Computer Science and Application 计算机科学与应用, 2016, 6(7), 415-422
Published Online July 2016 in Hans. http://www.hanspub.org/journal/csa
http://dx.doi.org/10.12677/csa.2016.67051

Hans汉斯

An Approach to Normalization of Dai Text for Speech Synthesis

Zhumei Wu, Jian Yang*, Zhan Wang

School of Information Science and Engineering, Yunnan University, Kunming Yunnan
Email: wucaptain@139.com, *nxryang@126.com

Received: Jul. 6th, 2016; accepted: Jul. 25th, 2016; published: Jul. 29th, 2016

Abstract

With the purpose of developing a Dai speech synthesis system, this paper focuses on the study of Dai numbers and special characters normalization. Both numbers and special characters are the non-standard words in Dai text. The main purpose of the text normalization is to represent the pronunciation of non-standard words with standard words. The normalization process includes non-standard words recognition, ambiguity judgment, disambiguation and non-standard translation. Firstly, the non-standard words are recognized and the ambiguous types of these non-standard words are determined using a method based on rule-based and context-keyword, in this paper. Then, the types of ambiguity are judged on regular expression. Lastly, the correct pronunciation of no-standard words is determined according to the transformation rules. Experimental results show that the correct rate of this normalization is more than 94.6%. This purposed method can fully satisfy the front-end text analysis in Dai text to speech conversion system, and has a good natural language processing application value.

Keywords

Dai Language, Speech Synthesis, Text Analysis, Normalization

傣语语音合成中的文本归一化方法

伍烛梅，杨　鉴*，王　展

云南大学信息学院，云南　昆明
Email: wucaptain@139.com, *nxryang@126.com
*通讯作者。

文章引用: 伍烛梅, 杨鉴, 王展. 傣语语音合成中的文本归一化方法[J]. 计算机科学与应用, 2016, 6(7): 415-422.
http://dx.doi.org/10.12677/csa.2016.67051

收稿日期：2016年7月6日；录用日期：2016年7月25日；发布日期：2016年7月29日

摘　要

本文以开发傣语语音合成系统为目的，重点研究傣语文本中的数字归一化和特殊字符归一化问题。数字和特殊字符都属于傣语文本中的非标准词，文本归一化的主要目的是用标准词表示非标准词的发音。归一化处理过程包括：非标准词识别、歧义判断、消歧处理和非标准词转换为标准词4个步骤。本文采用基于规则和上下文关键词相结合的方法识别非标准词，利用正则表达式判断其歧义类型，根据转换规则对非标准词进行消歧并确定其正确的傣文读音。实验结果表明，本文提出的文本归一化方法的正确率达到了94.6%，可以完全满足傣语文语转换系统前端文本分析的需求，并具有良好的自然语言处理应用价值。

关键词

傣语，语音合成，文本分析，归一化

1. 引言

中国傣族是一个历史悠久的民族，傣语是傣族人民使用的语言，全球约有 6600 万左右的人口使用傣语[1]。据不完全统计，中国少数民族语言有 80 多种，傣语是在汉藏语的基础上发展起来的一门独立语言，属于汉藏语系壮侗语族傣语支，在类型上同汉语一样是单音节为主的有声调语言，它是一种拼音文字，其字符为傣语字符[2]。由于西双版纳傣语的历史最为悠久，较多的保留了固有的语言习惯，所以，本文以西双版纳新傣文作为语音合成的研究对象。

虽然语音合成技术的研究已有较长时间，但大多仍集中在中英文语言的合成，对少数民族语言涉及较少[3]。近年来，少数民族语言的语音合成研究逐渐引起了研究者的关注，但主要集中在维吾尔语、蒙古语、藏语等语言中，对傣语研究仍是少有提及。而在我国云南，使用傣语的傣族同胞不在少数，为了促进傣族同胞的信息化发展，研究傣语语音合成的重要性不言而喻。

