朗读语料与自然口语的差异分析①

刘亚斌　李爱军

中国社会科学院语言研究所

lyabin@sina.com　liaj@linguistics.cass.net.cn

摘要

本文通过对朗读语音语料库 ASCCD、自然口语独白语音语料库 CASS 和自然口语对话语音语料库 CADCC 的统计分析，试图说明朗读语料与自然口语的主要差异。文章主要对二者在音节、声韵、副语言学和非语言学现象、语篇话题、话轮转换、基频变化以及音段音变现象等几个方面作了一些统计分析，并由此归纳出朗读语料与自然口语的几点不同。

1. 引言

近几年来，随着语料库语言学的兴起，各种各样的语料库也分别建立起来，其中主要是朗读语料库。随着研究的深入和语音处理技术的发展，朗读语料库已经不能满足需要，人们希望对真正的自然口语进行研究和处理，自然语料库的建立也就逐渐的开展起来，自然口语的研究日益成为语言学家和言语工程人员面对的一个重要课题。

为此，我们做了大量的工作，先后建立了朗读语篇语音语料库 ASCCD[1]、自然口语独白语音语料库 CASS[2]、和自然口语对话语音语料库 CADCC（包括 SET-1 和 SET-2）[3]，并且花费了大量的时间和人力进行了切分和音段、韵律标注，在此基础上开展了一系列的研究工作。本文的目的是通过对这三个语音语料库的主要数据的统计分析，来说明朗读语料和自然口语的主要差别。

2. 语料库简介

2.1. 语料库设计

朗读语篇语音语料库 ASCCD 共有 18 篇，文本语料是经过语言学家挑选的具有丰富语篇信息的议论体和叙事体语篇，由 10 个发音人（5 男 5 女）在录音室里朗读，每篇约有 300-500 字不等。

自然口语独白语音语料库 CASS 的原始语音是由清华大学广播站提供的录音磁带，内容主要是学校的讲座、学生自由讨论和一些公共会议，其中有对话，但大部分是独白。讲话人没有讲稿，所以是随意口语，因此含有大量口语现象和丰富的音变现象。录音环境是普通的教室、会议室和礼堂，录音设备也不是专业录音设备，所以背景噪音较大。该库共有 6 个小时的语音，我们选择了三个多小时的语音进行标注。

自然口语对话语料库 CADCC 包括两个子库，其中 SET 1 是电话对话库，SET 2 是正常通道对话库。表 1 给出了 CADCC 的详细信息。SET 2 中共有 13 对发音人，对话双方是同事或同学，有共同的爱好或话题，谈话内容不限，也就是语篇话题可以自由转换。其中有 8 位发音人曾参加过朗读语篇 ASCCD 的录音，这样以便详细对比朗读和自然口语的各种差异。录音在普通办公室或宿舍进行，对话者身带无线话筒，无线录音设备放置在另外的房间，这就保证了对话双方完全进入自然谈话状态。每一对发音人的谈话时间在 1 个小时左右。

①本课题得到国家社科基金、国家 973 基金和中国社会科学院语言所重点课题基金支持

2.2. 语料库转写和标注

我们对 ASCCD 和 CADCC 的 SET 1 都进行了音段和韵律标注，对 CASS 进行了音段标注，标注工具是 Praat 和 XWAVES+，音段标注采用 SAMPA-C 音段标注系统，韵律标注采用 C-ToBI 韵律标注系统，详见[4]。

ASCCD 和 CADCC 的标注有 7 层：正则的音节和声调标注、声韵母标注、韵律结构标注、重音结构标注、语句功能类型标注、杂类标注和话轮标注。CASS 标注信息有 3 层：音节层、声母/韵母层和杂类层。

在声韵母标注层，用 SAMPA-C 音段标注系统来标注实际发音，包括标注超音段特征（声调的变化、上上相连的变调和轻声变化）和音段特征（增音、减音、浊化、清化、喉化、送气化、成音节化、音素音变等等）；杂类层主要标注背景噪音、口语现象等非语言学和副语言学现象（见表 2），由于口语语气词和叹词在口语功能研究中的重要性，我们也在这层中标出。另外，我们还对 CADCC 的所有语音都进行了汉字转写，并将口语的非语言学和副语言学现象也按照表 2 的符号进行了转写，在文字转写中还标记了语篇话题（非句子话题）转换的位置。

