Language Resource Construction for Mongolian

Shipeng Xu[1] , Hongzhi Yu[1], Thomas Fang Zheng[2] and Gegeentana[1]
[1]Key Laboratory of National language intelligent processing, Gansu Province, Northwest Minzu University, Lanzhou, China
E-mail: xspxkrs@aliyun.com Tel: +86-18394170747
E-mail: yuhongzhi@hotmail.com Tel: +86-13909318273
E-mail: Gegeentana@foxmail.com Tel: +86-13893303181
[2]Center for Speech and Language Technologies, Tsinghua University, Beijing, China
E-mail: fzheng@tsinghua.edu.cn Tel: +86-13801012234

***Abstract*— Mongolian is a typical low-resource language. The resource limitation is in various aspects, from acoustic analysis, phonetic rules, lexicon, speech and text data. This paper describes our recent progression on Mongolian resource construction supported by the NSFC M2ASR project. Firstly, we collected the text data of Mongolian containing more than 60,000 sentences from the newspaper, internet and Mongolian books. Secondly, we built the initial dictionary of Mongolian based on the *Mongolian Chinese Dictionary*. All the resources are published following the M2ASR Free Data Program.**

I. INTRODUCTION

Mongolian is a nomadic nation mainly distributed in the east region of Asia. There are more than 10 million Mongolians in the world. It is the main nationality of the State of Mongolia, and in China, Mongolians is one of the major minority nations, with a population of more than 6 million people. Most of them live in the Inner Mongolia Autonomous Region, others are distributed in the neighboring provinces, including Jilin, Liaoning, Heilongjiang, Xinjiang, etc [1].

Mongolian belongs to the Altai language family, Mongolia branch [2]. It is the main communication tools of Mongolian people. Mongolian in China can be divided into three dialects: the inner Mongolia dialect (or the central dialect), the Baltu-Buryat dialect (or the northeast dialect) and the Weilate dialect (or the west dialect). The inner Mongolia dialect is most widely used compared to the other two dialects.

Mongolian is not only a practical language, but also an object of academic research, e.g., a tool to study the historical development of Mongolian and Altai nations and languages. Therefore, many people learn and study Mongolian in the world. As an example, the International Conference of Mongolian Scholars was held in Ulaanbaatar every five years [3]. In China, many institutes established specific research labs to study Mongolian, such as the Institute of Ethnology and Anthropology, Chinese Academy of Social Sciences [4] and the Inner Mongolia University [5].

In spite of the great importance of Mongolian, the research on this language largely lags behind the major languages such as Chinese and English. A key reason is that the linguistic and speech resources are very limited, and the existing resources are far from open and standard. Most of research institutes publish research based on their own private data, which makes the results incomparable and hence doubtable. In this paper, we will describe our work on Mongolian language resource construction, and release two resources for public usage: a text corpus that consists of 60k Mongolian sentences, and a Mongolian-Chinese dictionary that consists of 25000 Mongolian words. These resources are valuable for multiple studies, including acoustics, linguistics and speech processing.

Our work is part of the Multilingual Minor lingual Automatic Speech Recognition (M2ASR), which is supported by the National Fundamental Science of China (NFSC). The project is a three-party collaboration, including Tsinghua University, the Northwest National University, and Xinjiang University. The aim of this project is to construct speech recognition systems for five minor languages in China (Tibetan, Mongolia, Uyghur, Kazak and Kirgiz). However, our ambition is beyond that scope: we hope to construct a full set of linguistic and speech resources for the 5 languages, and make them open and free for research purposes. We call this the M2ASR Free Data Program. All the data resources, including the ones published in this paper, are released through the website of the project http://m2asr.cslt.org.

This paper is organized as follows. Section 2 presents some basic knowledge of Mongolian. Section 3 describes the collection of the text resource. The dictionary construction is presented in Section 4, and some concluding remarks are given in the final section.

II. CHARACTERS OF MONGOLIAN

A. Alphabets

The Mongolian writing system is alphabetic. It contains 35 alphabets, including 8 vowels, 17 basic consonants and 10 preposition consonants [6]. The alphabets involve a lot of variants. For example, depending on the position (top, middle and bottom) of the alphabet within the word, the alphabet may change. Moreover, the same word may pronounce differently in different contexts. Table I shows the 35 Mongolian letters defined by the GB 25914-2010 national standard [7]. Note that we have defined a Latin form for each Mongolian letter, which eases the text processing by computers.