语音合成包括前端文本分析和后端语音合成。前端文本处理是研究傣语语音合成的前提工作，在语言层、语法层、语义层的处理工作可以归结为前端的文本分析[4]。针对傣语的特点，傣语的文本分析的工作包括文本归一化、分词、傣语音素罗马化以及汉语音素罗马化四个部分。本文重点研究傣语文本归一化问题，主要从数字归一化和特殊字符归一化两方面进行详细阐述。

2. 傣语文本归一化流程与规则

傣文是一种拼音文字，正常形式为“声母+韵母+声调”的三音素音节模型。但是，文本中除了标准的音节外，还包括一些不规范的书写形式。比如：缩写词语，数学公式，数字，日期，特殊符号等。文本归一化就是对这些不规范的书写形式进行标准化转换，使傣语文本的书写形式与读音形式保持一致[5]。

文本归一化又称为文本正则化，是指对文本进行分析，将文本中的不规范文本识别出来，将其转换为规范文本的过程。这些不规范的文本我们称为非标准词[6]。非标准词在归一化时应进行标准化转换，否则合成语音将会出现语义不完整或者读音不正确的情况。文本中的非标准词，大部分只有一种读音，即不存在歧义。还有一部分文本，在不同的语境下会有不同的读法，即存在歧义。因此，文本归一化不仅仅只是简单的非标准词转换，还涉及到歧义判断与消歧处理。

傣语文本归一化主要包括两方面的任务，一是将非傣语字符转化为傣语字符并确定其读音，二是将傣文数字的字符形式转化为拼音形式，以使文本与录音保持一致性。处理步骤分为四步：非标准词识别、歧义判断、消歧处理、非标准词转换为标准词。其处理流程如图 1 所示。

2.1. 傣语非标准词分析

傣语的标准词必须由傣语声韵母构成，凡是不满足此条件的，都为非标准词。如：傣语数字字符“ᥐ，ᥑ，ᥒ，ᥓ，ᥔ，…”。虽然属于傣语字符，但它不是标准的声韵母形式，所以仍将其归为非标准词。傣语的非标准词主要由一些非标准字符组成。这些非标准字符可分为四类：

- 阿拉伯数字。比如：“0，1，2，3，4，…”。
- 傣语数字。比如：“ᥐ，ᥑ，ᥒ，ᥓ，ᥔ，…”。
- 特殊符号。比如：运算符(+，*，>，…)、网络符号(@，#，/，…)、单位(￥，$，…)。
- 英文字符。比如：“A，B，C，D，E，…”。

这些非标准字符，可以单独组成非标准词，也可以几种字符组合在一起。常见的形式包括：纯数字、数字与特殊符号组合、特殊符号、纯英文、英文与字符组合。

2.2. 非标准词识别

为了方便且准确识别出非标准词，我们建立了傣语标准字符表，包括 42 个声母，96 个韵母以及 2 个显性调符[7]。对非标准词的识别，我们的思路是，将傣文音节逐个与标准字符比对，在声韵母表中能直接匹配的音节标记为标准词，否则标记为非标准词。识别具体过程如下：

1) 建立非标准字符数据库。使用 Access 数据库来存储非标准字符。
2) 连接并读取数据库。使用微软的 OLEDB (Object Linking and Embedding Database)来访问数据库。整个过程包括：

a) 连接数据库。使用 OLEDB 中的类 OleDbConnection 来完成。
b) 读取数据库。根据 sql 语句来读取数据。用 OLEDB 的类 OleDbCommand 来执行 sql 语句。
c) 关闭连接。读取数据后，调用 OleDbConnection 对象的 close 方法关闭数据库连接。

3) 非标准字符比对。将数据库中的数据读取出来后，需要将傣文与标准字符比对。先比对非歧义字符。通过循环，将非歧义字符从变量中逐个取出，并与傣文比对，看傣文中是否包含该歧义字符。接着再比对歧义字符，若包含歧义字符，则后续需进行歧义判断并消歧。