3. 统计数据和分析

3.1. 音节统计

本文对 ASCCD 和 CADCC 两个库中音节的出现次数进行了统计，包括带声调和不带声调音节的统计。表 3 给出了出现次数最高的前 15 位音节（计声调和不计声调），其中 C-SET2 代表 CADCC 的 SET 2 子库（下同）。从表中可以看出，在口语中，“你、我、他、这、那”这几个代词的使用频率是相当高的，而朗读语料由于是人为设计的，就不存在这种现象。而“de0（的）,le0（了）,shi4(是),you3（有）,bu4（不）”在朗读语料和自然语料中都出现较多，这点也符合人们平时说话的习惯。另外，表中所列两库中的高频音节所对应的汉字与《现代汉语频率词典》[5]中的高频字表也极为吻合，几乎全在其前 35 个高频字之内。

3.2. 声韵母出现率统计

陈肖霞曾经对 863 朗读语料库的部分语料（1560 句）做过声韵母出现率统计[6]，本文对自然语料库 C-SET2 也做了声韵出现率统计，并与之作了对比（见表 4）。

表 1: CADCC 的详细信息

	Set1 (电话库)	Set2 (对话库)
内容	旅馆服务	不限
总时长	2 小时	16.2 小时
发音人	》200 人	16 男 & 10 女
口音	√	×
汉字转写	√	√
语音学标注	韵律和音段	韵律和音段
采样率	8 KHz	16 KHz
存储形式	.wav	.wav

表 3: 出现率最高的前 15 位音节 (计调和不计调)

计声调		不计声调	
ASCCD	C-SET2	ASCCD	C-SET2
de0	shi4	de	shi
shi4	de0	shi	de
le0	na4	yi	na
ren2	jiu4	you	yi
shi2	bu4	qi	ge
yi4	yi1	zhi	jiu
you3	wo3	le	zhe
zai4	zhe4	ren	bu
sheng1	ta1	bu	wo
bu4	ni3	ji	ta
da4	ge4	jiu	ni
yi2	le0	zai	le
he2	shuo1	guo	you
yi1	me0	zhe	er
wo3	you3	sheng	shuo

表 4: 出现率最高的前 10 位声母和韵母

声母		韵母	
朗读语料	C-SET2	朗读语料	C-SET2
j	s	e	e
sh	sh	i	a
zh	d	u	i
d	y	i2	uo
l	z	ong	en
x	n	ian	ai
g	j	ai	an
h	b	a	ou
b	g	uo	ao
z	zh	ing	ui

表 2: 非语言学和副语言学现象转写符号

NO.	现象		符号	
			开始	结束
1	副语言学现象	拉长	LE <	LE>
2		喘气	BR <	BR>
3		笑声	LA <	LA>
4		哭声	CR <	CR>
5		咳嗽	CO <	CO>
6		犹豫	DS <	DS >
7		口误	ER <	ER>
8		静音	SI <	SI >
9		含混音	UC <	UC >
10		语气词	MO <	MO >
11		咂嘴	SM <	SM >
12		非汉语词汇	NC <	NC >
13		吸气	SN <	SN >
14		打哈欠	YA <	YA >
15		叠接	OV <	OV >
16		插话	IN <	IN>
17		吞咽声	DE <	DE>
18		清嗓	HA <	HA>
19		打喷嚏	SE <	SE>
20		填充停顿	FP <	FP>
21		颤音	TR <	TR>
22	非语言学现象	噪音	NS <	NS >
23		平稳噪音	TN <	TN >
24		电话忙音	BP <	BP>

表 5: 每对发音人的语篇话题和时长分布

发音人	性别	话题数	总话题时长(秒)	话题平均时长
SUNXI	男	42	6524.859	155.3538
ZHNGJ	男	35	4197.666	119.9333
XINGY	女	14	1953.768	139.5549
CHENX	女	13	3062.002	235.5386
DUYU	女	27	4002.587	148.244
SONGW	男	11	2911.886	264.7169
LIUJI	男	10	2718.652	271.8652
LVJIN	女	27	4177.172	154.7101
XUCHA	男	15	6221.138	414.7425
DURUI	男	32	5859.053	183.0954
TANJI	男	7	2531.122	361.5889

