A study on the correspondence between IPA notation and pinyin script of Bai language

Jian Yang [1,3*], *Yujuan* Xue[2], *Liu* Sun[1,3], *Zengyong* Luo[1], and *Yan* Yang[4]

[1]College of Engineering, Yuxi Normal University, Yuxi, Yunnan 653100, China
[2]School of Mathematics and Computer, Dali University, Dali, Yunnan 671003, China
[3]Yunnan Key Laboratory of Smart City in Cyberspace Security, Yuxi Normal University, Yuxi, Yunnan 653100, China
[4]College of Chinese Language and Literature, Yuxi Normal University, Yuxi, Yunnan 653100, China

Abstract. The Pinyin script of Bai language is the official script currently used by the Bai ethnic minority in China. However, due to various reasons, the process of its promotion and usage is relatively slow. The lack of learning materials for comparing Pinyin script of Bai language with Chinese has become an important obstacle to the application and promotion of Pinyin script of Bai language. Currently, there is no dedicated Bai language corpus, and a large part of the linguistic resources in existing research archives are directly annotated with international phonetic alphabet (IPA) for Bai language with its corresponding lists of transliteration and paraphrase of Mandarin Chinese. In contrast, resources for comparing Pinyin script of Bai language with Mandarin Chinese are very scarce. Addressing this problem, the paper studies the conversion rules from Bai language IPA to Pinyin script. The research results will help convert the existing Bai language corpus in literature from IPA to Pinyin script, greatly increasing the bilingual comparison resources of Bai language and Chinese, which can actively promote the usage of Pinyin Bai script and the protection and development of Bai language.

1 Introduction

Bai language is the native language of the Bais, who are the 15th largest ethnic minority in China. In the 1950s, the government of P. R. China conducted a national survey of minority languages[1], then Bai language was divided into three major dialects: the Dali (southern) dialect, the Jianchuan (central) dialect, and the Bijiang (northern) dialect. The overall differences between the three dialects are not significant, but they have their own unique vocal features[2]. There are no written words handed down in Bai language. The script created after the founding of the People's Republic of China is generally referred to as the New Bai script or Pinyin Bai script, which is a phonetic script based on Latin letters. This script scheme for Bai language has been revised several times. This paper conducts research based on the "Bai language scheme (Draft)" formulated in 1993. The basic dialects of this scheme include the Jianchuan dialect and the Dali dialect, with Jianchuan Jinhua and

*Corresponding author: yangjian@yxnu.edu.cn

Dali Xizhou as two standard sound points. The existing corpus resources for comparing Bai language and Chinese are very scarce, with most of Bai language being passed down orally and only a small portion being preserved in both recorded audio and transcripts in other languages during the process of linguistic research. Most of the existing text resources are also recorded in IPA, which greatly limits research on Bai language.

This article conducts research on the Dali (southern) dialect of Bai language. Based on the existing corpus data[3] and grammatically annotated texts[4], it conducts comparative analysis and research. Through statistical analysis and empirical research, it summarizes the conversion rules from the phonetic text represented by IPA to the Pinyin Bai script, which is helpful to transcribe IPA notation in existing research literature into Pinyin Bai script, promote the inheritance of ethnic minority language culture and facilitate the research on the evolution of ethnic languages. This paper first summaries the general situation of the Bai language, then elaborates the conversion rules from the Bai language IPA to the Pinyin script, following by some examples of typical conversion. At last a summary of this paper is made and the future research and development of Bai language are looked forward.

2 Overview of Bai language

2.1 Sound, rhyme and tone system

The Dali (southern) dialect of Bai language has two local dialects: Dali and Xiangyun, with the former being the representative. Among them, the Dali dialect is further divided into two local regions, Wase Dacheng and Xizhou Meiba. The Xizhou dialect of Dali City is one of the two standard accent points of Bai language.

