Research Report on Automatic Word Segmentation and Tagged Corpus part

Planned Completion Date : 12/31/09
Title: Research Automatic Word Segmentation
Researchers: Ngodrup NyimaTrashi TashiJia Suonamjiancuo GuanBai PuDanzen
Presentation: National Standard，Papers and Documentations

Abstraction: Tibetan segmentation is the continual syllable sequence according to certain standard again group compound word sequence process. Automatic word segmentation is a basic subject in Tibetan Information Processing, also is the key subject in intelligent Tibetan information processing area. For resolve the Tibetan segmentation of the “word processing level” in Tibetan information processing, one of the precondition is the standards word classes and correct segmentation. In order to identify the automatic word segmentation and part-of-speech tags, firstly classify the Tibetan words according to requirements of Tibetan information processing and Secondly in accordance with Tibetan characteristics and its own formation rules and thirdly in accordance with Chinese segmentation, provide systemic and applied word segmentation criterions.

Based on processes the Tibetan real text in the computer, in this standard rules, must cover in the linguistics significance the word, but must cover the unit which is smaller than the word, like meets the ingredient (prefix), to meet the ingredient (decorates), meets the ingredient (suffix) and so on, as well as compared to a word bigger unit, like idiom, custom terminology, abbreviation, abbreviation as well as punctuation mark, non-Tibetan mark and so on. Only then like this, this standard can supply the information which Tibetan information processing needs, therefore this standard's part of speech mark collection and the above these is bigger than or information and so on unit which including Tibetan part of speech information the word as well as punctuation mark, non-Tibetan mark is smaller than the word. The current text corpora database contains about 1,000,000 words.

Status of Activity: Published research paper where providing word segmentation criterions.

The Contents for project:
(1) Modern Tibetan words and expressions classification system

(2) Part-of-speech tagging

(3)Tibetan word segmentation vocabulary standard formulation

(4)Tibetan word segmentation theory

(5) Tibetan word segmentation algorithm

(6) The parts-of-speech and tagging set standards of Tibetan Information Process (GB, National standards is applying)

