俄语语音识别语料库设计与研究

马延周[1]，易绵竹[2]

（1.解放军外国语学院基础部，河南洛阳，471003；

2.解放军外国语学院语言工程系，河南洛阳，471003）

摘要：现代俄语拥有超大的词汇量和复杂的组成成分，对语音识别研究带来巨大困难，而语音语料库建设又是语音识别的前提和基础。本文采用两种语音语料库的构建方法，从国内外媒体收集俄语语音并由人工标注，以及对选定的俄语文本语料进行朗读并录音，针对俄语语音识别的有关问题，详细阐述了俄语语音语料库的设计思路和建设过程，为俄语连续语音识别提供了研究支撑和真实的实验数据。

关键词：俄语语音识别；语音语料库；俄语声学模型；俄语语言模型

中图分类号：TP391　　**文献标识**：A

Design and Research onRussianSpeech Recognition Corpus

MA Yanzhou[1]，YI Mianzhu[2]

(1. Department of Basic, PLA University of Foreign Languages, Luoyang, Henan 471003, China；

2. Department of Language Engineering, PLA University of Foreign Languages, Luoyang, Henan 471003, China)

Abstract:Large vocabulary and complex composition ofmodern Russianhas brought great difficulties to speech recognition research, but the construction of speech corpus is indispensible to speech recognition since it is the premise and basis. This paper adopts two methods in constructing the speech corpus: Samples from Russian audios with artificial voice annotation from home and abroad,and selected transcripts read by native speakers. It focuses on related issues of Russian speech recognition, describes the design and construction process of the Russian speech corpus, providing a foundation and experimental support for the research of Russian speech recognition.

Keyword:Russian speech recognition；Speech corpus；Russian acoustic models；Russian language models

1 引言

借助计算机分析工具的帮助，语料库可以在语音研究、自然语言处理研究、传统语言研究等方面得到广泛应用，根据研究的用途和目的不同，可以确定不同的语料库类型，主要反映在文本语料库的采集方式和采集原则上，语料库已经成为语音与语言信息处理相关研究的一个重要的必备资源。随着语音技术研究[1]的深入发展，不仅需要大规模的文本语料库来建立语言模型，同样需要语音语料库来建立声学模型。语音语料库相比文本语料库而言，能够对语音的频谱图、声学的相关参数等进行详细记录，还能够对语言相关的句法、韵律等信息进行标注，在语音识别等相关研究领域广泛应用。

俄语语音识别面临的一个关键问题就是语言模型和语音模型的训练，这些都与语料密切相关，语料的选择应用遵循代表性和覆盖率的准则。通过实验已经证明，精心设计的语料比随意采集的语料的识别率能够提高 5%的效率[2]。俄罗斯国家语料库[3-4]的建设始于 2003 年，至 2012 年规模达到 3.64 亿词次，标注信息包括元文本标注、词法标注、句法标注等，包含多个子语料库如深度标注库、平行文本库等，其中口语语料库的建设相对滞后，且在建设的过程中存在诸多如标准不统一等因素，使得进展缓慢。

本文借鉴 RASC863[5]的设计思路和少数民族语音语料库的建设方法[6]，具体研究俄语语音识别语料库的建设方法及在建设过程中存在的问题，采用两种方法相结合：一是采集语音数据然后进行粗标注，选择俄罗斯和中国的电视台、广播电台等媒体，采用 Audition 录音结合 Praat 标注的方法，建设语音语料库[7]；二是根据一定的策略由朗读人进行朗读录音，选择 36 个朗读人对选定的文本语料进行朗读，并对其朗读的语音进行采集保存。针对连续语音识别中的相关问题，调整朗读文本的选择策略以尽可能多的覆盖俄语语音现象，以达到最佳的真实效果。

2 俄语语音识别系统结构

俄语语音识别系统的目的就是把输入的俄语声音信号识别成俄语文本，通常采用基于统计的建模

技术，在俄语语音建模和俄语语言建模同时使用，隐马尔科夫模型(Hidden Markov Model ， HMM)[8-9]方法应用于俄语语音建模，N-Gram 语言模型[10]方法用于俄语语言建模。