(1) The phonetic system of the Dali dialect has a total of 23 initial consonants, which can be arranged according to the pronunciation parts of the lips, the tip of the tongue, the middle of the tongue, the surface of the tongue, and the base of the tongue, combined with the voicing of the plosive, the affricate, the nasal, and the fricative as follows:

p	ph	m	f	v		
ts	tsh		s	z		
t	th	n			l	
tɕ	tɕh	ȵ	ɕ			j
k	kh	ŋ	x	ɣ		

(2) The Dali dialect has a total of 21 vowels, none of which are nasalized. There are 8 monophthongs, including one retroflex vowel ɐɹ, and 13 compound vowels, as shown below:

Monophthongs: i, e, o, a, u, w, v

Compound vowels: ieˈ, ia, iɔ, iou, iɯ, iv, ou, ui, ue, ueˈ, ua, uo

(3) It has a total of 8 tones, which are closely related to the initial consonants and the final vowels, and there is a distinction between tense and soothing tones. According to the five-degree tonal notation, they can be represented as shown:

soothing tones: 55, 35, 33, 31, 32

tense tones: 42, 44, 21 (where 44 is the tone with omitting annotation)

2.2 Speech features and related works

The phonetic system of Bai language is characterized by a concise structure of consonants and vowels, with different phonetic systems in various dialects [5]. The Jianchuan dialect and the Dali dialect have 20 to 23 initial consonants, while the Nujiang dialect generally has more than 30 initial consonants. In terms of vowels, due to the influence of loanwords

from Chinese, there are slight differences in nasal components among different dialects. In terms of tones, the number of tones varies among different dialects, usually ranging from 6 to 8 tones.

In recent years, numerous scholars have conducted research on the phonetics of Bai language, explaining the phonetic differences between loanwords from Chinese and fixed dialect words[6] and summarizing the voiced fricative of Bai language's initial consonants[7]. The construction of a speech corpus of Bai language[8] is also underway, and research on Bai language using computers and AI technologies is gradually progressing. A speech recognition scheme, based on HTK tools according to the phonetic characteristics of Bai language, has also been proposed[9]. With the deepening of cultural exchanges among various ethnic groups, the translation of Bai language into other languages has gained attention. Machine translation for Chinese and Bai language will also be a valuable direction for future research. Through the analysis of vocabulary and grammar in Bai language, applicable translation methods can also be obtained[10]. A automatic speech recognition system of Bai language based on deep learning was proposed[11]. However, most existing research results are based on the acoustic features of Bai language, and research on the application of Bai language translation and speech recognition to the linguistic features of the Bai language is scarce.

Most of the existing literature on Bai language research uses IPA to annotate the pronunciation, with a reference to the transliteration and paraphrase for Chinese or other languages. As far as we could tell, there is no research on the correlation between Pinyin Bai script and IPA notations of speech resources of Bai language. Therefore, it is necessary to summarize the rules for converting to the Pinyin Bai script through IPA. This conversion can enable the transformation of IPA script into Pinyin script, expanding the Bai language corpus while facilitating the study of linguistic features of the Bai language. This, in turn, can promote the work of bilingual translation between Chinese and the Bai language, broaden the channels of language and cultural inheritance, and better protect the endangered languages of Bais.

2.3 Features and current status of Pinyin Bai script

The Bai is a minority ethnic group deeply influenced by Han culture, and Mandarin Chinese has been used by the Bai people since ancient times. The Bai language has two scripts. Pinyin is a new Bai script created after the founding of the People's Republic of China, based on the Latin alphabet as the symbol system. In addition to the new script, there is also an old Bai script, which was used during the Tang and Song dynasties by imitating Chinese characters. The old Bai script consists of a symbol system composed of borrowed Chinese characters and newly created characters, hence it is also known as the ancient Bai script, square Bai script, and Chinese character Bai script. Due to the limitations and historical reasons, the old Bai script has almost been lost, and the existing old Bai script only appears on cultural relics such as Nanzhao tiles, ancient Buddhist scriptures and inscriptions. Currently, the official script of the Bai language is Pinyin Bai script, which was promoted after several revisions of the script scheme. The Pinyin script uses the Latin alphabet as the basis, representing characters based on their pronunciation. Due to the significant influence of Chinese on the Bai language, there are expressions in the Bai language that have meanings close to those in Chinese, and such expressions are usually translated using fixed forms of the Bai language. However, there are still many Chinese loanwords in Pinyin script, such as "caot yuip" for "草原" and "tsꞁ" for "子". In addition to Chinese loanwords, there are also some sounds in the Bai language that are the same or similar to those in Chinese, which are usually represented using the same letters in Chinese Pinyin.

