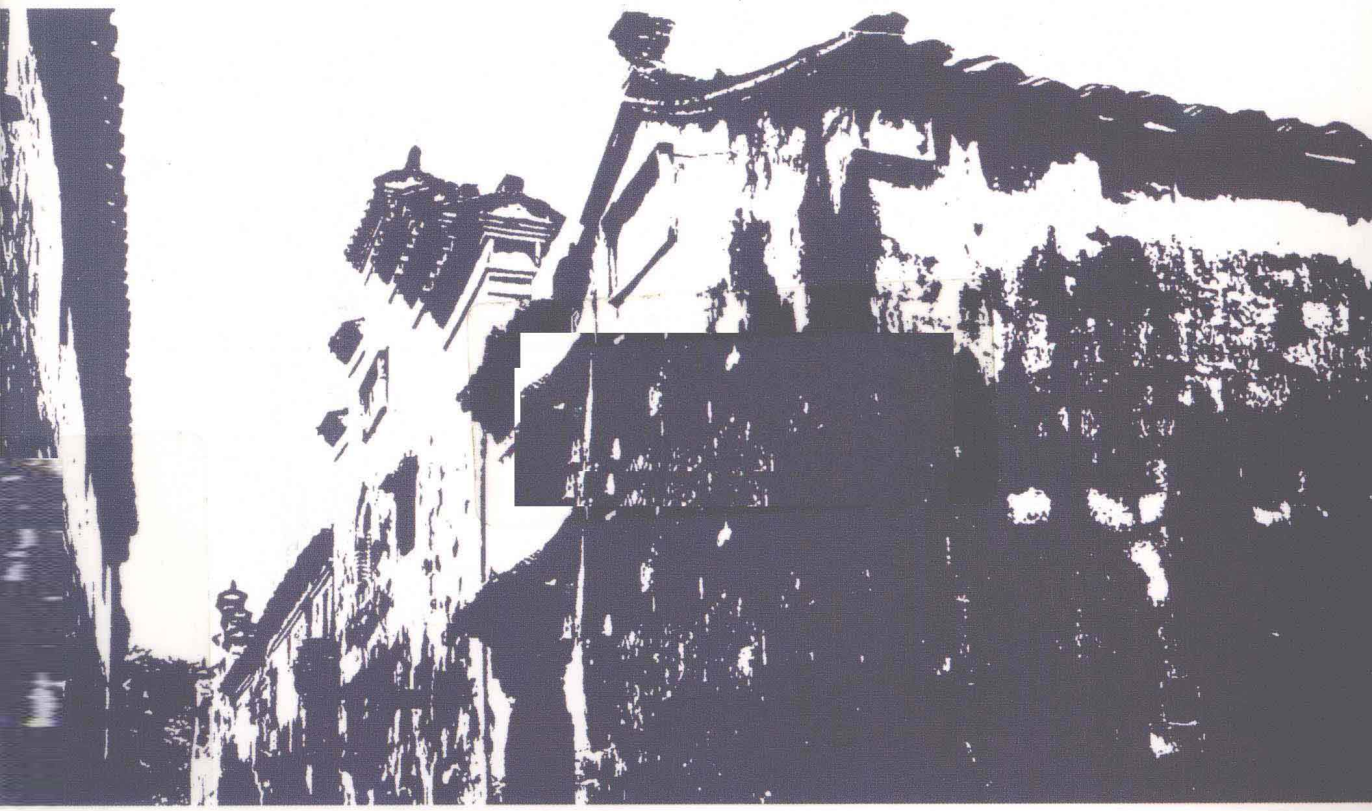
甘于恩　主编

南/方/语/言/学/丛/书

语言调查
语料记录与立档规范

范俊军　主编



暨南大學出版社
JINAN UNIVERSITY PRESS

暨南大学汉语方言研究中心
中国社会科学院民族学与人类学研究所

语言调查
语料记录与立档规范

主编　范俊军
审订　孙宏开　王　宁

编著　李云东　肖自辉　肖荣钦　甘辉云
　　张　帆　李义祝　王小娟

暨南大學出版社
JINAN UNIVERSITY PRESS
中国·广州

图书在版编目（CIP）数据

语言调查语料记录与立档规范/范俊军主编．—广州：暨南大学出版社，2011.1
（南方语言学丛书）
ISBN 978-7-81135-637-3

Ⅰ.①语… Ⅱ.①范… Ⅲ.①语言调查—规范—中国 Ⅳ.①H07-65

中国版本图书馆 CIP 数据核字(2010)第 176137 号

出版发行：暨南大学出版社

地　址：中国广州暨南大学
电　话：总编室（8620）85221601
营销部（8620）85225284　85228291　85228292（邮购）
传　真：（8620）85221583（办公室）　85223774（营销部）
邮　编：510630
网　址：http://www.jnupress.com　http://press.jnu.edu.cn

排　版：广州市天河星辰文化发展部照排中心
印　刷：河源市天才印务有限公司

开　本：787mm×1092mm　1/16
印　张：24.875
字　数：602 千
版　次：2011 年 1 月第 1 版
印　次：2011 年 1 月第 1 次
印　数：1—1000 册

定　价：55.00 元

总　序

2008 年 5 月，暨南大学汉语方言研究中心被批准为广东省教育厅人文社科重点研究基地，这既是莫大的荣誉，又是一种考验。中心全体研究人员皆深感责任重大，在各方有力的支援下，克服重重困难，竭尽全力开展科研工作，取得了一定的成绩。2009 年，《南方语言学》（第一辑）诞生了，她为省内外语言学者交流学术成果搭建了一个极佳的平台。该书出版不久，即告售罄。这说明《南方语言学》的出版不但非常及时，而且十分必要。