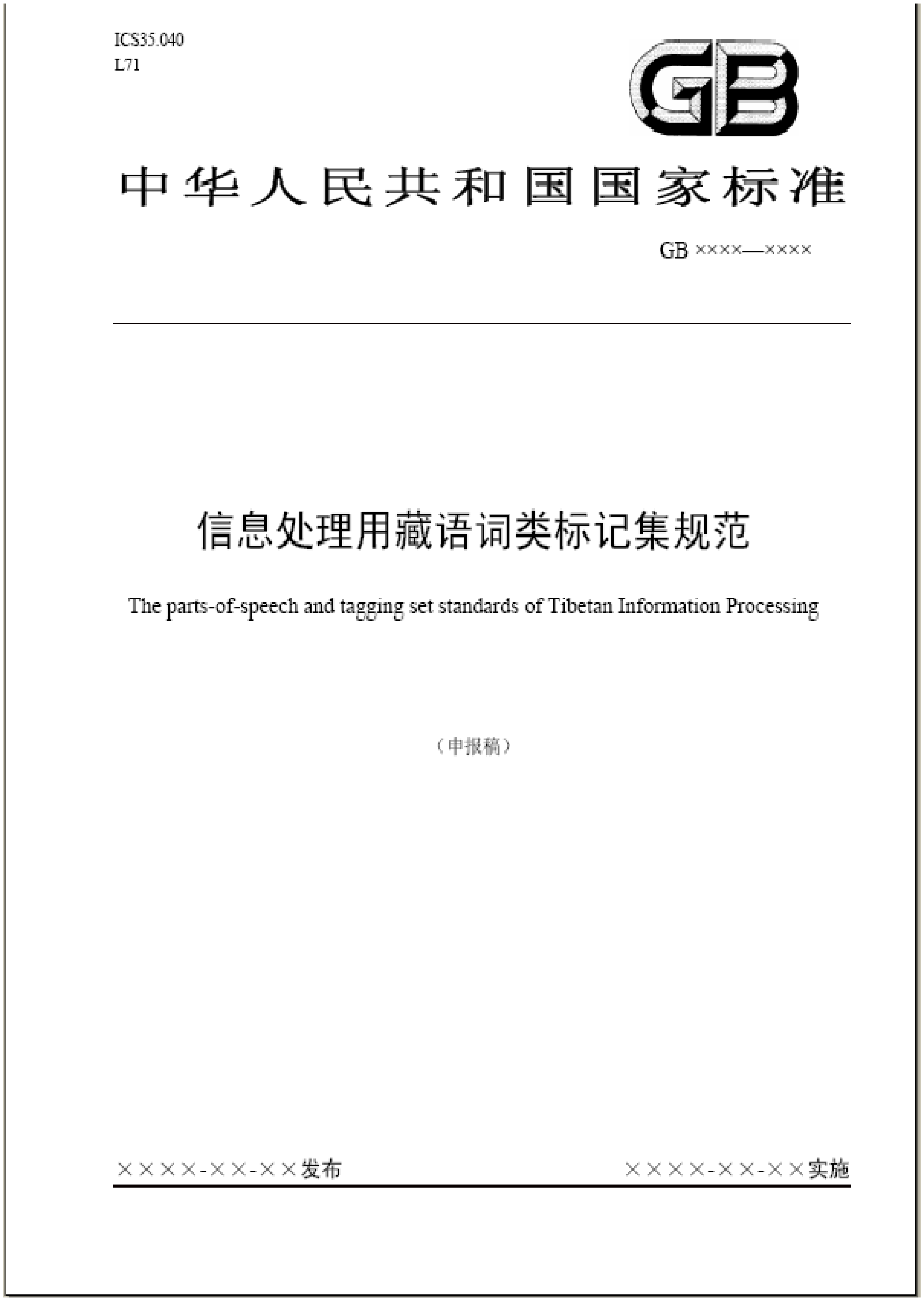
(7) Research on Tibetan Segmentation Criterion for Information Processing (GB,

National standards is applying)

Status of Activity: finished

Software： GangJian (Snow eyes) Tibetan word segmentation software is made in TU

Result:



ICS35.040
L71

中华人民共和国国家标准

GB ××××—××××

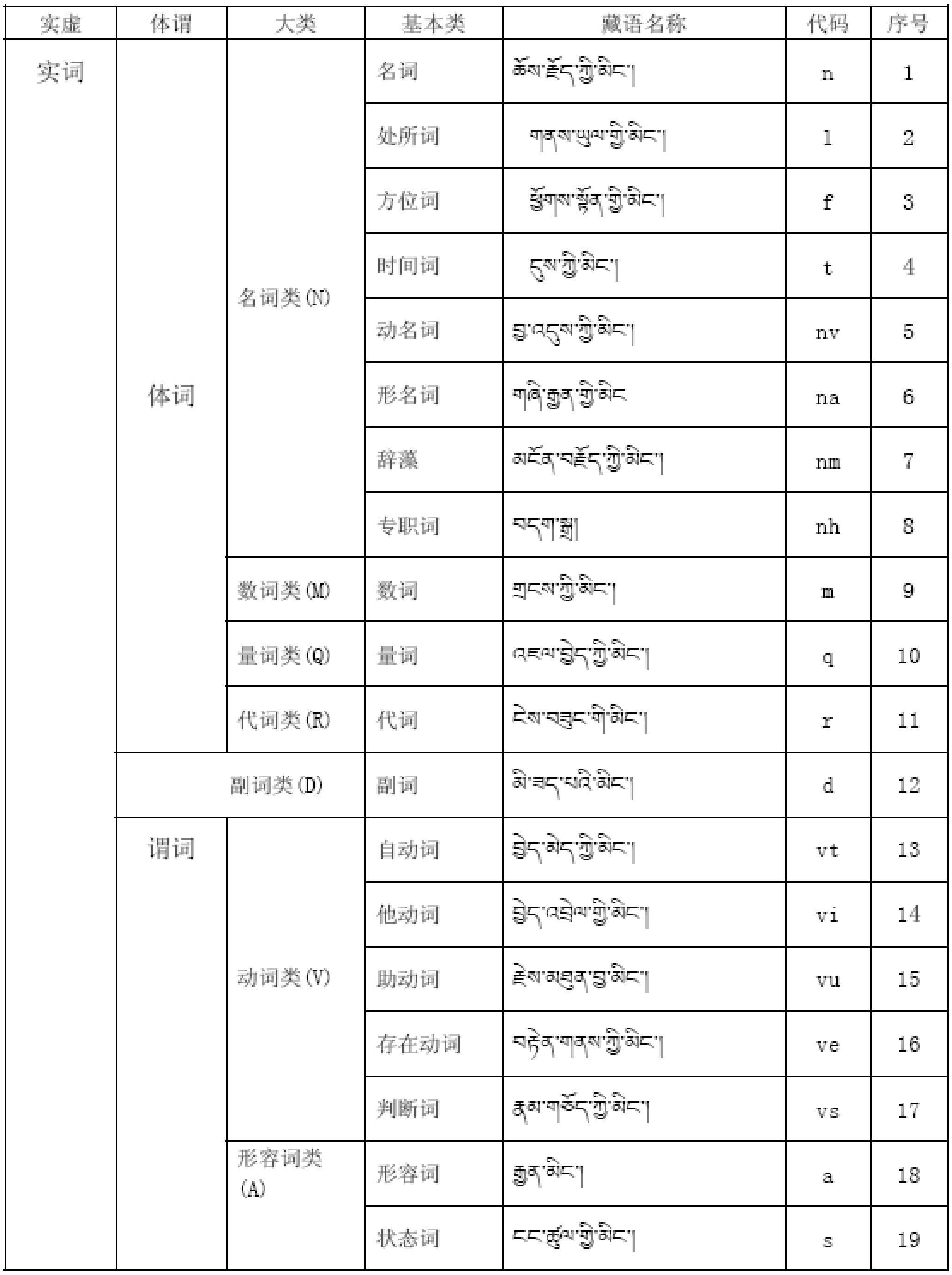
信息处理用藏语词类标记集规范

The parts-of-speech and tagging set standards of Tibetan Information Processing

（申报稿）

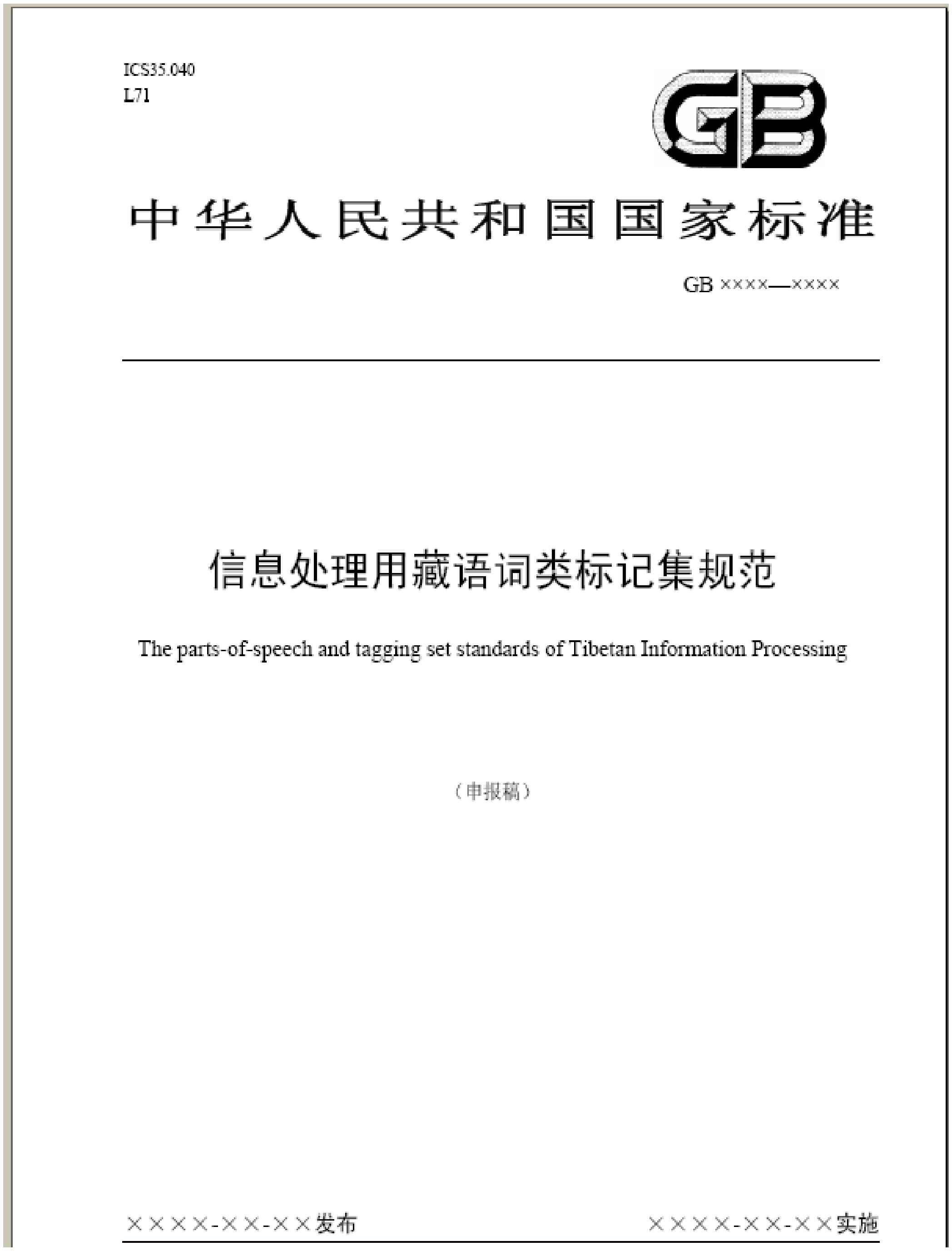
××××-××-××发布　××××-××-××实施

1．Main words classes and Code



实虚	体谓	大类	基本类	藏语名称	代码	序号
实词	体词	名词类(N)	名词	ཆོས་རྗོད་ཀྱི་མིང་།	n	1
			处所词	གནས་ཡུལ་གྱི་མིང་།	l	2
			方位词	ཕྱོགས་སྟོན་གྱི་མིང་།	f	3
			时间词	དུས་ཀྱི་མིང་།	t	4
			动名词	བྱ་འདུས་ཀྱི་མིང་།	nv	5
			形名词	གཞི་རྒྱན་གྱི་མིང	na	6
			辞藻	མངོན་བརྗོད་ཀྱི་མིང་།	nm	7
			专职词	བདག་སྒྲ།	nh	8
		数词类(M)	数词	གྲངས་ཀྱི་མིང་།	m	9
		量词类(Q)	量词	འཇལ་བྱེད་ཀྱི་མིང་།	q	10
		代词类(R)	代词	ངེས་བཟུང་གི་མིང་།	r	11
	副词类(D)		副词	མི་ཟད་པའི་མིང་།	d	12
	谓词	动词类(V)	自动词	བྱེད་མེད་ཀྱི་མིང་།	vt	13
			他动词	བྱེད་འབྲེལ་གྱི་མིང་།	vi	14
			助动词	རྗེས་མཐུན་བྱ་མིང་།	vu	15
			存在动词	བརྟེན་གནས་ཀྱི་མིང་།	ve	16
			判断词	རྣམ་གཅོད་ཀྱི་མིང་།	vs	17
		形容词类(A)	形容词	རྒྱན་མིང་།	a	18
			状态词	ངང་ཚུལ་གྱི་མིང་།	s	19