Currently, Pinyin Bai script, as the official script of the Bai language, is still in the stage of promotion and usage. Its users include not only residents of Bai ethnic areas but also some enthusiasts of Bai language research. There are also bilingual teaching experiments in Jianchuan, Dali, and other areas, and some Bai script teaching material and books are gradually being published [12], which has produced a positive effect on the promotion of Pinyin Bai script. However, the Bai language is still an endangered minority language that only a few people can understand, speak, and use. The popularization of Mandarin Chinese has also made the number of people who can speak Bai language very small, and most young people are unwilling to learn and inherit their own minority language. In addition, a large proportion of people who can speak Bai language don't know how to write Pinyin Bai script, resulting in extremely scarce resources for Pinyin Bai script corpus and comparing corpus with Chinese. Therefore, it is urgent to rescue, protect, learn, and inherit the Bai language. Addressing the problem, this paper uses existing IPA scripts of Bai language and Chinese comparing corpus to study the rules for converting IPA into Pinyin Bai script, transforming existing materials into bilingual materials of Pinyin Baiwen and Chinese, which is conducive to the promotion of Pinyin Bai scripts and the protection of the endangered minority language.

3 Conversion from IPA to Pinyin Bai script

Based on the spelling scheme in [3], text data research, and interviews with experts, as well as the IPA corpus in [4], a complete conversion rule from IPA to Pinyin Bai script is summarized in this section. The mentioned spelling rules of the Bai language are based on the book "Bai Nationality Script Plan (Draft)" published in 1993. According to the expression habits, the single syllable 'v' is written as 'vu', and the missing vowel 'iei' in the basic spelling rules is added. On this basis, some special syllable conversion cases are constructed to obtain the complete rules about the conversion from IPA to Pinyin Bai script. The conversion rules will be listed one by one as followed according to different situations. The statistic features related to the frequency of syllables involved in the conversion rules are all derived from the corpus data in the literature [4].

3.1 Basic conversion of the sound, rhyme, and tone system

(1) Conversion of initial consonants

Table 1. Conversion table of initial consonants.

IPA	Bai script	IPA	Bai script	IPA	Bai script	IPA	Bai script
p	b	th	t	ɣ	hh	ɳ	ni
ph	p	n	n	ŋ	ng	ts	z
m	m	l	l	tɕ	j	tsh	c
f	f	k	g	tɕh	q	s	s
v	v	kh	k	ɕ	x	z	ss
t	d	x	h	j	y		

(2) Conversion of vowels

Table 2. Conversion table of final consonants.

IPA	Bai script	IPA	Bai script	IPA	Bai script	IPA	Bai script
a	a	e	ei	iou	iou	ua	ua
o	o	ɯ	e	iɯ	ie	uo	uo
eˡ	er	v	v	ieˡ	ier	ueˡ	uer
i	i	ia	ia	ie	iei	ɔ	ao

u	u	iɔ	iao	ui	ui	ou	ou

(3) Conversion of tone

Table 3. Tone conversion table.

tone name	tone symbol	tone value	tone name	tone symbol	tone value
x tuning	x	33	p tuning	p	42
omitting annotation		44	f tuning	f	35
l tuning	l	55	d tuning	d	21
t tuning	t	31	z tuning	z	32