基于 HMM 的俄语语音识别系统，主要由语音信号预处理、俄语声学模型、俄语语言模型、俄语语音识别单元等四个模块组成[6,11]，如图 1 所示。

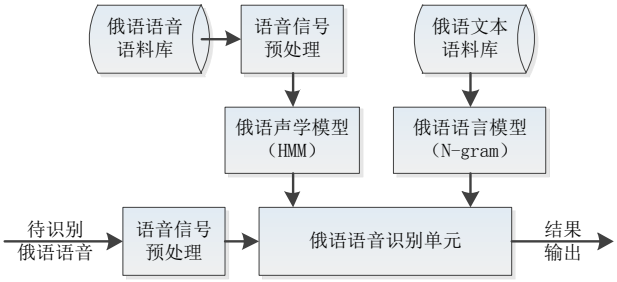

图 1 俄语连续语音识别系统

语音信号预处理与特征提取。预处理主要采用信号处理技术，在可能的范围内，降低环境噪声和信道的影响，同时完成对语音信号的压缩。在综合考虑性能、响应的时间、成本和计算量的前提下，采用什么方式提取特征参数，是研究的重点问题。当前常用的两种提取方式是线性预测（LP[12]）技术和 MEL 倒谱参数[13]技术。梅尔倒谱技术，因为它可以模拟人耳听觉感知的特点，得到广泛的推广和应用。实验结果表明，使用该技术可以提高语音识别系统的性能，Mel 频率倒谱系数已逐渐取代线性预测倒谱，因为它考虑到人类声带和接收声音的效果。

声学模型与模式匹配。声学建模单元和语言建模单位是由发音字典设置并确定它们之间的关系，它包含一组可以处理与识别的单词集及单词的发音。声学模型是系统的重要组成部分，也识别系统的最底层模型，它主要计算语音的特征矢量序列和每个发音模板之间的距离。声学模型的设计与俄语语言学特点有密切的关系，模型单元的大小对俄语语音系统的识别率、训练数据量大小有较大的影响。基于隐马尔可夫模型的统计模型λ (N,M,π ,A,B)是建立识别系统的基础，还包含重新估计参数和识别算法等理论。

语言模型与语言处理。语言建模研究普遍采用是是基于统计的 N-gram 及其变体，包括由语音命令构成的语法网络或由统计方法构成的语言模型，可以对语法、主义等进行处理。其中对大词汇量、连续语音识别系统特别重要，根据语言模型、语法结构等对可能的分类错误进行纠正，特别是必须通过上下文结构才能确定的词义，基于三元统计语言模型是比较成功的模型。它可以限定不同词之间的连接关系，减少搜索空间，有利于减少系统的误识率。

3 俄语语音语料库的设计与建立

面向俄语识别的语音语料库[14-15]的建设方法大致可以分为两种，一种是从不同的渠道采集俄语语音，然后使用语音标注工具进行文本的标注；另一种是根据一定的策略，选择具有代表性的、尽可能全部覆盖俄语语言现象的文本，然后人工朗读这些文本并采集语音。

3.1 采集俄语语音再进行标注

本方法的基本思路与实现过程如图 2 所示：

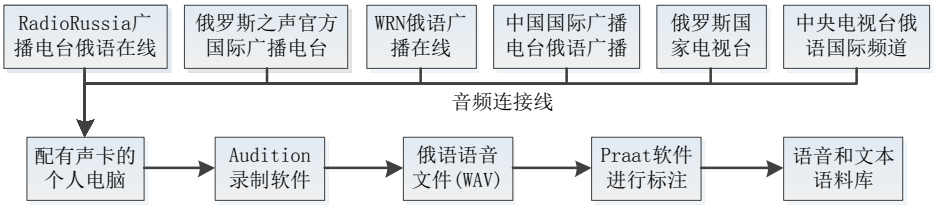

图 2 采集俄语语音再进行标注过程

3.1.1 音源的选择及录制

音源的选择考虑到口音、国别等因素，分别从国内和国外选择具有代表性的俄语广播电台和电视台。分别为：俄罗斯之声 ГолосРоссии、WRN 俄语广播在线、Radio Russia 广播电台、中国国际广播电台俄语

广播、俄罗斯国家电视台、中央电视台俄语国际频道 6 个语音源。通过音频连接线将音源设备与配有声卡的个人电脑进行连接，采用 Audition3.0 软件录音，保存为 wav 格式的音频文件。