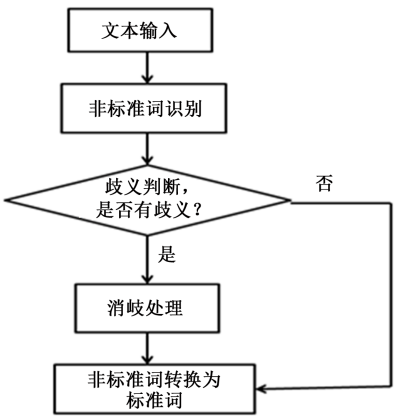

Figure 1. The process of normalization
图 1. 归一化流程

2.3. 歧义判断

歧义非标准词一般是阿拉伯数字与歧义字符的组合。通过非标准词识别，若傣文中包含歧义字符，则对其进行歧义判断，并结合规则与语境进行消歧处理。根据现有的语料及以往的经验，我们总结出了四种常见的歧义类型，包括：纯数字(123)、数字：数字(12：13)、数字－数字(12~13)、数字/数字(12/13)。对于每一种歧义类型，我们通过正则表达式(Regular Expression)与关键字相结合的方式来判断。正则表达式是对字符串操作的一种逻辑公式，指用事先定义好的特殊字符及普通字符组合成一个“规则字符串”，通过这个“规则字符串”来对文本内容进行匹配与筛选[8]。

2.3.1. 纯数字

纯数字的歧义判断主要是判断该数字是数值还是数码。大多数情况下，纯数字按数值来读。但有时数字需要读为数码，如：手机号码、邮编、QQ 号码等。因此需要将数码读法的情况列举出来，并进行判断。常见数码数字的正则与关键字如表 1 所示。

2.3.2. 歧义字符

当数字与歧义字符组合时，数字都是按数值来发音的，我们主要判断歧义字符的读法。常见的组合有以下三种(n1，n2 表示阿拉伯数字)：

- n1：n2。这种形式既可读为时间，又可读为比分。如：12：15。由于其形式比较固定，在进行正则判断时是比较简单的。其正则可编写为："(^|[^0-9])\d+:\d+($|[^0-9])"。
- n1 － n2。这种形式既可以读为范围，又可以读为减法。如：12~15。依据其形式，可编写正则为："(^|[^0-9])\d+−\d+($|[^0-9])"。根据正则匹配，即可判断傣文中是否包含该歧义类型。
- n1/n2。这种形式既可以读为时间，又可以读为分数。如：12/15。依据其形式，可编写正则为："(^|[^0-9])\d+/\d+($|[^0-9])"。根据正则匹配，即可判断傣文中是否包含该歧义类型。

2.4. 消歧处理

通过正则表达式及关键字，我们判断出傣文中哪些地方有歧义，然后对其进行消歧处理。对于纯数字，如果判断出该数字应该读为数码，那么就将其转写为数码读法的傣文；如果是其他歧义字符，则依据歧义判断结果，对其进行消歧。下面将重点阐述数字消歧处理和字符消歧处理。

2.4.1. 数字消歧处理

傣文数字包括阿拉伯数字、傣语数字、傣语读音三种形式。它们三者之间的对应关系如表 2 所示。

如果歧义判断为数码，只需根据表 2 所示的转换规则一一转换即可。如果歧义判断为数值，除了按规则将数字转换成傣语数字外，还需在相应位置添加权重。傣语中的数值权重标识符包括：“ᦉᦲᧇ(十)”、“ᦌᦱᧁ(二十)”、“ᦣᦾᧉ(百)”、“ᦘᦱᧃ(千)”、“ᦖᦹᧃ(万)”。数值的读音形式分为一般形式和特殊形式。对于一般的数值，其读法为“数字+权重”的方式，主要包括以下三种形式：