3.2 Conversion of simple final syllables

In the Bai language spelled by IPA, there are often cases of simple finals without initial consonants. A syllable is formed by an independent vowel, which can express a complete meaning. At this time, it is necessary to consider comprehensively according to the pronunciation of the Bai language, and add the initial consonant 'w' before some finals. According to the statistical survey results, there are three types of simple finals that need to be added with the initial consonant 'w': u, ua, and ue. Not all simple finals without initial consonants need to be added. Here are some examples (the first three are cases where the simple finals are added with the initial consonant 'w', and the last two are cases not need):

u^{31}→wut (4 times)
ua^{55}→wual (45 times)
ue^{33}→wuex (86 times)
a^{31}→at (219 times)
ou^{42}→oup (8 times)
Example 1:
IPA: $ke\iota^{55}$ $ȵi^{44}$ a^{33} $tɯ^{44}$ $nɔ^{31}$ $pe^{44}thu^{33}$ ua^{55} pu^{31} xua^{35} tsa^{35}
Pinyin Bai script: gerl ni ax de naot bei tux wual but huaf zaf.
Paraphrase: I noticed you were in a bit of a hurry when you walked today.
Example 2:
IPA: ti^{21} a^{31} sua^{44} $pe\iota^{42}$ $ts\gamma^{55}$ $ke\iota^{35}$ $tɔ^{31}$ $xɯ^{44}$ nv^{21} $tɕhie\iota^{55}$ $tɯ^{44}$.
Pinyin Bai script: did at sua berp zil gerf daot he nvd qierl de.
Paraphrase: As long as one speaks the Bai language, they fear that the Great Black Dragon might hear them.

3.3 Conversion of monophthongs

In the Bai language represented by the IPA, there are cases where the initial consonant 'v' appears as a single consonant, meaning that there is no vowel following it. In such cases, the single consonant 'v' is converted to 'vu', and the tone conversion is carried out as usual. In other cases where there is a vowel following the consonant 'v', the conversion can be carried out according to the common rules. Here are some examples:

v^{33}→vux
va^{32}→vaz

3.4 Conversion of the double i structure caused by the initial consonant ȵ

When the initial consonant ȵ is converted to Pinyin, it becomes ni. However, in the representation of Bai language in the IPA, the initial consonant ȵ appears 865 times, all of which are followed by the final consonant i. This situation results in the "ȵi" being converted to the Pinyin script as the "nii", causing a script of double i. In the real writing of

Pinyin script, this double i structure is automatically simplified by removing one i, that is to convert ŋi- to ni-. Here is an example:

ŋia^{55}→nial (55 times)
ŋi^{21}→nid (374 times)
ŋiou^{44}→niou (88 times)
Example 3:
IPA: pɔ^{31} ue^{44} xa^{31} le^{21} nɔ^{44} ŋiou^{44} tɕia^{33} le^{21} jɯ^{44} ŋi^{55} ?
Pinyin script: baot wue hat leid nao niou jiax leid ye nil ?
Paraphrase: Why does he eat like this?

3.5 Syllable conversion of vowel iv

The vowel iv does not have a corresponding Pinyin script. According to the statistical analysis of existing IPA script data, iv only appears in the structure ŋiv- (appears 72 times). Considering the conversion of the initial consonant ŋ to remove one i from the double i structure and the conversion of the vowel v, the ŋiv is converted to niv directly, shown as follows:

ŋiv^{33}→nivx (70 times)
Example 4:
IPA: ŋeɪ^{21} mi^{32} nɔ^{44} ŋiv^{33} ŋi^{21} sɯ^{33} tɯ^{44} xɔ^{55},
Pinyin script: ngerd miz nao nivx nid sex de haol,
Paraphrase: The daughter of Haiyin Village understood,

3.6 Syllable conversion of vowel sound ue

The vowel "ue" does not have a corresponding Pinyin script in the Bai language. Based on the pronunciation of the Bai language and the consideration of aspirated and unaspirated consonants, there are generally three scenarios when converting the vowel ue to Pinyin Bai script: (1) converting to "uer"; (2) converting to "uei"; (3) no conversion, still written as ue. Additionally, there is a special case where ue is a simple final syllable, that is, the syllable ue has no initial. This situation can be referred to as the method of adding the initial consonant w in front of ue mentioned earlier, which converts "ue" to "wue". According to different scenarios, the conversion rules for syllables containing the vowel ue can be summarized as shown in the Table 4.