TABLE I MONGOLIAN ALPHABET

vowel	Latin	consonant	Latin	consonant	Latin	consonant	Latin
ᠠ	a	ᠨ	n	ᠱ	š	ᠺ	k
ᠡ	e	ᠩ	ng	ᠲ	t	ᠯᠺ	lk

	i		b		d		c
	o		p		č		z
	u		x		ǰ		h
	ö		g		y		zr
	ü		m		r		lh
	ee		l		w		zh
			s		f		ch

B. Syllable

Syllable is the unit of Mongolian pronunciation. A syllable may consist of one or several phonemes. Most of the syllables are structured in one of six forms: V, VC, VCC, CV, CVC and CVCC, where V represents a vowel, and C represents a consonant. Taking account the foreign words, the possible syllable structures could be richer. Table Ⅱ shows the typical syllable structures.

TABLE Ⅱ MONGOLIAN ALPHABET

Type	Word	IPA	Type	Word	IPA
V		ɑ	CV		tɑ
VC		æl	CVV		bi:
VV		ɑɪ	CVC		nɑr
VVC		ɑ:b	CVVC		næædz
VCC		ɑlt	CVCC		nɛbʧ
VVCC		ɑ:ldʒ	CVVCC		ny:rs

C. word

Mongolian words can be categorized into 13 classes: noun, adjective, quantifier, time and location, pronoun, verb, adverb, modal, imitated words, postposition, modal particle, conjunction and emotional words. Table III presents some examples.

TABLE Ⅲ MONGOLIAN WORD TYPES

Type	Example	Mean	Type	Example	Mean
noun		daytime	modal		about
adjective		good	Imitate d woeds		ha ha
quantifier		one	postposition		same
time and location		south	modal particle		query
pronoun		i	conjunction		and
verb		eat	emotional		Wake up
adverb		fairly			

III. TEXT CORPUS

We collected a Mongolian text corpus that contains more than 60,000 sentences. The data sources involve newspaper, internet and books, and the sentences are in various domains. The total number of words is about 35,000. The shortest sentences only contain one word and the longest sentences contain more than 200 words. The coding of the text is UTF-8. The text file is 14.7 M in the disk.

A number of language-specific normalization steps were conducted. The purpose was to convert non-standard words to standard ones. In the Mongolian text, non-standard words are the words containing non-Mongolian characters, such as Arabic numbers, English characters and other functional symbols (e.g., ^, $). We found that there were about 2500 sentences that contain non-standard words. Most of the non-standard words are numbers, such as time and telephone numbers.

IV. MONGOLIAN-CHINESE DICTIONARY

The main contribution of this study is that we constructed a Mongolian-Chinese dictionary that more than 25k words. As far as we known, this is the first open and free electronic dictionary for Mongolian-Chinese pairs. This resource will greatly benefit related research, for example speech recognition.

A. Building the initial dictionary

We choose the *Mongolian Chinese Dictionary* as the initial dictionary for our study. This dictionary is very famous and has a great impact on the study of Mongolian. A particular reason to choose it as our initial resource is that this dictionary gives the IPA spelling of each word.

We manually input the paper-version of this dictionary into computer to obtain the electronic version. The input was reviewed by several Mongolian students to ensure the quality. This leads to our initial dictionary that contains more than 25,000 words. For each word, the information includes the prototype, the Latin form, the API pronunciation and the Chinese meaning.

B. The conversion of IPA

The IPA table is a general and standard representation of pronunciations of a language, whoever using IPA directly is not convenient as the letters used by IPA are not easy to be processed. Therefore, we designed a conversion rule that change IPA letters to simple Latin character compositions. This new representation can be called as NIPA. Table Ⅳ shows the conversion rule. Applying this rule, the IPA-based pronunciation is converted to NIPA-based pronunciation.

