SI TANG MEUNG VAX
KAWN NYAWM PA HU LAIG DAUX KRAX PA YUH DAUX HAK TIEX HAMX YUING
NANG SEUNG PHU EUX 48 KAUX CAW DEEH ING DAUX VENG PANG KHAM

语言记录 中的

语料库建立

Corpus construction

in Language Documentation

蘇沙

马克思普朗克 心理语言学研究所

Hilário de Sousa

Max Planck Institute for Psycholinguistics

hilario@bambooradical.com *hilario.desousa@mpi.nl*

ELDP 语言记录培训 *ELDP School on Language Documentation* 玉溪师范学院 *Yuxi Normal University*

语言描述 *Language Description*	≠	语言记录 *Language Documentation*
语料库语言学 的 语料库 *Corpus* *for Corpus Linguistics*	≠	语言记录 的 语料库 *Corpus* *for Language Documentation*

语言描述
Language Description

≠

语言记录
Language Documentation

透过语言资料描述语言结构

Describing language structure through language data

- 收集语言资料並非主要目的
- *Collecting language data is not the primary goal*
- 收集语言资料种类不一定要很广
- *Range of language data do not necessarily need to be very wide*

语言记录的焦点是 可以观察的语言行为与知识。语言记录的目的是对于［在某一个时间某一个话语社团观察到的的语言行为、和话者对这些语言行为的知识］的持续、多功能纪录。

“Language documentations [...] focus on observable linguistic behavior and knowledge. The goal is a lasting, multifunctional record of the linguistic practices attested at a given time in a given speech community and the knowledge speakers have about these practices.” (Himmelmann 2008: 346)

语言描述 *Language Description* ≠ 语言记录 *Language Documentation*

透过语言资料描述语言结构

Describing language structure through language data

- 收集语言资料并非主要目的
- *Collecting language data is not the primary goal*
- 收集语言资料种类不一定要很广
- *Range of language data do not necessarily need to be very wide*

全面性纪录一个话语社团的语言行为（与话者对之的认识）：

Comprensive record of the linguistic practices of a speech community:

- 词汇 *lexicon*
- 语用方程式 *speech formula*
- 对话 *conversation*
- 独白 *monologue*
- 语言艺术／语言游戏 *linguistic art/ language game*
- 仪式语言 *ceremonial language*

等等 etc.

语言描述 *Language Description* ≠ 语言记录 *Language Documentation*

透过语言资料描述语言结构

Describing language structure through language data

- 收集语言资料并非主要目的
- *Collecting language data is not the primary goal*
- 收集语言资料种类不一定要很广
- *Range of language data do not necessarily need to be very wide*

多功能纪录 *multifunctional record*

对象 *target*:

- 提供资料的语言社团 *speech community that provided the data*
- (不同种类的语言学家和其他学者 *various types of linguists and other scholars*)

等等 etc.

语言描述 *Language Description* ⇆ 语言记录 *Language Documentation*

提供资料的语言社团 *speech community that provided the data*

- 语言文化保育／复兴 *maintenance/ revival of language and culture*

语言学及其他相关学科 *linguistics and related disciplines*

- 语音学／音韵学 *phonetics/phonology*
- 词法学／句法学 *morphology/syntax*
- 语义学 *semantics*
- 语用学 *pragmatics*
- 社会语言学 *sociolinguistics*
- 心理语言学 *psycholinguistics*
- 认知语言学 *cognitive linguistics*
- 计算语言学 *computational linguistics*
- 语料库语言学 *corpus linguistics*
- 语言教学 *language teaching*
- 语言政策 *language policy*
- 文体学 *stylistics*
- 人类学 *anthropology*
  - 语言人类学 *linguistic anthropology*
  - 民族音乐学 *ethnomusicology*

等等 etc.