Table 4. Conversion rules for syllables containing the vowel sound "ue"

IPA	Bai script	IPA	Bai script	IPA	Bai script
khue-	kuer-	lue-	luei	thue-	tue-
kue-	guer-	sue-	suei-	tshue-	cue-
nue-	nuer-	tsue-	zuei-	ue-	wue-
xue-	huer-	tue-	duei-		

Examples of different conversion rules in the above table are as follows.

khue^{55}→kuerl lue^{31}→lueit thue^{55}→tuel
kue^{32}→guerz sue^{33}→sueix tshue^{31}→cuet
nue^{55}→nuerl tsue^{31}→zueit ue^{33}→wuex
xue^{35}→huerf tue^{32}→dueiz

3.7 Conversion of special symbol ɳ

IPA indicates that a large portion of the vocabulary in the Bai language has been influenced by Chinese, resulting in the emergence of many function words and suffixes.

The phonetic character ɿ is often used to form these function words and suffixes, with common ones including tsɿ44 (noun suffix), tsɿ55 (auxiliary word), etc. When converting syllables containing the character ɿ, the character ɿ is consistently converted to the letter i. Examples are as follows:

$ts\text{ɿ}^{44}$→zi (135 times)
$ts\text{ɿ}^{55}$→zil (246 times)
$s\text{ɿ}^{35}$→sif (47 times)
Example 5:
IPA: $ts\text{ɯ}^{33}$ $m\text{ɯ}^{44}$ $ts\text{ɯ}^{35}$ $tse\text{ɹ}^{21}$ $ts\text{ɿ}^{55}$ kuo^{32} kua^{35}.
Pinyin Bai script: zex me zef zerd zil guoz guaf.
Paraphrase: Once served as an official in Mengzhou city.

4 Conversion instance

Based on the study of texts represented in the IPA of the Dali dialect of the southern Bai language, the literature [4] contains 21 long dialogues of the Bai ethnic group, including rich mythological stories and legends of the Bais' local deities (named as Benzhu). According to the conversion rules proposed, these corpora can all be converted into Pinyin Bai script. Below are some examples of selected corpus (excerpts from "The Story of the Great Black Dragon").

(1) Example 6.
IPA: $pu\text{ɯ}^{55}$ $tu\text{ɯ}^{21}$ $m\text{ɯ}^{35}$,
Pinyin Bai script: bel ded mef,
Paraphrase: Once upon a time,
(2) Example 7.
IPA: ŋ̍ia^{55} kou^{32} $xu\text{ɯ}^{31}$ $ts\text{ɯ}^{33}$ nv^{21} ou^{21} $p\text{ɔ}^{35}$ ŋ̍i^{21}.
Pinyin Bai script: nial gouz het zex nvd oud baof nid.
Paraphrase: In our Fengwei brook, there is a dragon king.
(3) Example 8.
IPA: nv^{21} ou^{21} $p\text{ɔ}^{35}$ $tu\text{ɯ}^{31}$ ŋ̍i^{21} ou^{44} ɕui^{33} le^{55} ou^{44} mu^{33},
Pinyin Bai script: nvd oud baof det nid ou xuix leil ou mux,
Paraphrase: The dragon king doesn't make rain either,
(4) Example 9.
IPA: $ke\text{ɹ}^{33}$ v^{33} le^{55} $ke\text{ɹ}^{33}$ mu^{33},
Pinyin Bai script: gerx vx leil gerx mux,
Paraphrase: It's also not raining,
(5) Example 10.
IPA: $ku\text{ɯ}^{32}$ kuo^{32} thu^{33} ŋ̍i^{21} $ke\text{ɹ}^{35}$ kha^{44} ɕui^{33} $ne\text{ɹ}^{55}$ ɣɯ^{33} tɕhiɯ^{55} $ts\text{ɿ}^{44}$.
Pinyin Bai script: dez guoz tux nid gerf ka xuix nerl hhex qiel zi.
Paraphrase: If any passerby is thirsty, offer them a drink.