TABLE Ⅳ CONVERSION RULE FROM IPA TO NIPA

API	NAPI	API	NAPI	API	NAPI	API	NAPI
a	a	ə	e1s	l,	l2	ɳ:	i2l
ă	as	e:	el	m	m	t	t
a:	a1l	əː	e1l	n	n	ʧ	ch
æ	a1	əʳ	e1r	n,	n1	u	u
æ:	a12	f	f	ŋ	ng	ʊ	u1
aɪ	ai1	g	g1	o	o	u:	ul
aʊ	au1	g,	m	ö	o2	ʊ:	u11
b	b	i	i	ŏ	os	ui	ui
c	c	ɪ	i1	o:	ol	ʊɪ	u1i1
ɔ	a2	ɪː	i11	ø:	e3l	w	w
ɔː	a2l	i:	il2	œ	e4	x	x
d	d	iɛ	ie2	œ:	e4l	x,	x1
dz	dz	ɪʊ	iu1	p	p	ʏ	y1
ʣ	dzh	ɪʊː	iu11	r	r	y:	yl
dʒ	dg	j	j	ɳ:	i3l	y	y1y
e	e	k	k	s	s	ʐ	z
ɔ̆	a2s	l	l	ʂ	sh1	r,	r1
ə	e1	ɬ	ll	ʃ	sh	ɳ	i2

C. Dictionary coverage

With the initial dictionary, we checked the coverage of the dictionary for the text corpus built in the last section. The result shows only less than 6,000 sentences (about 10%) were completely covered by the dictionary. Further analysis shows that only about 5,000 words in the dictionary appeared in the text data. The reasons we found are mainly the following two:

1. The morphology rule of Mongolian is rather flexible, which results in many inflections that were not covered by the initial dictionary.
2. There are many foreign words (words borrowed from other languages, mainly from Chinese) in the text data, such as geographic name and name of person. These words cannot be well covered by the initial vocabulary.

D. Morphology rules for dictionary improvement

For the foreign words, it is possible to label them manually. We are doing so at present. For the flexible morphology, it is much more difficult. It is not possible to label all the inflections by hand, so an automatic approach is necessary. As the first step, we have collected a full set of morphology rules, and utilize them to generate the pronunciations and meanings of inflection words.

Table Ⅴ ~ table Ⅹ summarize some of the most important rules. Table Ⅴ presents the eight cases in Mongolian. The nominative, freeze-frame and objective cases are the most well-known. The Xiangwei case represents the relationship between indirect object and the predicate, as well as the relationship between the adverbial and the discourse. The Pingjie case and Congci cases can represent both indirect object and adverbial. The Hetong case describes the relationship between indirect objects and verbs. The Lianhe case appears very few in modern Mongolian.

TABLE Ⅴ MONGOLIAN CASE

Name	Additional ingredients	Usage	IPA
nominative			
freeze-frame	ᠨ yin ᠨᠦ un/ün ᠤ u/ü	ᠨ After the word ending with vowels; ᠨᠦAfter the consonant (except n) at the end of the word; ᠤ Only after the word ending with n consonants;	[i:n] [æ:] [ə:]
Xiangwei case	ᠳᠤ du ᠲᠤ tu	ᠳᠤ After the ending with vowels and l, m, n, ng consonants of the word; ᠲᠤ After the word ending with b, g, r, s, d consonant;	[d] [t]
objective case	ᠶᠢ yi ᠢ i	ᠶᠢ After the word ending with vowels; ᠢ After the word ending with consonants;	[i:] [iːg]
Pingjie case	ᠪᠠᠷ bar/bᠡᠷ iyar/iyer	ᠪᠠᠷ After the word ending with vowels; ᠡᠷ After the word ending with consonants;	[a:r] [ə:r] [ɔ:r] [o:r]
Congci case	ᠴᠠ ača/eče	ᠴᠠ After the word ending with both of vowels and consonants;	[a:s] [ə:s] [ɔ:s] [o:s]
Hetong case	ᠲᠠᠢ tai/tei	ᠲᠠᠢ After the word ending with both of vowels and consonants;	[tæ:] [tə:] [tœ:]
Lianhe case	ᠯᠤᠭᠠ luɣ-a ᠯᠦᠭᠡ lüge	ᠯᠤᠭᠠ After the positive word (compact vowel) ᠯᠦᠭᠡ After the negative word (loose vowels)	[læ:] [lə:]

Table Ⅵ presents the rule of plural forms. The plural form in Mongolian is different from that in English. The first four cases just need to add the suffix as a single word behind the target word. The other two cases append the suffix to the target word to form a new word.