			区别词	ཁྱད་མིང་།	b	20
虚词		助词类(U)	时态助词	དུས་གསལ་མིང་།	ut	21
			语气助词	བརྣན་གསལ་གྱི་མིང་།	un	22
			祈使助词	སྐུལ་སྒྲོན་མིང་།	ap	23
			比喻助词	མཚུངས་གསལ་གྱི་མིང་།	ui	24
			原因助词	རྐྱེན་གསལ་མིང་།	uc	25
			目的助词	དམིགས་གསལ་མིང་།	ue	26
			终结助词	རྫོགས་ཚིག།	uf	27
	格助词类(P)		格助词	རྣམ་དབྱེའི་རྐྱེན།	p	28
	连词类(C)		连词	ཚིག་ཕྲད།	c	29
叹词类(E)			叹词	འབོད་སྒྲ།	e	30
拟声词类(O)			拟声词	རྗེས་སུ་བྱེད་པའི་སྒྲ།	o	31
附加类	大于词的单位		成语	དཔེ་ཆོས།	i	32
			习惯语	རྒྱུན་འཇགས་མིང་།	y	33
			简略语	བསྡུས་གཟུགས།	j	34
	小于词的单位		前接成份	སྔ་རྐྱེན།	h	35
			中接成份	བར་གྱི་གསལ་བྱེད།	zh	36
			后接成份	མིང་མཐའ།	k	37



ICS35.040
L71

中华人民共和国国家标准

GB ××××—××××

信息处理用藏语词类标记集规范

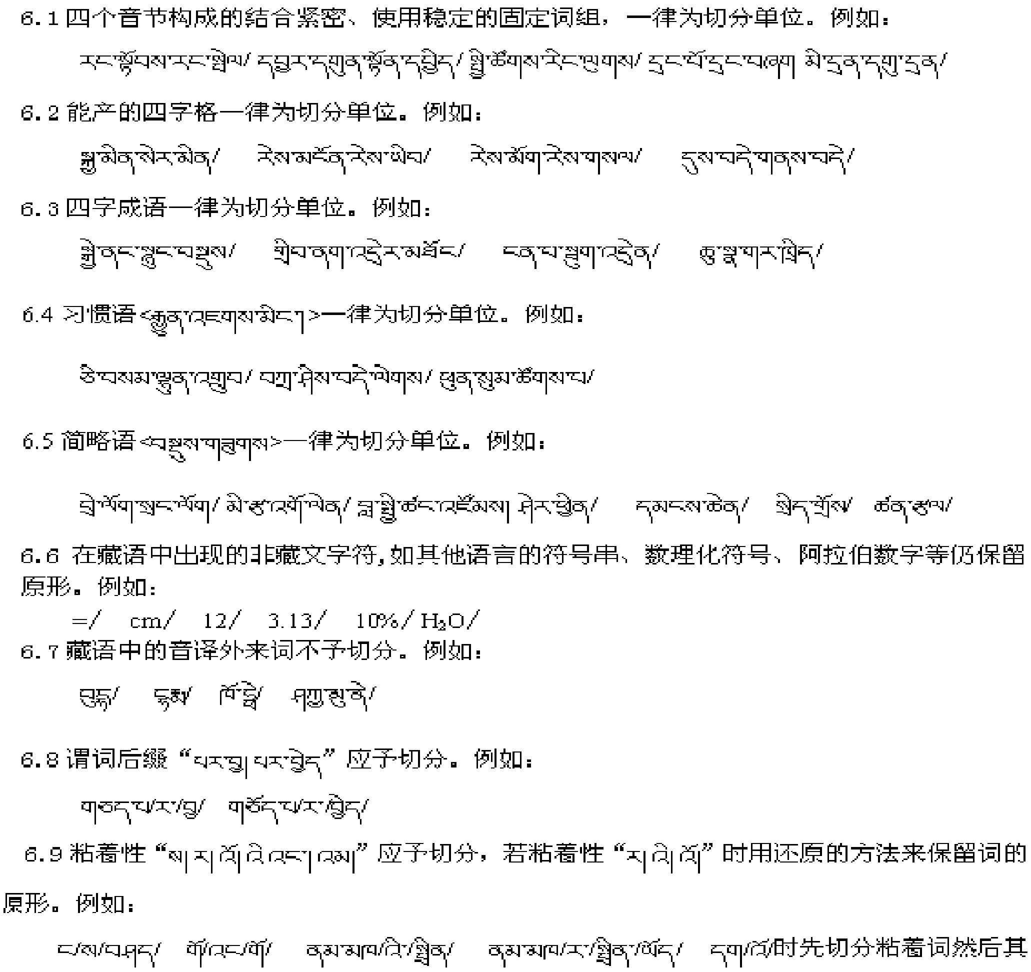
The parts-of-speech and tagging set standards of Tibetan Information Processing