5 Conclusion

Since its inception, the promotion of Pinyin Bai script has been slow, with corresponding learning materials available only in some regions, and relatively few publications in Bai script. Coupled with the popularization of Chinese, more and more young people are reluctant to learn and use Pinyin Bai script, leading to the endangered heritage status of the Bai language. Based on existing comparative corpus resources, the paper proposed the conversion rules from IPA to Pinyin Bai script, enabling the expansion of existing phonetic and textual corpora to be converted into Pinyin Bai script corpora. This promotes the use

and inheritance of ethnic languages to a certain extent and also facilitates the research on the linguistic features of the Bai language and the translation between Bai and Chinese, protecting and inheriting endangered languages of ethnic minorities while deepening the exchange of ethnic cultures.

The transmission of the Pinyin Bai script is not only an evolutionary study of the national language, but also a crucial foundation for promoting the prosperity and development of national culture. The use of emerging technologies to disseminate and translate the Bai language, thereby achieving bilingual translation between Chinese and the Bai language, is just around the corner. Based on the rules proposed in this paper, our future research will build a new text corpora composed of Pinyin Bai script and corresponding Chinese text based on the existing text corpora, which is composed of IPA and its Chinese translations. At the same time, with the help of deep learning technology, a system for speech recognition of the Bai language and translation between the Bai language and Chinese and other languages will be designed and constructed.

Fundings: This work was supported in part by the National Natural Science Foundation of China under Grant No. 62266048, and in part by the Yunnan Fundamental Research Project No. 202101AU070007.

References

1. S. Lianke, Digital methods and basic theoretical research on the protection and inherita nce of the vocal language and oral culture of the Yi people. J. Yuxi Norm. Univ. **30**, 18 -22 (2015). https://doi.org/10.3969/j.issn.1009-9506.2015.01.004
2. X. Lin, Dali series: Bai language (Yunnan Nationalities Publishing House, Kunming, 2 008)
3. W. Feng, 366 Bai language conversation sentences (Social Sciences Academic Press, B eijing, 2014)
4. W. Feng, Annotated text of Bai language grammar (Social Sciences Academic Press, B eijing, 2016)
5. D. Wenfei, A comparative study of Bai dialects and vernaculars in the surrounding are as of Erhai Lake, Master Thesis, Yunnan Normal University, College of arts (2013)
6. W. Anqi, The phonology and attribution of the Bai language. National Languages. **30 (4)**, 3-22 (2009)
7. Y. Xiaoxia. Study on the aspirated fricative in Bai l-anguage, Master Thesis, Yunnan N ormal University, College of arts (2007)
8. Y. Jian, L. Haiguang, Z. Xiaoling, Research on the construction of Bai language speec h corpus. J. Dali Univ. **02(012)**, 21-26 (2017)
9. Z. Lingtong. Speech recognition method of the Bai ethnic group based on HTK. J. Dali Univ. **12(10)**, 27-32 (2013). https://doi.org/10.3969/j.issn.1672-2345.2013.10.007
10. G. Jingfang, Z. Lingtong, Comparative analysis of machine translation methods applica ble to Bai-Chinese. Digital Tech. & Appl. **38(5)**, 224-225 (2020). https://doi.org/10.19 695/j.cnki.cn12-1369.2020.05.126
11. G. Jingfang. Speech recognition system for the Bai language based on deep learning, Master Thesis, Dali University, Faculty of Engineering (2021)
12. Z. Xia. The use, development, and status of Pinyin scripts (Dali Daily Press, Dali, 2017)