TABLE Ⅵ MONGOLIAN PLURAL FORM

Suffix	IPA	Usage	Notes
ᠨᠤᠳ ᠨᠦᠳ	[nɔ:d] [nu:d]	After the noun describing people and things ending with vowels or n consonant;	Separate with the noun
ᠴᠤᠳ	[ɔ:d] [u:d]	After the consonant (except n) at the end of the word describing people and things;	Separate with the noun
ᠨᠠᠷ	[nar] [nər]	After the noun describing people	Separate with the noun
ᠱᠤᠳ ᠱᠦᠳ	ʃʊ:d][ʃu:d] [ʃʊ:l][ʃu:l]	After the adjective expressing the adjectives of human attributes and some noun (such as ethnic names)	With the noun attached together
ᠳ	[d]	After the noun ending with n (especially qin), occasionally added to a few nouns ending with r, l, i, or su (with voice off)	With the noun attached together
᠄	[s]	After some of the nouns ending with the vowel. Occasionally added to a few words ending with the n, l, i, and there are voice off phenomenon	With the noun attached together

Table Ⅶ shows the possession form of nouns. The possession of a noun reflects the affiliation of a subject. In Mongolian, the possession form is formulated by adding a suffix composition as a single word behind the target.

TABLE Ⅶ THE POSSESSION FORM

	Form	IPA	Means
Person genitive	ᠮᠢᠨᠤ	[min]	First person (singular)
	ᠴᠢᠨᠤ	[ʧin]	Second person (singular)
	ᠨᠢ	[ni]	Third person (singular)
	ᠮᠠᠨᠤ	[man]	First person (majority)
	ᠲᠠᠨᠤ	[tan]	Second person (majority)
Reflexive genitive	ᠠᠨ	[a:n][ə:n]	
	ᠬᠠᠨ	[xa:n] [xə:n]	

Table Ⅷ is the mood verb suffix. Mood represents the speaker's attitude to the relationship between the behavior and the object. The mood of Mongolian verb can be categorized into two types: the first is the imperative mood and the other is statement mood.

TABLE Ⅷ THE MOOD VERB SUFFIX

Type		Form	IPA
imperative mood	First person	᠂ ᠬᠦᠭᠡᠢ ᠂ ᠬᠦᠭᠡ	[j] [sʊgæ:][sugə:]
	Second person	ᠮᠤᠭᠤᠨ ᠪᠤᠳᠤᠨ ᠪᠢᠴᠢᠭᠤᠨ ᠪᠤᠳᠤᠨ ᠪᠢᠴᠢᠭᠦᠨ ᠪᠦᠳᠦᠨ	[gtʊn] [gtun] [a:ræ:] [ɔ:rœ:] [ə:re:] [o:re:]

			[a:tʃ] [ɔ:tʃ] [ə:tʃ] [o:tʃ]
	Third person	ᠠᠭ ᠴ ᠥᠭᠡ ᠥᠭᠡᠭᠡ ᠲᠤᠭᠠ ᠭᠠᠭᠠ ᠭᠤᠤ ᠭᠡᠭᠡ	[ag] [əg] [əg] [og] [tʊgæ:] [tuge:] [a:sæ:] [ɔ:sæ:] [ə:se:] [o:se:] [gʊ:dʒæ:] [gu:dʒe:]
statement mood	past tense	ᠵᠠ ᠵᠠ ᠵᠠ ᠵᠠ ᠪᠢ	[dʒ][tʃ] [dʒe:] [tʃe:] [w]
	Present tense	ᠨᠤ ᠪᠡ ᠪᠠᠢᠨᠠ/ᠨᠤ	[n] [dʒæ:n]
	perfect tense	ᠰᠠᠨ	[la:] [lɔ:] [lə:] [lo:]

Table IX presents the vice verb, which is the connection between two verbs. The main characteristic of vice verb is the connection and it can be categorized into three types, as shown in table Ⅸ. Finally, Table Ⅹ describes the participle suffixes. Participle can express behavior, action and the nature characteristic of an object. Therefore, a participle has both the nature of verbs and adjectives.