3.1.2 语音的切分

语音文件的切分是处理大文件的一个重要步骤，通常采用基于能量和过零率的端点检测算法[16]来进行，最终要将一个大的语音文件切分成时长为 8 秒左右的单独文件并保存。端点检测是俄语语音识别过程中的重点研究内容，它能够确保采集或录制的语音信号是正确的真实的语音信号，而不是噪声信号，这样可以大大减少有效数据量和计算量，并能够提高计算的效率。

3.1.3 标注及保存

对上一步切分的语音进行人工标注，标注采用 Praat 软件，标注信息的准确与否，将影响到识别的结果。因此需要确定符合俄语语言学的规范，如俄文到拉丁文的转换对应标准如表 1 所示、语音数据的标注层级和文本的选择原则。

表 1 俄文字母和拉丁字母转换对照表

俄文	拉丁	俄文	拉丁	俄文	拉丁
А	A	К	K	Х	H 或 KH 或 CH
Б	B	Л	L	Ц	C
В	V 或 W	М	M	Ч	CH 或 TCH
Г	G	Н	N	Щ	SCH 或 SC 或 SHTCH 或 STCH 或 SHCH
Д	D	О	O	Ш	SH
Е	E 或 JE	П	P	Ъ	Ъ 不用写拉丁字母
Ё		Р	R	Ь	J 或不用写拉丁字母
Ж	ZH 或 J 或 V	С	S	Ы	Y
З	Z	Т	T	Э	E
И	I	У	U	Ю	JU 或 IU
Й	J 或 I	Ф	F	Я	JA 或 IA

标注信息：语音信号的采样率，采样数，编码格式信息；语音对应的文本内容；语音段和非语音段的端点位置等。俄语的最小发音单元是音节，一个音节一定包括一个元音。音节可以由一个元音构成，也可以和辅音构成，词有几个元音就有几个音节构成。本系统面向识别研究，采用手工标注的方式只需标注到词和音节两个层次。

本方法选择 36 名俄语专业的学生进行，每 6 名学生一组，采集同一媒体语音，时长大约在 2 个小时，共计约 72 小时的音频文件和标注文件。其中音频文件和文本文件的命名一致，仅后缀名不同，文件名共有 7 位数字组成。1-2 位表示朗读人编号（取值 01-36），第 3 位表示性别编号（取值 0-1），4-6 位表示切分后的单句的编号（取值 000-999）。选用 Praat 软件，标注词和音节两个主要层次。标注 160213.wav 文件如图 3 所示。

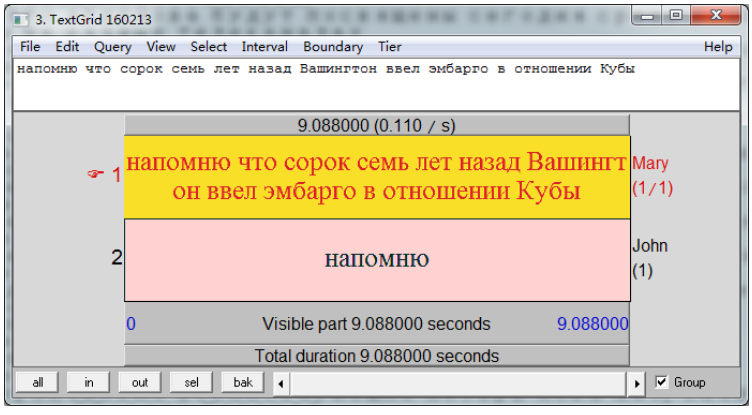

图 3 标注 160213.wav 文件

3.2 选择俄语文本并朗读采集语音

按照某种策略选择俄语文本，然后对其进行朗读并采集语音，建立过程如图 4 所示。俄语语音语料库的构建可分为前期准备、俄语文本的采集、朗读人的选择、音频录制等几个环节。

俄语语音语料库的生产过程可分为准备阶段，俄语文本采集，环境设置和音频录制等步骤。

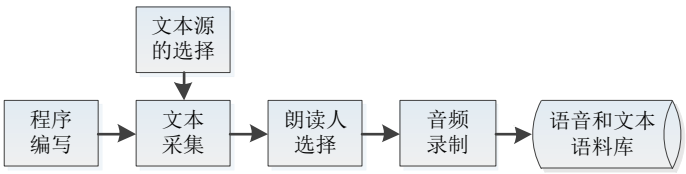

图 4 俄语朗读语音语料库的建立过程

3.2.1 俄语文本源的选择

在对俄语文本进行收集以前，最重要的工作是俄语文本源的选择和程序的编写。由于俄语网站及俄语电子信息的不断增多，利用计算机程序来收集大规模的俄语文本语料成为可能，首先选择俄罗斯具有代表性的网站，对其新闻语料进行选定。如表 2 所示：