Table 1. Digital regular recognition and keywords
表 1. 数码的正则识别与关键字

数码类型	实例	正则表达式	关键字
手机号码	13586894560	(^|[^0-9])1[3,4,5,7,8]\d{9}($|[^0-9])	无
邮政编码	650000	(^|[^0-9]) [1-9]\d{5} ($|[^0-9])	ᦝᦲᧇᦵᦟᦂᦲᧃᦵ
QQ 号码	10000	(^|[^0-9]) [1-9]\d{4,} ($|[^0-9])	qq

Table 2. Three forms of Dai digital
表 2. 傣语数字的三种形式

阿拉伯数字	傣语数字	傣语读音	阿拉伯数字	傣语数字	傣语读音
0	᪀	ᩈᩪᩁ	5	᪅	ᩉ᩶ᩣ
1	᪁	ᩉ᩠ᨶᩧ᩠ᨦ	6	᪆	ᩉᩫ᩠ᨠ
2	᪂	ᩈᩬᨦ	7	᪇	ᨧᩮᩢ᩠ᨯ
3	᪃	ᩈᩣ᩠ᨾ	8	᪈	ᨸᩯ᩠ᨯ
4	᪄	ᩈᩦ᩵	9	᪉	ᨠᩮᩢ᩶ᩣ

- 个位数
  - 字符形式与读音形式一一对应，直接替换即可。
- 十位数
  - 十位标志词是 ᩈᩥ᩠ᨷ (/sip55/)，读法为“十位标志词+个位数字”，如 ᪁᪂ (12)为 ᩈᩥ᩠ᨷᩈᩬᨦ。
  - 30~99 读法均为“十位数字+十位标志词+个位数字”，如 ᪄᪆(46)为 ᩈᩦ᩵ᩈᩥ᩠ᨷᩉᩫ᩠ᨠ。
- 百千万
  - 百位标志词为 ᩁᩬ᩠ᨿ(/hɔi11/)，千位标志词为 ᨻᩢ᩠ᨶ(/pan41/)，万位标志词为 ᩉ᩠ᨾᩨ᩠ᨶ(/mɯn35/)，读法均为“数字 + 标志词”，如 ᪂᪉᪄᪉(2949)为 ᩈᩬᨦᨻᩢ᩠ᨶᨠᩮᩢ᩶ᩣᩁᩬ᩠ᨿᩈᩦ᩵ᩈᩥ᩠ᨷᨠᩮᩢ᩶ᩣ。

而对于特殊读法，其转换规则分为以下四类：

- 数字 1 与任意数字组合，若为个位数，则需要变调读为 ᩋᩮᩢ᩠ᨯ(/ʔet55/)。如 ᪁᪁(11)读音为 ᩈᩥ᩠ᨷ ᩋᩮᩢ᩠ᨯ，而不是 ᩈᩥ᩠ᨷ ᩉ᩠ᨶᩧ᩠ᨦ。
- 20 为特殊读法，标志词为 ᩈᩣ᩠ᩅ(/sa:u41/)，读法为标志词+个位数字，如 ᪂᪃(23)为 ᩈᩣ᩠ᩅᩈᩬᨦ，而不是 ᩈᩬᨦ ᩈᩥ᩠ᨷ ᩈᩣ᩠ᨾ。
- 当 1 出现在最高位时，必须省略不读，直接读标志词即可。如 ᪁᪆᪂(162)为 ᩁᩬ᩠ᨿ ᩉᩫ᩠ᨠ ᩈᩥ᩠ᨷ ᩈᩬᨦ。
- 当一个 0 或连续几个 0 出现在除个位以外的其他位上时，均用 ᨸᩣ᩠ᨿ(/pa:i55/)代替，意为“多”，如 ᪂᪀᪀᪆(2006)为 ᩈᩬᨦ ᨻᩢ᩠ᨶ ᨸᩣ᩠ᨿ ᩉᩫ᩠ᨠ。