语言描述 *Language Description* ≠ 语言记录 *Language Documentation*

透过语言资料描述语言结构

Describing language structure through language data

- 收集语言资料<u>并非</u>主要目的
- *Collecting language data <u>is not</u> the primary goal*
- 收集语言资料种类<u>不一定</u>要很广
- *Range of language data <u>do not necessarily</u> need to be very wide*

持续纪录 *lasting record*

- 档案格式／媒体 *file format/ medium*
- 档案命名 *file naming*
- 原数据 *metadata*

语料库语言学 的

语料库

Corpus

for Corpus Linguistics

≠

语言记录 的

语料库

Corpus

for Language Documentation

代表性 & 平衡性

Representativeness & Balance

对象 *Target*:

- 学者 *scholars*

语料库建立 *Corpus construction*:

- （资料来源的版权问题 *copyright issues of source data*）
- 学者的需求 *scholars' needs*

对象 *Target*:

- 学者 *scholars*
- 话语社团 *speech community*

语料库建立 *Corpus construction*:

- （资料来源的版权/隐私问题 *copyright/privacy issues of source data*）
- （学者的需求 *scholars' needs*）
- 话语社团的需求 *speech community's needs*

语料库语言学 的
语料库
Corpus
for Corpus Linguistics

⇆

语言记录 的
语料库
Corpus
for Language Documentation

语料库语言学的代表性与平衡性

Representativeness & Balance in
Corpus Linguistics

- 数量上的严格控制(统计学原因) *tight quantity control (statistical reasons)*
- 资料有数量上的上限 *upper limits in quantity of data*
- 标写必须划一(&计算机可读) *transcriptions must be consistent (and machine-readable)*

语言纪录的代表性与平衡性

Representativeness & Balance in
Language Documentation

- 平衡各人的需求 *Balancing the needs of various people*
  - 话语社团各人 *various people of the speech community*
  - (不同学者 different scholars)
- 平衡性可以因应代表性稍为调节 *Balance can be slightly adjusted due to representativeness*
- 资料无上限 *no upper limits of data*
- 标写的划一... *consistency of transcriptions...*

语料来源

Source of data

- 田野调查 *fieldwork*
    - 一手资料 *first hand data*
    - 二手资料 *second hand data*
        - 准不准你用 *permission for usage*
        - 可不可靠 *is it reliable*
        - 「适不适合」 *is it ‘suitable’*
- 公开的文字／录音／录像（注意版权问题） *public text/audio/video (beware of copyright issues)*
    - 已出版的文章 *published articles*
    - 网上发言／对话（注意隐私问题） *internet speech/conversation (beware of privacy issues)*
    - 电台／电视台／网上广播 *radio/television/internet broadcast*
        - 准不准你用 *permission for usage*
        - 可不可靠 *is it reliable*
        - 「适不适合」 *is it ‘suitable’*

