million words, we found 240,000 unique words. At the same time, among this real corpus, we found that the words that appeared more than two times occupied 44%. The rest appeared only once.

The original electronic dictionary has about 120,000 word-roots, but in the real corpus of 4 million words, 140,000 word-root appeared. Only about 20%of the 10,000 suffixes that we had appeared two or more times. The rest in this corpus appeared only once or did not appear at all. This explains that only 20% suffixes are commonly used in the written language. The rest are used in spoken language.

10. The sub system of Uighur tagging

When building the Uyghur corpus, we also developed an automatic word tagging system based on the Uyghur information electronic dictionary, and the Uyghur suffix structure. In addition, to improve the correctness of the electronic word tagging, we created a tool to manually correct the corpus. Through these methods, we can guarantee the correctness and consistance of tagged results. The function of these auxiliary tools includes tagging, restricting the dictionaries and online checking of the compatible words (antonyms or synonymous). It also makes corrections using some rules. Nowadays, the subsystem of the Uyghur tagging is under testing.

11. Conculusion

Through nearly 20 years of development, Xinjiang minority nationality information processing work obtained some promising results. However, compared to other languages , it still needs a lot of work, especially on the basic theory studies and basic application technique development. As far as Uyghur grammar and semantic research of information processing are concerned, it is still in the primary stage. It still does not have its own theoretical system. Also, this is a very big project that needs too much investment. This is also the main obstacle in the other minority nationalities information processing research. Apart from these, Uyghur information processing also has a lot of work to be done. For example: Uyghur grammar information dictionary, reference database for Uyghur phonetics, linguistics studies of Uyghur corpus, Uyghur grammar attribution studies of information processing, Uyghur semantic studies of information processing, Uyghur machine translation etc. I hope the research of the Uyghur language information processing will get more and more support from researchers and officials.

References

[1] Yusup Aibaidulla and Kim-Teng Lua, The development of Tagged Uyghur Corpus, Proceedings of PACLIC17, 1-3 October 2003, Sentosa, Singapore, P228-234.
[2] Yusup Aibaidulla etc. Contemporary Uighur Corpus Managing, China Artificial Intelligent Developing 2003, Beijing Post University Press, P1007-1010.
[3] Yusup Aibaidulla, Semantic Problem Solution Research of Uighur Sentence Grammar Analyzer, Computer Application and Software, Vol 4. 2002, P59-62.

[4] Yusup Ebeydulla, Progress In System Design of Contemporary Uyghur Corpus. Journal of Chinese Language and Computing, Volume 13, 2003, P283—301.
[5] Yusup Ebeydulla, An Uyghur Syntax Parser. Journal of Chinese Language and Computing. Volume 13, 2003. P273-282.
[6] Yusup Ebeydulla，Askhar, The transform Subsystem of Uighur File Style of the Contemporary Uighur Corpus System, Journal of Xinjiang Normal University（Natural Science Version）vol. 1, 2003, P16-19.
[7] Yusup Abaidula，Abdiryim Sali, Discussion on the Corpus Linguistics, Journal of Xinjiang Normal University（Social Science Version）, Vol. 4, 2003, P76-79.
[8] Yusup Abaydul Research on System of Contemporary Uyghur Word Frequency Statistics and High Frequency Words, Procceedings of the International Conference on Chinese Computing 2005, 21-23 March 2005 Singapore.
[9] Xinjiang Normal University Information processing with contemporary Uyghur words tagging sets standard version 1.0.