（申报稿）

××××-××-××发布　　××××-××-××实施

Standard general rule:

This standard take Tibetan information processing as a goal, according to Tibetan glossary construction word rule and characteristic, provide Tibetan segmentation principle. In this standard Tibetan syllable symbol “.” does not make the segmentation unit's mark, but “/” takes the segmentation unit's mark.



6.1 四个音节构成的结合紧密、使用稳定的固定词组，一律为切分单位。例如：

རང་སྟོབས་རང་གྲོལ/ དབྱར་དགུན་སྟོན་དཔྱིད/ སྒྱུ་ཚིགས་རིང་ལུགས/ དྲང་པོ་དྲང་པ་ཞིག། མི་དྲན་དགུ་དྲན/

6.2 能产的四字格一律为切分单位。例如：

སྐྱ་མིན་སེར་མིན/ རིས་མཛོན་རིས་ཡིབ/ རིས་སོག་རིས་གསལ/ དུས་བདེ་གནས་བདེ/

6.3 四字成语一律为切分单位。例如：

སྐྱེ་ནང་སྐྱུང་བསྡུས/ གྲིབ་ནག་འཛེར་མཐོང/ ངན་པ་སྡུག་འཛིན/ ཆུ་སྣ་གར་ཕྱིད/

6.4 习惯语<རྒྱུན་འཇུགས་མིང།>一律为切分单位。例如：

ཅི་བསམ་ལྷུན་འགྲུབ/ བཀྲ་ཤིས་བདེ་ལེགས/ ཕུན་སུམ་ཚོགས་པ/

6.5 简略语<བསྡུས་གཟུགས>一律为切分单位。例如：

ཐེ་ཤོག་སྒང་ཤོག/ མི་རྩ་འགོ་ཡིན/ ཟླ་སྤྲི་ཚོང་འཛོམས། ཤེར་ཕྱིན/ དམངས་ཆེན/ སྲིད་གྲོས/ ཚན་རྩལ/

6.6 在藏语中出现的非藏文字符,如其他语言的符号串、数理化符号、阿拉伯数字等仍保留原形。例如：

=/ cm/ 12/ 3.13/ 10%/ H_2O/

6.7 藏语中的音译外来词不予切分。例如：

ཨརྻ/ ཧྲཾ/ ཀོ་ཊི/ ཤཀྱ་མུ་ནེ/

6.8 谓词后缀 “པར་བྱ། པར་བྱེད” 应予切分。例如：

གཅད་པ/ར་/བྱ/ གཅོད་པ/ར་/བྱེད/

6.9 粘着性 “ས། ར། འོ། འི འང། འམ།” 应予切分，若粘着性 “ར། འི། འོ།” 时用还原的方法来保留词的原形。例如：

ང/ས/བཤད/ གོ/འང/གོ/ ནམ་མཁ/འི/སྤྲིན/ ནམ་མཁ/ར་/སྤྲིན་/ཡོད/ དག/འོ/时先切分粘着词然后其