由表可见，863 朗读语料中的高频声母与自然口语中的基本相同，而韵母则有较大不同，863 朗读语料由于经过人为的挑选，其高频韵母较为平均的由单元音、二合元音和三合元音构成，而自然口语中的高频韵母则几乎全是由单元音和二合元音构成，其中二合元音居多。

3.3. 非语言学和副语言学现象统计

表 6 是非语言学和副语言学现象的统计结果，黑体是出现率较高的现象。由于在朗读语料中极少出现这些口语现象，故本文仅对自然语料库 CADCC 进行了统计。由表可见，自然语料中的口语现象非常丰富，并且出现的频率也相当高。而其中出现最多的是语气词、拉长和叠接，三者之和分别占到总数的 66.20%和 51.79%。另外，笑声和犹豫在对话库中的出现率也较高，分别占到 4.48%和 17.11%。这就是说，人们平时说话时的表现形式是丰富多彩的，我们建自然口语库时必须充分考虑到这一点，这些现象是口语自然度的标志之一。

3.4. 语篇话题统计

我们在口语对话库 C-SET2 的文字转写中标记了语篇话题转换的位置。所谓语篇话题，是指对话双方所说的某一段话所围绕的一个中心话题，而不是每句话的句子话题。语篇话题的组织呈非线性形式，即一个语篇话题可以被另一个打断，然后在一定的位置重新开始。

表 5 中给出了每对发音人的语篇话题数（子话题除外）和时长分布。假定不考虑语速等个人特性，图 2 给出了不同长度的语篇话题数分布情况，其平均长度是 185.96 秒。

表 7 是语篇话题出现的累计频率，80%的话题短于 4-5 分钟，90%的话题短于 5-6 分钟。

由上述数据可知，自然口语具有很强的随意性，对话双方为了维持谈话的继续，思维具有很大的跳跃性，因此话题转换的速度很快。

3.5. 话轮统计

话轮交替机制是话语分析的一个重要课题。

我们在 CADCC 中发现了各种话轮交替机制。图 1 中左边一列是考察两个话轮 AB 的出现模式图，右边一列是考察话轮 ABA 的出现模式图。表 8 统计了两个子库的话轮出现情况。在 SET 1 中总共有 3256 个话轮，其中有 315 个是有叠接（overlap）的话轮，40 个是插入式叠接的话轮，所以叠接的话轮占总话轮的 20.6%（(2*315+40)/3256）；SET 2 中有叠接的话轮占 10.6%(986*2/18608)。

这里，电话库的覆盖现象比会话库要多，大概是因为人们打电话时交流的信息相对集中，为了节省一些电话费，说话速度很快，只要听明白对方的主要信息，不等对方说完马上回应。同时这也表现出自然话语很强的上下文相关性。

这种叠接，是一种复杂的口语现象。说它复杂，是因为其产生原因是多方面的，大致有以下几种：一是谈话双方地位悬殊，地位高者就有权依自己意愿而随意中断地位低者的话语；二是插话者急于表述自己的观点，而对另一方所说的话没兴趣或没有耐性听完；三是插话者已经从对方说话时表达出来的各种信息（包括表情、体态、语调等）推断出来对方的意图或观点，认为没有必要或是没有时间（如打电话时）继续听对方把话说完；四是该叠接仅是一种必要的反馈信号，其功能根据具体情况可分为表理解、表态度或表收到[7]。之所以说这是一种口语现象，是因为我们还发现在朗读语料和人与机器之间的对话中不存在这种叠接现象。朗读语料里自然不会出现叠接现象，而在人与机器的对话中也没有叠接，这里也有两个原因：一是由于人在潜意识中不把机器看作是与自己同等地位的说话对象，因此不会像与人对话那样自然，也不会和机器抢着说话（认为抢了也是白抢）；二是目前机器说话的智能性较差，在表情、体态和语调等方面还十分欠缺，并且不能提供及时有效的反馈信号，人们很难从前面的话推断出后面的内容或者得到必要的反馈信息，从而也就不会和机器抢话了。