然而，《南方语言学》的容量毕竟有限，她要面向广大语言学者，不能只容纳、发表暨南大学汉语方言研究中心的学术成果。为了全面检视中心近年来的研究成果，经过充分酝酿，我们决定编纂“南方语言学丛书”。本次推出的是第一系列，以岭南方言研究为重点，包括以下七种：

甘于恩《广东四邑方言语法研究》

伍巍《方言研究集稿》

范俊军《语言调查语料记录与立档规范》

彭小川《广州话助词研究》

陈晓锦、翁泽文《粤语西翼考察——广西贵港粤语之个案研究》

刘新中《广东、海南闽语若干问题的比较研究》

钟奇《汉语方言的重音模式》

第一系列出版后还将征集相关研究成果，陆续推出第二系列、第三系列等。我们既重视基地骨干研究人员成果的整理与出版，也欢迎兼职研究人员提交成熟的研究成果与读者分享。我们衷心希望汉语方言研究中心真正成为团结、整合南方方言研究力量的一方园地，不断推动语言基础研究和应用研究的深入开展。

广东社会氛围宽松，语言丰富复杂，学术思想活跃，经济蓬勃发展，这些都是广东省语言学繁荣的重要条件。汉语方言研究中心现已成为广东省人文社科的重点研究基地，并获得了省里及各方的有力支持。如此良好的学术环境，请诸君勿负之！让我们加倍珍惜，共同努力吧，广东方言学乃至语言学事业一定可以结出更加丰硕的果实！

最后，欢迎读者朋友对“南方语言学丛书”的编纂、出版提出宝贵意见，使它成为南方语言学的一个闪亮品牌。

甘于恩

2010 年 7 月 8 日深夜草于方言调查途中

说　明

一、范围

本规范制订了语言调查记录的词汇集、句子集、话语主题集、话语转写规范、语料数据格式以及采录设备技术要求，拟订了普通语言描述主题条目、中国语言标准代码和语言资源立档元数据。本规范适用于中国语言调查、语料采录、语档建设、语言教学与研究、濒危语言资源保存、语言工程以及语言资源开发利用。