根据上述规则，傣语数值的归一处理主要分为两阶段：阿拉伯数字转换成傣语数字及添加权重。

2.4.2. 字符消歧处理

对于歧义字符的消歧，我们先设置一个默认读法，将读法更普遍的视为默认读法。如：“12~20”作为范围的情况更多，因此将“-”的默认读法设为 ᨳᩧ᩠ᨦ。这样确保了在没有明显的关键字进行判断的情况下，字符能转换为标准傣语。消歧的思想是利用正则表达式，根据上下文关键字来确定其歧义类型，根据转换规则将非标准词转写成标准词。部分字符消歧规则如表 3 所示。

这里以时间“12：30”为例说明，消歧过程如下：

1. 依据歧义字符“：”，将上述字符串拆分为“12”，“30”。
2. 将歧义字符消歧为“ᨳᩣ”，并合并出字符串“12 ᨳᩣ 30”。
3. 将傣文中的“12：30”作为整体替换为“12 ᨳᩣ 30”。

2.5. 非标准词转换为标准词

傣文的非标准词识别和消歧处理后，需要对傣文中的非标准词进行标准化转换。转换分为两类：数字转换和特殊字符转换。对于数字转换，难点在于对数值权重的添加，除了将数字归一化为标准傣语外，还要在相应位置添加权重。傣语的特殊字符主要包括英文字母、网络符号和数学符号等，部分字符转换

Table 3. Disambiguation rule of characters
表 3. 字符消歧规则

歧义类型	正则表达式	关键字	消歧后的傣语读音
n1：n2	(^|[^0-9])\d+:\d+($|[^0-9])	ᦔᦲᧈ(比)默认读法：时间	n1 ᦎᦱᧉ n2n1 ᦎᦱ n2(时间)
n1/n2	(^|[^0-9])\d+/\d+($|[^0-9])	ᦶᦎᧉ(等于)默认读法：日期	n2 ᦶᦜᧂ n1ᦡᦲᧂ n1 ᦟᦹᧃᦶᦜᧉ n2 ᦓᦲᧉ(日期)
n1-n2	(^|[^0-9])\d+-\d+($|[^0-9])	ᦔᦲᧈᧉ(比)默认读法：范围	n1 ᦟᦹᧄ n2n1 ᦶᦑᧉ n2(范围)

规则如表 4 所示，经过消歧处理后，特殊字符都是不存在歧义的。所以，特殊字符的归一化只需按照转换规则一一替换即可。

3. 傣语文本归一化实现

经过对傣语语料文本的统计，我们发现本文语料文本包括数字和特殊符号两种形式的非标准词，所以，本文主要研究数字归一化和特殊符号归一化。

傣文数字包括阿拉伯数字、傣语数字、傣语读音三种形式，数字归一化的目的就是将数字的字符形式转换成读音形式。阿拉伯数字有两种语义：数值和数码。对于数码形式的归一化，只需根据转换规则直接转换即可，如手机号码“13598689545”其读音形式为“ᦵᦀᦲᧆᦉᦱᧄᦠᦱᧉᦵᦂᦱᧉᦶᦜᧆᦶᦜᧆᦵᦂᦱᧉᦠᦱᧉᦉᦲᦠᦱᧉ”。而对于数值形式的归一化，除了将单个数字转换成读音形式外，还需添加权重标识符。下面我们以傣文“ᦉᦱᧄᦂᦲᧂᦶᦑᧉ，᧒᧐᧑”为例，阐述具体实现过程，上述傣文包含傣语数字᧒᧐᧑(201)其归一化流程如图 2 所示。