3.6. 基频统计

图 2 是发音人 WJC 的一段话在朗读和自然口语两种情况下，其主要韵律短语的 F0 最大值和最小值的变化情况。可见，自然口语的主要韵律短语的基频变化范围比朗读语料大，基频上限的变化比下限的变化更大。

3.7. 音段音变统计

自然口语的语速较快，又兼口语现象较多，因此自然口语中含有大量的音变现象，如增音、减音、音素音变、央化、鼻化等。

表 9 是对自然口语语音语料库 CASS 近 4 个小时的标注结果进行的音段音变统计结果[2]。表中数据是指实际发音和正则发音相比，有音变的单元所占的百分比。由表可见，在 CASS 中，从声韵来看，声母的音变率较高，平均有 27.46%发生了音变，而韵母也有 12.02%发生了音变；从音节来看，有 29.24%发生了音变。这表明，自然口语中的音变现象非常多，而声母较之韵母更容易发生音变。

产生音变的原因归结为以下几方面：

⑴语速：讲话速度是影响音变的一个很重要的因素，一般来说语速越快，音变现象越多。

⑵个人讲话方式：比如有些北京人，发音习惯很“懒”，舌和唇在发音时不到位，致使音变出现率高。

⑶方言的影响：有的发音人的某些发音，受到自己方言背景的影响，会产生音段和超音段的音变。

⑷韵律和语境的影响：比如零声母如果处于韵律边界的起始位置，一般会变成一个擦音声母、无擦通音或喉塞音，很少出现真零（吴宗济 刘铭杰，1991）；协同发音的影响[8]。

⑸词汇影响：高频词较易发生音变。

表 7: 语篇话题出现的累计频率

时长(秒)	话题出现频率
<160	60%
<200	72.3%
<260	83%
<320	93.6%
<620	100%

表 6: 副语言学和非语言学现象统计

	现象	符号	C-SET1	C-SET2
1	**拉长**	[LE]	**669**	**1302**
2	**喘气**	[BR]	43	**646**
3	**笑声**	[LA]	40	**1082**
4	**哭声**	[CR]	0	0
5	**咳嗽**	[CO]	8	90
6	**犹豫**	[DS]	17	**4135**
7	**噪音**	[NS]	**284**	**596**
8	**静音**	[SI]	**不计**	**685**
9	**含混音**	[UC]	**633**	**2223**
10	**语气词**	[MO]	**2057**	**8310**
11	**咂嘴**	[SM]	80	**568**
12	**非汉语词汇**	[NC]	11	10
13	**吸气**	[SN]	5	175
14	**打哈欠**	[YA]	6	12
15	**叠接**	[OV]	**315**	**2904**
16	**插话**	[IN]	68	**1337**
17	**吞咽声**	[DE]	5	52
18	**清嗓**	[HA]	9	40
19	**打喷嚏**	[SE]	6	2
20	**电话忙音**	[BP]	**338**	0
21	**话题开始**	[TP]	**未标**	206