表 2 网络文本语料选定范围

门户网站	http://news.rambler.ru/	俄罗斯门户网
搜索网站	http://www.yandex.ru/	俄罗斯 Yandex 搜索引擎
新闻网站	http://www.ruvr.ru/	俄罗斯之声
	http://www.izvestia.ru/	俄罗斯消息报
	http://www.rutv.ru/	俄罗斯国家电视台
	http://rian.ru/	俄罗斯新闻社
	http://www.itar-tass.com/	俄通社-塔斯社

从这些网站上确定搜集的内容，如政治、经济、军事、社会、互联网等方面新闻内容，然后分析网页构成，利用 C#编写爬虫程序，通过过滤、去格式化、查重、替换等步骤，提取其中的文本内容，以 utf-8 的格式保存为纯文本文件，一个页面内容保存为一个 txt 文件，然后分类保存在一个文件夹中。

3.2.2 文本语料的采集

需重点考虑以下几个问题：(1)音节的覆盖，识别系统中的每一个最小识别单元都应该出现在所设计的语音语料中。要保证声学模型训练的精确，也可以要求识别系统中每一个最小的识别单元在语料中出现的次数要大于一定值。(2)音节的均衡，每个音节单元在语料中出现的次数与别的音节单元相比较，不能出现太大偏差。合理的音节平衡能够在确保音节覆盖率的基础上，有效地控制语音语料库的规模。(3)为了确保文本语料中句子的连贯和自然，语音语料库中的文本应该最大限度的采用真实语料。

基于以上考虑，收集了大约 80M 的俄语纯文本语料。语料的来源主要包括俄文文学小说电子版、俄语商务口语文本语料、日常使用对话文本语料、大学俄语教辅材料和俄罗斯主要网站的新闻，其构成比例如图 5 所示。从收集的语料中筛选尽可能包含所有俄语词、词根、词缀、元音、辅音、音素、音节等具有代表性的句子 7200 个。

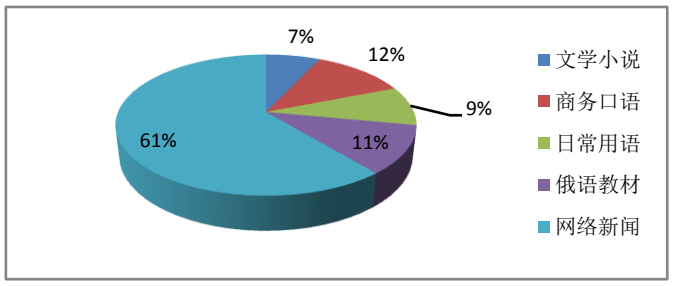

图 5 俄语文本语料的组成

3.2.3 朗读人的选择

朗读人的选择要考虑其性别，年龄，教育背景等因素，因为这些因素会影响语料库的整体质量。本数据库包含 36 位不同朗读人，来自国内不同省份、年龄分布在 18 到 56 岁。所有朗读人均为母语为汉语，大学本科专业为俄语的在校学生或老师。性别比例均衡，男女比例为 20：16。

将 7200 个句子分成 36 组，每组 200 个句子。其中 140 个句子由长约 15 个俄语单词构成，60 个句子为俄语中的人名、地名、日期、时间。每个句子朗读 1 遍，保存为 1 个音频文件。文件命名同前一种

方法。如 560222.wav(txt)。表示编号为第 56 号的朗读人，性别为女，朗读第 222 号句子。

3.2.4 音频录制

本次录音使用 Windows XP 操作系统，联想笔记本自带的集成声卡，软件采用 Audition3.0 录音。打开软件，切换到录音界面，将录音系统连接好，正式录音开始前，测试朗读人音量大小及口语流利程度，并根据朗读人的音量大小调试设备到最佳录音状态。音频录制过程中，实时监控朗读人的音量大小及朗读的准确度。图 6 为录制 531225.wav 音频文件的截图。