TABLE Ⅸ THE VICE VERB SUFFIXES

Type		Form	IPA
Simple connecti-on vice verbs	Tied vice verbs	ᠪᠡ ᠵᠡ	[dʒ][tʃ]
	Separation vice verbs	ᠲᠤᠭᠠ/ ᠣᠭᠠ	[a:d] [ɔ:d] [ə:d [o:d]
	Joint vice verbs	᠂	[n]
Restrict connecti-on vice verbs	Immediately vice verbs	ᠮᠠᠭᠴᠠ/ ᠮᠣᠭᠴᠠ ᠨᠠᠭᠠ/	[magtʃ] [məgtʃ] [məgtʃ] [mogtʃ] [na:ra:n][nə:rə:n]
	Follow vice verbs	ᠮᠠᠯᠠᠷ ᠬᠠᠯᠠᠷ ᠮᠣᠯᠣᠷ ᠬᠣᠯᠣᠷ	[xla:r] [xlɔ:r] [xlə:r] [xlo:r] [xlæ:] [xle:]
	premise vice verbs	ᠮᠡᠨᠳᠢᠨ/ ᠮᠦᠨ/	[mæ:ndʒin] [me:ndʒin] [mæ:n] [me:n]
	Assuming vice verbs	ᠪᠠᠯ ᠪᠣᠯᠪᠠ	[wal] [wɔl] [wəl] [wol] [was] [wɔs] [wəs] [wos]
	Concessions vice verbs	ᠪᠡᠴᠡ ᠲᠡᠴᠡ	[wtʃ] [ja:tʃ] [jɔ:tʃ] [jə:tʃ] [jo:tʃ]
	Meet vice verbs	ᠲᠠᠢ	[tal] [tɔl] [təl] [tol]
	Purpose vice verbs	ᠮᠠᠷ ᠬᠠᠷ ᠮᠣᠤᠵᠠ ᠬᠣᠤᠵᠠ	[xa:r] [xɔ:r] [xə:r] [xo:r] [xʊja:] [xuje:]
	Opportunity vice verbs	ᠩᠠᠭᠠ ᠂ ᠩᠣᠭᠣ ᠩᠤᠭᠤᠲ/ ᠂ ᠩᠤᠭᠤᠯ/	[ŋga:] [ŋgɔ:] [ŋgə:] [ŋgo:] [ŋgʊ:t] [ŋgu:t]
Mixed connecti-on vice verbs	Continuation vice verbs	ᠰᠠᠭᠠᠨ ᠰᠣᠭᠣᠨ	[sa:r] [sɔ:r] [sə:r] [so:r]

TABLE Ⅹ THE PARTICIPLE SUFFIXES

Type	Form	IPA
past tense participle	ᠰᠠᠨ/ ᠰᠣᠨ/	[san] [sɔn] [sən] [son]
Present and Future Tenses participle	ᠬᠤ ᠂ ᠺ	[x]
Regular body participle	ᠳᠠᠭ/ ᠳᠥᠭ	[dag] [dɔg] [dəg] [dog]
Continuous body participle	ᠠᠭᠴ ᠂ ᠣᠴ	[a:] [ɔ:] [ə:] [o:]
Likelihood participle	ᠮᠤᠷ ᠸᠴ	[ma:r] [mɔ:r] [mə:r] [mo:r] [m]
Main body participle	ᠭᠲᠠᠯ ᠂ ᠺᠲᠣᠯ	[gtʃ]

E. The current release

This construction of the resource is still undergoing. At present, our latest release is a dictionary containing 26462 words that include 25962 words inherited from the initial dictionary, and 500 words constructed using the morphological rules. This dictionary has been double checked by native Mongolians.

V. CONCLUSIONS

We present our recent progress on Mongolian linguistic resource construction, and release two resources: a text corpus and a Mongolian-Chinese dictionary. All the resources are free for research usages. In the future, we will keep on releasing larger text corpus and dictionary, plus other language and speech resources. Particularly, we will release more than 100 hours of Mongolian speech signals to assist the speech recognition community.

REFERENCES

[1] Population Census Office under the State Council, "Epartment of Population and Employment Statistics of National Bureau of Statistics", *Tabulation on the 2010 Population Census of the People's Republic of China*. *China Statistics Press*. 2012.

[2] Qing Gelier, "Mongolian Grammar", *Inner Mongolia People's Publishing House*, 1991.

[3] Д.Jegmeide, Xuwei Gao, "The First International Conference of Mongolian Scholars", *Mongolian Studies*, 1990 (3) 65-66.

[4] He Hu, "Acoustic Phonetics Clues in the Evolution of Vo wels of Mongolian", *Journal of central university for nationalities (philosophy and social sciences edition)*, 2015 (4), pp. 142-151

[5] Wu Lan, Dabhurbayar, GUAN Xiaoda and ZHO Qiang, "Phrase Structure Prasing of Mongolian", *Journal of chinese information processing,* 2014,28(5), pp. 162-169.

[6] GB 25914-2010: "information technology of traditional Mongolian nominal characters, presentation characters and control characters using the rules" , 2011-01-10, publish (2011)

[7] Inner Mongolia University, Mongolian Language Institute, "Mongolian and Chinese Dictionary", Inner Mongolia University Press, 1999.