表 8: C-SET1 **和** C-SET2 **中的话轮统计**

话轮		出现次数
C-SET1	A	1635
	B	1621
C-SET2	A	9284
	B	9324

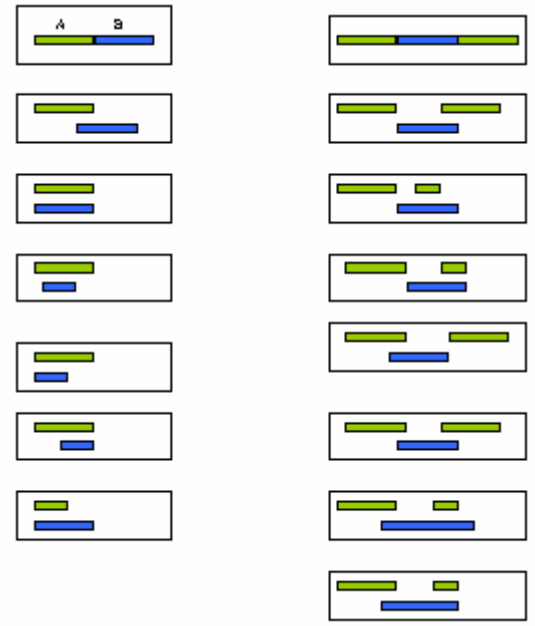

图 1. CADDC **中的话轮交换图。左边一列是考察话轮** AB **的出现模式图，右边一列是考察话轮** ABA **的出现模式图。**

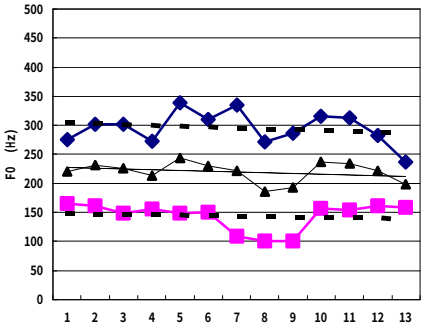

图2. A **朗读方式下韵律短语**F0**变化**

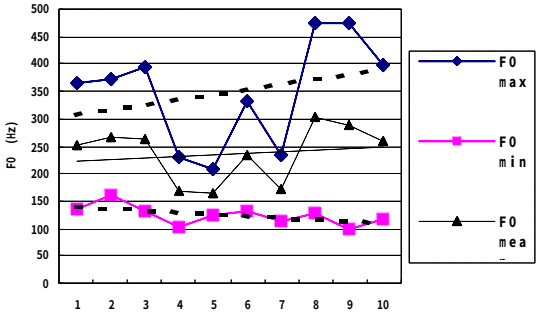

图2. B **自然口语方式下韵律短语**F0**变化**

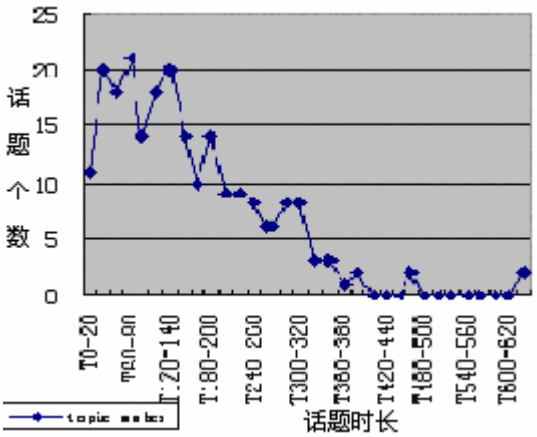

图 2. **话题长度的分布情况**

表 9: CASS **中不同单元的音变统计结果**

发音人	声母音变(%)	韵母音变(%)	音节音变(%)
F03	29.5	25.4	20.9
F04	34.4	22.7	24.1
M01	17.7	20.3	14.2
M02	31.3	22.8	22.0
M03	40.2	23.6	27.8
M04	45.5	24.5	32.9
M12	25.5	24.9	17.6
平均	27.46 (11040/40209)	12.02 (5647/46999)	29.24 (13742/46999)

4. 结论

本文以朗读语音语料库 ASCCD、自然口语独白语音语料库 CASS 和自然口语对话语音语料库 CADCC 为基础，对语料中的音节、声韵、副语言学现象、语篇话题、话轮、韵律以及音段音变等方面进行了对比分析，初步得出了朗读语料和自然口语的几点差异。

(1) 音节出现频率：“你、我、他、这、那”几个代词在自然口语中出现频率较高，“de0（的），le0（了），shi4(是)，you3（有），bu4（不）”在朗读语料和自然口语中使用都较多。

(2) 声韵母出现频率：朗读语料与自然口语中的高频声母基本相同，而韵母则有较大不同。朗读语料中高频韵母较为平均的分别由单元音、二合元音和三合元音构成，而自然口语中的高频韵母则几乎全是由单元音和二合元音构成，其中二合元音居多。