1. 非标准词识别

利用正则匹配识别非标准词为“，᧒᧐᧑”。

2. 歧义判断与消歧

通过歧义判断，判断其是数码还是数值，若为数码，按照表 2 所示的规则将其转换成傣语读音形式即可；若为数值，就按照以下步骤进行归一化。

3. 添加权重标识符

利用正则表达式匹配到字符串“，᧒᧐᧑”。确定该字符串索引位置为 12(从 0 开始)。由于该数值是三位数，所以，第一个权重标识符“ᦣᦾᧉ(百)”应添加于“᧒”后，即傣文索引位置为 12 + 2，即 14。第二个权重标识符“ᦉᦲᧇ(十)”应添加于“᧑”之后。由于前面已添加了“ᦣᦾᧉ(百)”，这时傣文为“ᦉᦱᧄᦂᦲᧂᦶᦑᧉ，᧒ᦣᦾᧉ᧐᧑”，所以索引为 20。数值中添加的多个权重可通过循环的方式插入傣文字符串，但是需要计算下次添加的索引位置。这里的 20 等于 14 加上 ᦣᦾᧉ 的字符数 3，加上两个空格符 2 及 ᧑ 的字符数 1。若数值处于句首，由于匹配到的字符串不含“，”索引位置偏移不同。因此，数值处于句中还是句首应分开处理。

4. 处理个位 1(᧑)

添加权重后，首先要处理个位 1。在处理前，首先得判断哪些 1 处于个位。这里也可利用正则进行判断。正则表达式能否准确的匹配出我们需要的字符，关键是能否抓住该字符的核心特征。若我们仅用“᧑”来匹配个位 1，则会将不处于个位的 1 挑选出来。添加权重之后，个位 1 有一个关键特征，其前面必定是标识符“ᦉᦲᧇ (十)”。据此可编写正则：“ᦉᦲᧇ\s᧑[^᧐᧑᧒᧓᧔᧕᧖᧗᧘᧙]”。找出个位 1 后，就可将这里的 1(᧑)替换为“ᦵᦀᦲᧆ”。

5. 处理中间 0

最后处理中间 0，为提高效率，先判断傣文中是否包含“᧐(0)”，包含了才处理。若数值中包含中间

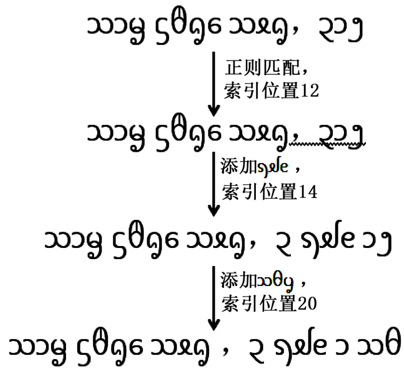

Figure 2. The normalization process of ᦉᦱᧈᦵᦎᦲᧈᦷᦉᧅ，ᦷᦉᧂ
图 2. ᦉᦱᧈᦵᦎᦲᧈᦷᦉᧅ，ᦷᦉᧂ 的归一化流程

Table 4. The transliteration rules of special character
表 4. 傣语特殊字符转写规则

特殊字符	傣语读音	特殊字符	傣语读音
>	ᦜᦴᧂᦜᦱᧃ	A	ᦒᦲᧃᦈᦲᧃ
+	ᦗᦸᧂᦎᦲᧃ	C	ᦏᦲᧃ
=	ᦵᦑᦱᦵᦂᦲᧃ	E	ᦃᦲᧃ
@	ᦒᦲᧃ	G	ᦋᦲᧃ
*	ᦉᦲᧃ	I	ᦚᦲᧃ

0，如：“102”、“1002”、“1022”。为了便于处理，仍然要抓住其关键特征。即不在个位的 0，它们可作为十位，百位，千位，万位，在添加权重之后，相应的字符为“ᦀᦲᧉᦏᦲᧃ”、“ᦀᦲᧉᦣᦸᧂ”、“ᦀᦲᧉᦘᦲᧃ”、“ᦀᦲᧉᦉᦱᧃ”。只要出现上述字符串，那么肯定属于中间 0。只需将上述字符串替换为“ᦜᦲᧂ(多)”即可。若是连续的几个 0，那么替换完后会出现多个“ᦜᦲᧂ”。再将“ᦜᦲᧂᦜᦲᧂᦜᦲᧂ”、“ᦜᦲᧂᦜᦲᧂ”等，进一步替换为“ᦜᦲᧂ”即可。