二、规范

本规范引用或参照了下列标准和文件：

（1）GB/T 4880.1－2005；GB/T 4880.2－2000　语种名称代码

（2）GB/T 16159－1996　汉语拼音正词法基本规则

（3）开放语档联盟的标准和文件

（4）民族语：语言条目，美国暑期语言研究院

（5）RT－04 转写方案，语言数据联盟，2004

（6）话语转写层次细分规则，美国语言学会，2008

三、编码

本规范的调查表条目以 7 位数字和字母组合编码。编码格式如下：

（1）词汇：词汇标识符（V）+义类序号（2 位数字）+下位义类序号（1 位字母）+词条序号（3 位数字）。例如：V01A001 天。

（2）语法例句：句子标识符（S）+语法项大类序号（2 位数字）+语法项次类序号（1 位字母）+例句序号（3 位数字）。例如：S01A001 花开了。

（3）日常用句：句子标识符（S）+ 话题序号（2 位数字）+ 日常用句标识符（N）+句子序号（3 位数字）。例如：S01N005 进屋里坐坐。

（4）话语主题：话语标识符（D）+话语体裁类型序号（2 位数字）+体裁次类序号（1 位字母）+话题序号（3 位数字）。例如：D01A001 招待客人就餐时的闲聊。

四、分级

词汇集条目按核心词（S）、最常用词（T）、常用词（X）、次常用词（Y）、非常用词（Z）分为 5 级，句子集按基本例句（T1）、扩充例句（T2）、自选例句（T3）分为 3 级。

五、立目

语言描述条目、词汇集、句子集和话语主题集等四种集表的条目以普通话立目，个别概念用方言立目。所有条目有英文对照，其中词汇集的部分生物条目列出拉丁学名。

六、软件

配套软件 FieldSound V3.0（田野之声，2009SR018518）具有制表、分条目录音播放、数据处理和语图分析等功能。下载网址：http://www.clarc.cn（中国语言有声资源联盟）。

INTRODUCTION

Ⅰ. Coverage

This book protocols lexicon, sentence list and discourse topics, discourse transcription convention, and the technical requirements of audio and video equipments for language data recording. It also specifies language standard codes, general language description entries, metadata set for language resource archiving. It is applicable to language fieldwork, language data archiving, language teaching, linguistic research, endangered language preservation, language engineering, and the development and utilization of language resources in China.

Ⅱ. Criteria

This book quotes or refers to the following standards and documents:

1. GB/T 4880.1 - 2005; GB/T 4880.2 - 2000
2. GB/T 16159 - 1996
3. OLAC standards and documents
4. Ethnologue: language entries, SIL
5. RT - 04 Transcription, LDC, 2004
6. Transcription Delicacy Hierarchy for Discourse Transcription, LSA, 2008

Ⅲ. Encoding

Each entry in all the sets is encoded into 7 letter-and-number codes. The form is as below:

1. Lexical set: V + semantic classification code (2 digits) + sub-classification code (1 letter) + serial number of entries (3 digits).

2. Grammatical sentence: S + grammatical classification (2 digits) + sub-classification (1 letter) + serial number of sentences (3 digits).

3. Everyday sentence: S + topic classification (2 digits) + N + sentences serial number (3 digits).

4. Discourse topic set: D + serial number of data type (2 digits) + sub-type (1 letter) + serial number of topic entries (3 digits).

Ⅳ. Classification

The lexicon entries are divided into five grades: Swadesh words (S), top common words (T), common words (X), sub-common words (Y), and uncommon words (Z). And the sentence entries are put into three grades: the basic, the extensional and the optional.

Ⅴ. Entry representing

Entries are generally represented in Putonghua, but a few entries in Chinese dialect words as necessary. All entries of the four sets are also translated into English, of which some creatures are given in Latin Scientific Name.

Ⅵ. Software

The software FieldSound_V3.0 (p/n: 2009SR018518) with the book is a useful tool for language fieldwork, such as table-creating, audio-recording, data-processing and spectrum-analyzing, etc. Download website: http://www.clarc.cn.