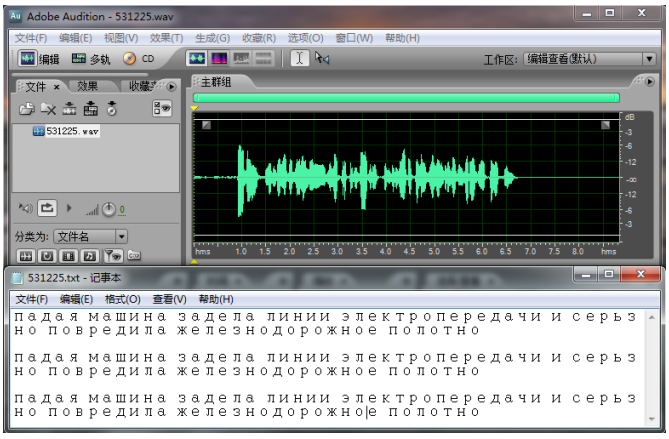

图 6 录制 531225 音频文件的截图

4 俄语语音语料库建设关键问题

语料库在俄语语音识别过程中起着举足轻重的作用，语料库建设的质量也关系着识别效果的高低，在俄语语音语料库的建设过程中也存在着以下几个关键问题：

1）俄语的词汇复杂多变。

俄语属于东斯拉夫语支，词汇量大且形式复杂来源广泛，词形的变化反映了词之间的复杂关系和词在整个句子中语法功能。名词词形有 12 个，单数和复数共有 12 个格；动词有 100-200 个形式，包括时、态、形动词、副动词等；形容词有近 30 种形式，二十多个格。实词分为词干（语义）和词尾（语法），且词尾的意义包含多个。

从俄语语法，构词特征可以看出，俄罗斯拥有大量的词汇，对俄语语音识别的研究带来了很大的困难，尤其是在语言模型的建立时很难获得足够的词汇。

2）语言模型和声学模型的建立需要考虑俄语语言学和语音学的特征。

建立俄语声学模型和语言模型，需要考虑俄语语音学和语言学的相关知识。俄语分为元音（6 个）和辅音（36 个），主要特点有：少量元音，大量辅音；辅音中浊清相对和硬软相对如表 3 所示；非重音节中元音发音会发生弱化现象，有时很模糊；不同的单词的重音落在不同的音节，位置不固定，形态变化时重音位置可以发生移动。

表 3 俄语清浊辅音对照表

清辅音-发音时声带不振动
[п]、[п’]、[ф]、[ф’]、[к]、[к’]、[т]、[т’]、[с]、[с’]、[ш]、[щ]、[х]、[х’]、[ц]、[ч]
浊辅音-发音时声带振动
[б]、[б’]、[в]、[в’]、[г]、[г’]、[д]、[д’]、[з]、[з’]、[ж]、[м]、[м’]、[н]、[н’]、[р]、[р’]、[л]、[л’]、[j]

俄语拼音规则的特殊性，如重音的变化、元音的弱化、清浊辅音的转化、某些音组在词中的不规则发音和句子的语调等，造成了复杂多变的语音现象，所以在选择俄语文本语料时，尽可能包含所有这些音节和特殊的语音现象。进行俄语语音识别时，采用 HTK 工具和方法进行特征提取、建立语言模型和声学模型，同时必需考虑俄语语言学和语音学的特征，特别是重音位置的判断。

3）俄语文本语料的设计与收集。

文本语料库的建立需要在目标明确的前提下，经过精心的设计才能达到预期。在对语料的内容、语料的来源、语料库中的文本类型、语料库的结构进行设计时，应以是否有助于俄语语音识别为准则进行。理想的情况是，语料库的统计样本覆盖所有可能的语音语料库，以尽可能小的数据涵盖语言现象。然而，

这样的语料库建设是不容易的，它需要有足够的语料作为取样的基础。

语料库不能无限大，有目的地设计并选择恰当的语料，对于训练有较强鲁棒性的语言模型和声学模型有着重要的作用。确保包括尽可能多的语言和语音现象，提高语音模型的鲁棒性，同时避免声学模型训练数据稀疏的问题，因此要确保系统中每个有可能出现的音素单元在声学模型的训练中都最大限度的得到足够充分的训练。由于俄语期刊、报纸等文本数据收集比较困难，又没有足够的历史数据可供参考，通过网络下载的文本语料需要经过程序来转换提取，所以这项任务非常艰巨。