(3) 副语言学现象：自然口语中的口语现象非常丰富，而朗读话语很少出现咳嗽、笑声、犹豫等副语言学和非语言学现象。

语篇话题：自然口语具有很强的随意性，话题转换较快，且呈非线性变化。

(4) 话轮转换方式：自然口语中出现了各种各样的话轮转换形式，并且有大量的叠接现象，大量的叠接的产生原因是多方面的。

(5) 韵律的声学表现：自然口语的基频变化范围比朗读语料的变化范围大，上限的变化更大；口语的语速较快，在我们的语料中，自然口语的语速是朗读语料的 4 倍。

(6) 音段音变现象：和朗读语料相比，自然口语中存在大量的各种各样的音变现象，其中声母的音变率比韵母更高。

(7) 除了上面这些差异以外，另外一个很显著的区别表现在句法方面：口语中存在大量的省略句，而朗读语料中出现最多的句式是 SPVO (subject phrase + verb + object)。

由于自然口语语料库 CADCC 刚刚建成，因此本文只是做了一些初步的统计分析，下一步我们将用机器自动挑选部分语料进行详细的音段和韵律标注，以便展开更为深入的研究。

参考文献

[1] Li Aijun, Chen Xiaoxia, et al (2000), "The phonetic labeling on read and spontaneous discourse corpora", *ICSLP'2000*.

[2] Li Aijun, Zheng Fang, etc. “CASS: A phonetically transcribed corpus of spontaneous speech”, *Report of Phonetic Research*, 2000.

[3] 李爱军，徐波等，“口语对话语音语料库 CADCC 和其语音标注”，第五届全国现代语音学会议，2001

[4] Chen, Xiaoxia Zu, Yiqing & Li Aijun), "A cardinal labeling system for Standard Chinese", *the proceeding of the 4th National Modern Phonetics*, edited by Lv shinan, published by JinCheng Publishing House,1999.

[5] 《现代汉语频率词典》，北京语言学院出版社，1986

[6] 陈肖霞，祖漪清，“基于连续话语语料库的语音音段的初步统计分析”，语音研究报告，1998

[7] 吴平，“汉语会话中的反馈信号”，《当代语言学》，2001，第 2 期

[8] 孙国华，“连续话语中的减音研究”，第五届全国现代语音学会议，2001。

[9] Li Aijun, etc.,"A national database design for speech synthesis and prosodic labeling of standard Chinese", In *proceedings of oriental COCOSDA'99*, TaiPei, TaiWan.

[10] Li Aijun, "The study on the phrasal and sentential accent", *the proceeding of the 4th National Modern Phonetics*, edited by Lv shinan, published by Jin Cheng Publishing House,1999.

[11] Lin Maocan, "Breaks and prosodic phrase in Standard Chinese", *Contemporary Linguistics, No.4,2000.* (In Chinese)

[12] Lin Maocan, "Hierarchical Stress and F0 Restructuring in Utterance of Standard Chinese---One of the Cues to Chinese", *Intonation proc. of SFSSLP'2000.*

[13] www.praat.org

[14] Gillian Brown and George Yule, *Discourse analysis*, Cambridge University Press,1983

COMPARATIVE ANALYSIS BETWEEN READ AND SPONTANEOUS SPEECH

Liu Yabin, Li Aijun

Institute of Linguistics, CASS

lyabin@sina.com, liaj@linguistics.cass.net.cn

ABSTRACT

From the development of language, spontaneous speech is an archaic, common used and typical form of the language. In the past decades from 50s to 80s of the 20th century, we focused on read speech to do our research in three fields: acoustics, psychology and physiology. In the recent 10 years, the research on spontaneous speech is becoming more and more important for the speech applied technology and the associated theories. Spontaneous speech rather than read speech is one of the unresolved problems faced by many speech recognition systems.

Many differences exist between read and spontaneous speech in Chinese on linguistic and phonetic aspects, such as prosodic and segmental variability, turn-taking, discourse topics and paralinguistic phenomena. This paper gives some illustrations and then depicts the research on read and spontaneous speech by analyzing the annotated read speech corpus ASCCD and spontaneous speech corpus CASS and CADCC.

Keywords ： speech corpus; spontaneous speech; read discourse; prosody; segment