目　录

CONTENTS

壹　普通语言描述主题条目

Entries for general language description

LD01　语言名称（Language name）

LD01A　主要名称（Primary language name）

LD01A01　中文名称（Chinese name）：_____

LD01A02　拼音名称（Pinyin name）：_____

LD01A03　国际音标名称（IPA name）：_____

LD01A04　拉丁转写名称（Roman transcription）：_____

LD01B05　民族文字名称（Ethnic writing name）：_____

LD01B　语言代码（Language code）

LD01B01　国际标准代码（ISO 639－3）：_____

LD01C02　国标或其他代码（GB or other code）：_____

LD01C　语言别称（Alternative names of language）

LD01C01　中文名称（Chinese alternative name）：_____

LD01C02　拼音名称（*Pinyin alternative name*）：_____

LD01C03　别称注音（IPA alternative name）：_____

LD01C04　拉丁转写别称（Roman transcription）：_____

LD02A05　民族文字别称（Ethnic writing name）：_____

说明：

1. 语言名称用本族人对语言的称呼。如有几个名称，以通用名称为主要名称。别名包括：本族人对语言的其他称呼，外族对该语言的称呼，语言研究者给语言的命名。
2. 如本族没有专门的语言名称，而外族人的称呼在本族人听来带有贬义，则以族群名称或传统居住地命名，不得将贬称作为语言主要名称。外族的贬称可在别名中列出，后面注明“（贬）”。

LD02　人口及分布（Population and distribution）

LD02A　县内人口（Population in the county）

LD02A01　民族人口及分布（Population & distribution of ethnic groups）

LD02A02　各种语言使用人口及分布（Speakers & areas of each language）

LD02B　调查点总人口和语言使用人口（Total population and speakers of the fieldwork place）
LD02B01　调查点人口（乡/村/街）（Total population of the place）
LD02B02　语言使用人口（Speakers of the language）
LD02C　调查点双语状况（Bilingualism in the place）
LD02C01　单语人口（Monolingual speakers）
LD02C02　双语和多语人口（Bilingual and multilingual speakers）
LD02C03　双语和多语的性别分布（Bilingualism classified by sex）
LD02C04　双语和多语的年龄分布（Bilingualism classified by age）
LD02C05　语言使用人口增减趋势（Trend of speakers increasing or decreasing）
LD02D　语言跨国境状况（The language in other countries）
LD02D01　国名及具体分布（Country name and geographical distribution）

LD03　语言系属分区（Classification of linguistic affiliation）

LD03A　语系（Linguistic phylum or language stock）：_____
LD03B　语族（Linguistic family）：_____
LD03C　语支（Linguistic branch）：_____

LD04　方言状况（Dialects of the language）

LD04A　方言区片归属（Classification of dialect affiliation）：_____
LD04B　方言内部互懂程度（Dialect intelligibility）：_____
LD04C　语音词汇相近程度（Lexical similarity）：_____
说明：1. 语言系属分类以国内学者的划分为准。
2. 互懂度、词汇相近度用百分比表示。

LD05　语言使用范围（Domains of language use）

LD05A　语言地位（Language prestige）
LD05A01　法定民族语言（Ethnic language authorized by government）
LD05A02　县级官方语言（Official language within the county）
LD05A03　乡镇通用语言（Official language within the town）
LD05B　语言使用范围（Domains of language use）
LD05B01　家庭范围（Language use in the family）
LD05B02　社区范围（Language use in the community）
LD05B03　对外相邻社区（Language use while with other community nearby）

LD06　宗教信仰与民俗（Religious affiliation and folk rituals）

LD06A　宗教活动中的语言使用（Language in religious activies）

LD06A01　经书、读经、宗教歌曲（Religious books，chanting，chorus）
LD06A02　教徒活动（Religious community member's activities）
LD06B　民俗活动中的语言使用（Language in folk activities）
LD06B01　法师、道士等神职人员的吟唱、诵读或咒语（Incantation，sacred speech，Taoists's words）
LD06B02　仪式活动参与人的语言（Participant's speech）