5 总结

本文初步建立了一个适合俄语语言识别研究的语音语料库，为俄语语音技术的研究提供了真实的数据资源，语音语料库和文本语料库建成后，可以使用 HTK 工具建立相应的俄语声学模型和语言模型，搭建连续语音识别系统。

俄语语音语料库的建设是一个漫长复杂而艰巨的工作，需要在多个方面不断加以完善。首先在俄语语料选取中，要加大原始真实语料的来源范围和信息量，使选取的语料能覆盖更多的语音现象，以扩大整个语音语料库的信息量。在朗读人方面，选择更多的来自不同地域、文化知识背景、生活习惯、年龄段的朗读人。在语音录制方面，增加录音文件的保存格式与采样频率，以增加语音库容量，使语音语料库更具有实用性。最后在语音标注时，要开发集成录音和标注等功能的 B/S 或 C/S 模式软件，提高标注的自动化程度，增加相应的字段来描述文件的质量和用户的评价，以方便后续的应用。完成这些方面的工作，需要语言和语音学、心理学、计算机科学等多学科的相互配合。

参考文献

[1] 韩纪庆,张磊,郑铁然.语音信号处理[M].北京:清华大学出版社,2013:241-255.
[2] 吴华,徐波,黄泰翼.基于三音子模型的语料自动选择算法[J].软件学报,2000(2):271-276.
[3] 李绍哲.俄语语料库和基于语料库的语法研究[D].[哈尔滨]:黑龙江大学,2012.
[4] ЕА Гришина.Устная речь в Национальном корпусе русского языка[J].Национальный корпус русского языка,2003,2005(2):94-110.
[5] 李爱军,王天庆,殷治纲.863 语音识别语音语料库 RASC863—四大方言普通话语音库[C]//第七届全国人机语音通讯学术会议（NCMMSC7）论文集,2003:274-277.
[6] 那斯尔江·吐尔逊,吾守尔·斯拉木,麦麦提艾力.维吾尔语大词汇量连续语音识别研究——语音语料库的建立[C]//民族语言文字信息技术研究——第十一届全国民族语言文字信息学术研讨会论文集,2007:379-385.
[7] UtpalBhattacharjee,KshirodSarmah.Development of a Speech Corpus for Speaker Verification Research in Multilingual Environment[J].International Journal of Soft Computing and Engineering (ijsce),2013,2(2):443-446.
[8] Lawrence Rabiner.A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition[J].Proceedings of the Ieee,1989,77(2):257-286.
[9] JD Ferguson.Hidden Markov Analysis: an Introduction[J].Hidden Markov Models for Speech,1980(1):8-15.
[10] Peter F Brown,Peter V Desouza,Robert L Mercer, et al.Class-based N-gram Models of Natural Language[J].Computational Linguistics,1992,18(4):467-479.
[11] Steve J Young,PCWoodland.State Clustering in Hidden Markov Model-based Continuous Speech Recognition[J].Computer Speech & Language,1994,8(4):369-383.
[12] HynekHermansky.Perceptual Linear Predictive (plp) Analysis of Speech[J].The Journal of the Acoustical Society of America,1990,87(4):1738-1752.
[13] VivekTyagi,IainMccowan,HemantMisra, et al.Mel-cepstrum Modulation Spectrum (mcms) Features for Robust Asr[C]//Automatic Speech Recognition and Understanding, 2003. Asru'03. 2003 Ieee Workshop on,Ieee,2003:399-404.
[14] Ирина Сергеевна Кипяткова,Алексей Анатольевич Карпов.Автоматическая обработка и статистический анализ новостного текстового корпуса для модели языка системы распознавания русской речи[J].Информационно-управляющие системы,2010(4):1-7.
[15]ОН Ляшевская,ВА Плунгян,ДВ Сичинава.О морфологическом стандарте Национального корпуса русского языка[J].Национальный корпус русского языка,2003,2005(2):111-135.
[16] 路青起,白燕燕.基于双门限两级判决的语音端点检测方法[J].电子科技,2012(1):13-15+19.

马延周（1977-），博士研究生，研究方向为计算语言学和语言信息处理，E-mail：myz827@126.com。
通讯地址：河南洛阳涧西区解放军外国语学院基础部计算机与网络教研室　471003　13526980276

易绵竹（1964-），教授，博士生导师，研究方向为计算语言学和语言信息处理，E-mail：mianzhuyi@gmail.com。