6. 将傣语数字转为傣语读音

依照规则将数值权重添加后，就需将相应的傣文数字转换为傣文发音。由于两者之间一一对应，且不存在任何包含关系，将傣文数字逐个替换即可。

除了将数字归一化为标准傣语外，还要将一些特殊字符转为标准傣语。经过消歧处理后，剩下的特殊字符都是不存在歧义的。所以，特殊字符的归一化只需按照表 4 所示的转写规则一一替换即可。

4. 实验结果与分析

本文主要讲述了傣语归一化的实现过程，主要分为四个步骤：非标准词识别，歧义判断，消歧处理，非标准词转换为标准词。归一化的难点在于歧义判断与消歧处理。这里歧义判断是依据正则表达式来判断。依据编写好的正则，来匹配出存在歧义的非标准词。当匹配出歧义时，又需要依据相关规则与关键字来消歧。对于纯数字类非标准词，主要分为数码与数值两种读法。大部分的纯数字是读数值的，因此在归一化时还需要依据数值的大小添加“万”、“千”、“百”等权重，这也是归一化时的一个难点。

本文语料文本包括非标准词 3750 个，其中基本非标准词的比例为 62.6%，歧义非标准词的比例为 37.4%。集外测试的语料文本共有 3288 句，其中含有非标准词 8360 个，基本非标准词的比例为 58.5%，

歧义非标准词的比例为41.5%。

采用本文方法，集内测试正确消歧非标准词1335个，消歧正确率为95.2%；正确归一化非标准词3638个，归一化正确率为97.0%。集外测试正确消歧非标准词3230个，消歧正确率为93.1%；正确归一化非标准词8025个，归一化正确率为96.0%。

另外，傣语文本中共包含2571个字符形式的傣语数字，因不存在歧义现象，按规则进行转写后，正确率为100%，达到预期要求。

从实际应用的效果表明，为提高傣语文语转换系统合成语音的质量，本文提出的归一化方法还有进一步改进的空间。在后续研究中不仅可以考虑增大词典的规模，还可以考虑采用跳跃式的词串匹配方式来进行关键词的匹配。

致　谢

首先，感谢云南民族大学的傣族同胞玉腊光罕以及她的同学们，谢谢他们帮助我们校对傣语文本。其次，感谢云大语音实验室的所有成员的关心和帮助，谢谢你们给了我一个温暖的集体和轻松愉快的学习环境。最后还要感谢国家自然科学基金项目对实验室的支持。

基金项目

获国家自然科学基金(61262068)资助。

参考文献 (References)

[1] 戴红亮, 张公瑾. 西双版纳傣语基础教程[M]. 北京: 中央民族大学出版社, 2012.

[2] 玉康, 张秋生, 岩温龙. 西双版纳傣语基础教程[M]. 昆明: 云南民族出版社, 2006.

[3] Gao, L., Chen, Q., Li, Y.H., *et al*. (2010) Several Problems of Text Analysis in Tibetan Speech Synthesis. *Journal of Northwest University for Nationalities* (*Natural Science Edition*), **2**, 1-7.

[4] Hopkins, H. and Edmunds, T. (2016) Broadcast System Using Text to Speech Conversion. United States Patent 9263027.

[5] Haunschild, R. and Bornmann, L. (2016) Normalization of Mendeley Reader Counts for Impact Assessment. *Journal of in Formetrics*, **10**, 62-73. http://dx.doi.org/10.1016/j.joi.2015.11.003

[6] Sproat, R., Black, A.W. and Chen S. (2001) Normalization of Non-Standard Words. *Computer Speech & Language*, **15**, 287-333. http://dx.doi.org/10.1006/csla.2001.0169

[7] 戴红亮. 西双版纳傣语数词层次分析[J]. 民族语文, 2004(4): 22-26.

[8] 邱涛, 王斌, 杨晓春. 利用关键因子过滤的正则表达式匹配算法[J]. 计算机科学与探索, 2016(3): 326-337.