LD07　语言发展（Language development）

LD07A　语言教育（Language education）
LD07A01　幼儿园的教学语言和日常语言（Language use in kindergarten）
LD07A02　小学的教学语言、语言课程和日常语言（Language use in primary school）
LD07A03　初中的教学语言、语言课程和日常语言（Language use in middle school）
LD07A04　语言培训学校（Sparetime language-training school）
LD07B　语言产品和传媒语言（Language products and language use in new media）
LD07B01　纸本语言产品（教材/工具书/报刊/书籍）（Printed or written language publications）
LD07B02　其他语言产品（磁带/光盘/器件/软件/网站）（Language products as tapes，CD，MP3，software and web）
LD07B03　影视文艺作品（电影/电视剧/歌曲/娱乐产品）（Visual products as film，movies，TV series，songs）
LD07B04　电台节目频道（新闻/娱乐/语言学习/其他频道）（TV and broadcast programs in the language）
LD07B05　公共信息广播和通讯服务（车站/公共交通/医院/电话/其他）（Public information broadcast and communicational services）

LD08　文字和扫盲（Wrting system and literacy）

LD08A　语言的文字系统（Wrting system of the language）
LD08A01　语言的传统文字（有/无；名称；表音/表意；官方地位）（Traditional writing system）
LD08A02　语言的新创文字（有/无；名称；表音/表意；官方地位）（New official writing system）
LD08B　文字通行面和识字状况（Domains of the writing system and literacy）
LD08B01　传统或新文字的使用范围（官方文书/公共标牌/报刊/其他）（Domains of traditional or official writing system）
LD08B02　识字群体比例（公务员/教师学生/商业人士/农民/其他）（Classification of the literacy members）

LD08C 语言社群的汉语扫盲状况（Chinese abilities and literacy）
LD08C01 老中青扫盲比例（普通话听说/汉字认读）（Chinese abilities and literacy by ages）
LD08C02 男女扫盲比例（普通话听说/汉字认读）（Chinese abilities and literacy by sex）

LD09 社群成员的语言态度（Language attitudes of the community members）

LD09A 态度积极的人口类别（Positive speakers by age，sex，other factors）
LD09A01 态度积极的老中青比例（Positive speakers by age）
LD09A02 态度积极的男女比例（Positive speakers by sex）
LD09A03 态度积极的职业比例（Positive speakers by occupation）
LD09B 持中间态度的人口类别（Neutral speakers by age，sex，other factors）
LD09B01 持中间态度的老中青比例（Neutral speakers by age）
LD09B02 持中间态度的男女比例（Neutral speakers by sex）
LD09B03 持中间态度的职业比例（Neutral speakers by occupation）
LD09C 持消极态度的人口类别（Negative speakers by age，sex，other，factors）
LD09C01 态度消极的老中青比例（Negative speakers by age）
LD09C02 态度消极的男女比例（Negative speakers by sex）
LD09C03 态度消极的职业比例（Negative speakers by occupation）

LD10 语言活力（Language vitality）

LD10A01 充满活力，安全（safe，of great vitality）
LD10A02 活力脆弱，不安全（unsafe，vulnerable）
LD10A03 活力衰退，确有危险（of declined vitality，definitely endangered）
LD10A04 活力很弱，严重濒危（of poor vitality，severely endangered）
LD10A05 活力极弱，极度濒危（of poorest vitality，critically endangered）
LD10A06 无活力，灭绝或几近灭绝（no vitality，extinct）

LG11 语言社群地域分布（Geographical distribution of language community's residency）

LD11A 与中心城镇距离及交通状况（Distance away from central towns and vehicles）
LD11A01 与中心城镇距离（县城/中心集镇，交通工具）（Distance away from central towns）
LD11A02 地理开放程度（山区/丘陵/平原/沿江；闭塞/开放；聚居/散居/杂居）（Geographical openness of the area）
LD11B 语言地图（标记到自然村）（Language map）
LD11B01 语言地图标记符号（Labels in map）

贰　语料描述与立档元数据术语

Metadata terms of language data description & archiving

说　明

1. **本元数据集参照或引用的标准和文件**

（1）DCMI Metadata Terms，2008
（2）OLAC Role Vocabulary，2006
（3）OLAC Linguistic Data Type Vocabulary，2006
（4）OLAC Metadata，Version 1.1，2008
（5）ISLE Metadata Initiative，version 3.04，2003

2. **本集包括以下标准和文件**

（1）都柏林核心元数据
（2）contributor 元素扩展：参与者角色词汇
（3）type 元素扩展：语料类型词汇
（4）medium/extent 限定元素：语料格式词汇

3. **用途**

本元数据术语集适用于语言调查记录、语料汇集、处理、分类以及语言资源数字化立档。

表 1　都柏林核心元数据

元素代码	中文名称	说明及举例
title	题名	语料资源的名称。题名中可以包括语言名称词。例如：广东畲语情景话语。
creator	创建者	创建资源的主要责任人。如：调查项目主持人、发音合作人、作者。
contributor *	贡献者	对语料资源作出贡献的其他个人或机构。详细的责任分工，参照下文的《contributor 元素扩展：参与者角色词汇》。
date	日期	创建或获得语料资源的时间。如：语言调查日期、录音或摄像日期。
coverage	涵盖范围	语料资源内容覆盖的时间和空间范围。例如：语料资源的发生地、内容发生的时段。在数字化立档中，应使用地理信息系统的标准地名和时域名。
description *	描述	对语料资源实体的简介或概述，可以采用摘要、表格、图示或其他形式呈现的内容，呈现方式不限。
format *	格式	语料资源数字化格式和介质载体，包括文件格式、介质形态、容量、技术参数，等等。具体内容可参照下文的表 4。
identifier	标识码	标识语料资源的字符串或数字代码。最好使用统一资源标识符（URI）。
language *	语言	描述语料资源的语言。例如：用汉语描述少数民族语言语料资源，这里的语言即汉语。请参照第叁部分的《中国语言标准代码》。
publisher	发布者	使语料资源成为可利用和可获取形态的个人或机构。如：出版社、杂志社、图书馆、档案馆、网络服务商。
relation	关联	语料资源与其他资源的关系或关联。可以使用统一资源标识码与相关内容链接。
rights	权限	语料资源的所有权或使用权限。如：著作权、出版权、使用权。
source	来源	语料资源的出处。如：来自某个机构或个人、某本著作、某个网站。
subject *	主题	语料资源的内容主题。用关键词或短语、简单句子表述，建议使用专业主题词或控制词。
type *	类型	语料资源的特征和类型。如：声音、文献、图像，数字化格式、体裁。可使用表 3 的类型词汇。
provenance	出处	贯穿语料资源始终的有关其所属权和监管权任何变化的陈述，关系到该资源的权威性、完整性及可解释性。
rightsHolder	权属者	拥有或管理语料资源的个人或机构。

注：加“*”的元素有限定元素或扩展词汇。

表 2　contributor（贡献者）元素扩展：参与者角色

元素代码	中文名称	说明及举例
annotator	标注者	为语料做标注的个人或机构。
author	作者	参与原始语料资源采集和记录的人员。
compiler	汇编者	将语料各部分汇编成集的个人或机构。如：制作语料库，将单首歌汇成 CD，把有关软件收集打包。
consultant	咨询顾问	为语言描写和分析提供参考意见者，通常是该语言族群的文化人士，但不是发音人。
data inputter	录入员	语料数据打字录入或排版的个人或机构。
depositor	存档者	收藏语料资源的个人或机构。
developer	技术开发员	为语料采集、创建或处理开发技术工具的个人或机构。
editor	编校员	对语料编辑加工、校对、核查的个人或机构。
illustrator	绘图员	语料中插图和插表的绘制者。
interpreter	口译员	话语记录中的现场翻译或解说人。
interviewer	采访人	实地采访并获得一手访谈语料的采访人员。
participant	参与人	语料产生或创建时除话语主角以外的在场人员。如：观众、活动组织者、突然插话者。
performer	演出者	仪式或演出中除主持人和主角以外的演出人员。如：仪式或表演活动的附和者、配角、配乐者。
photographer	摄影者	照片和视频材料等语料资源的摄制人员。
recorder	录音者	操作录音设备现场录制声音语料资源的人员。
researcher	研究者	产生语料的某项研究成果的研究人员。
research participant	项目参与者	参与研究项目的个人或机构，项目内容是语料的基本组成部分。
responder	应答者	语料话语事件的次要参与人，特指会话中只有“嗯”、“是的”等简单应答的次要说话人。
signer	主角	录音、视频或文献语料中的话语主角。
singer	歌唱者	音频和视频语料资源中的歌唱者，包括独唱人、领唱人和合唱人。
speaker	发音人	音像语料或文本记录材料的主要发音合作人。
sponsor	资助人	为创建语料资源提供财物支持的个人或机构。
transcriber	抄录员	语料或有关材料的抄录员、誊写员。
translator	笔译员	语料或相关材料的文字翻译人员。

表 3　type（类型）元素扩展：语料类型

受控词	中文名称	说明及举例
collection	合集	语料资源的汇集，它的各个组成部分可单独分开描述。如：“瑶语词汇语料集”，有声音、文字等语料。
sound	声音	可听感的语料资源。
image	视图	除文字以外的视觉语料资源。如：电影、动画、照片和图画。
movingImage	动态视图	一系列连续移动的视频语料。
stillImage	静态视图	静止的可视化语料资源。
text	文本	以文字呈现的可读语料资源。
primaryText	原始文本	未经加工修饰的第一手文字语料。如：民间藏稿、古籍、碑刻、铜刻、竹刻、皮刻等，田野调查现场笔记。
descriptionText	描写文本	对语言或其中某方面做系统结构分析或研究的文本资源。
phonemeLexicon	音素词表	以正式文字或其他书面符号形式展现，反映语言音素系统的词表。
wordList	词汇表	以正式文字或其他书面符号形式展现的语言词汇表。
sentenceList	句表	以正式文字或其他书面符号形式展现的语言句子表。
discourseType	话语类型	以正式文字或其他书面符号形式展现的话语文本。
drawing	草图	田野调查中现场绘制的有助于理解语料内容的图示。
dataset	数据集表	以数字化形式展现的表格文件，如：excel 表和 access 表单。
software	软件	语料采集、播放或展示的软件工具和技术平台。

表 4　　medium/extent 限定元素：语料数据介质

限定元素	受控词	中文名称	说　明
medium		介　质	语料资源的物质材料或物理载体
	audioTape	录音带	
	videoTape	录像带	
	CD	激光唱盘	
	DVD	激光视盘	
	imageFilm	胶片	
	epigraphEntity	铭文载体	
	bookPaper	纸本	
extent		范　围	可使用 MIME 协议受控词
	audio/wav	wav 音频文件	
	audio/aiff	aiff 音频文件	
	audio/au	au 音频文件	
	audio/ra	ra 流媒体文件	
	audio/mid	mid 电子音乐文件	
	audio/mp3	mp3 音频文件	
	video/avi	avi 视频文件	
	video/mpeg1	mpeg1 视频文件	
	video/mpeg2	mpeg2 视频文件	
	video/mpeg4	mpeg4 视频文件	
	video/mov	mov 视频文件	
	video/asf	asf 视频文件	
	video/wmv	wmv 音视频文件	
	image/jpeg	jpeg 图片文件	
	image/bmp	bmp 位图文件	
	image/tiff	tiff 图形文件	
	text/doc	doc 文档文件	
	text/pdf	pdf 文档文件	
	text/txt	txt 纯文本文件	
	text/html	html 网页文本文件	
	Text/xml	xml 文本文件	
	sampling	采样率	
	size	大小/容量	
	duration	时长	