语料种类 - 语言行为与知识

Type of data – Linguistic Behaviour and knowledge

沟通活动 *communicative events*

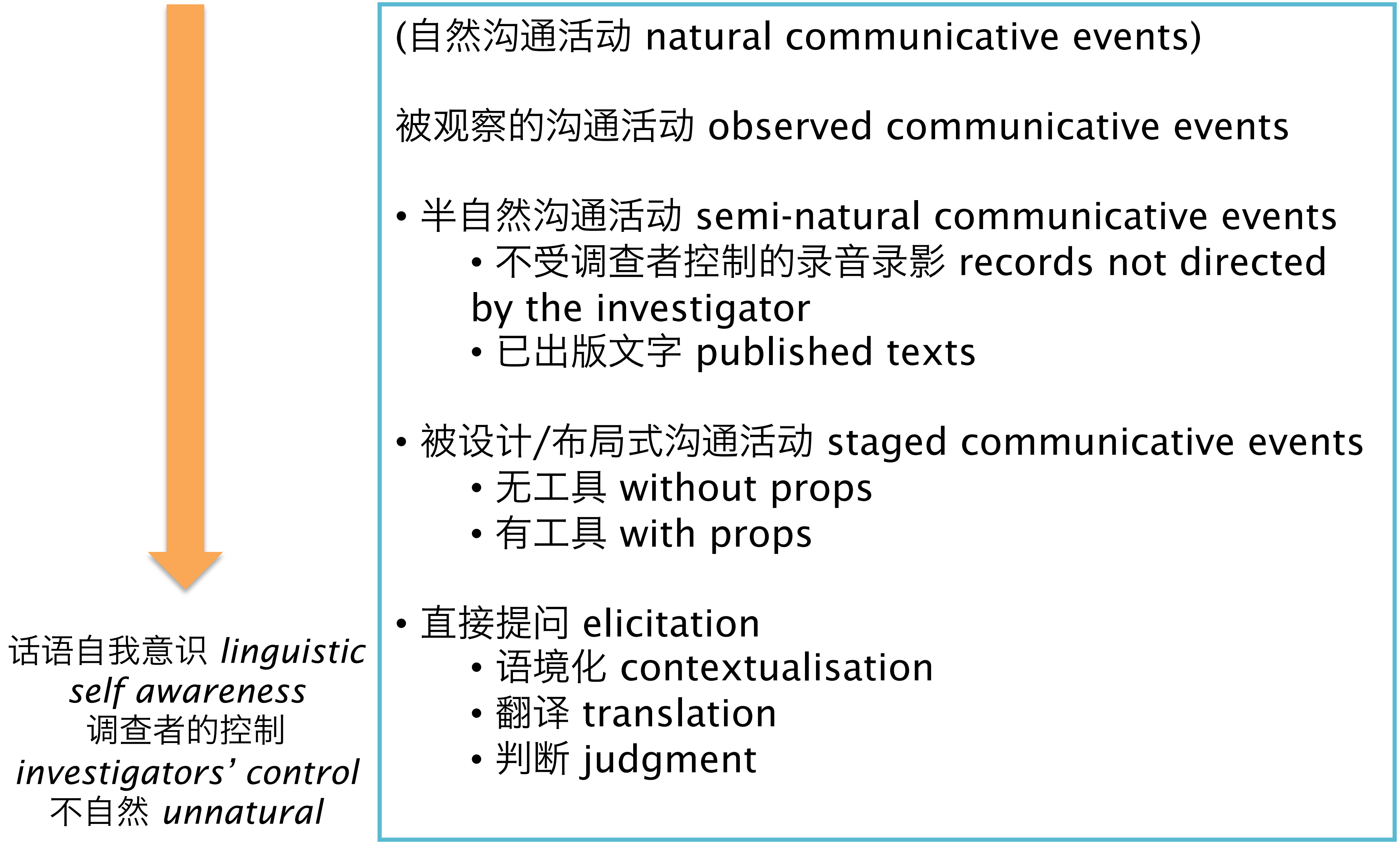
列表式资料 *lists*

- 词形系列 *paradigms*
- 民俗分类 *folk taxonomy*

分析资料 *analytic matters*

要有语言学分析／翻译／注释 *with linguistic analysis/translation/ commentary*

语料种类 - 沟通活动的自然程度
Type of data – naturalness of communicative events



(自然沟通活动 natural communicative events)

被观察的沟通活动 observed communicative events

- 半自然沟通活动 semi-natural communicative events
  - 不受调查者控制的录音录影 records not directed by the investigator
  - 已出版文字 published texts

- 被设计/布局式沟通活动 staged communicative events
  - 无工具 without props
  - 有工具 with props

- 直接提问 elicitation
  - 语境化 contextualisation
  - 翻译 translation
  - 判断 judgment

话语自我意识 *linguistic self awareness*
调查者的控制 *investigators' control*
不自然 *unnatural*

语料种类 - 计划程度
Type of data – level of planning

	大类 *major types*	例 *examples*
无计划 *unplanned*	感叹 *exclamative*	「哎呀！」 *‘Aiya!’*
		「救命呀！」 *‘Help!’*
	指令 *directive*	「叉烧包！」 *‘Porkchop bun!’*
		问候 *greetings*
		寒暄 *small talks*
	对话 *conversational*	聊天 *chat*
		讨论 *discussion*
		访问 *interview*
	独白 *monological*	叙述 *narrative*
		描述 *description*
		演说 *speech*
		官方通报 *formal address*
有计划 *planned*	仪式 *ritual*	连祷 *litany*

准许

Consent

合作人要清楚明白录音／录影可能的用途；合作人有资料发放的权控制。

The consultant needs to clearly understand the potential usage of data; the consultant has rights over the distribution of data.

同意书（签名）／ 口头同意的录音录像

consent form (signiture)/ audio-video recording oral consent

典藏机构的同意书样本

example consent form from an archiving agency

- 资料索取条件 *condition for data access*
- 可能的发布形式 *possible method of data dissemination*
- 学术／教学用途 *academic/ educational use*

等等 *etc.*

隐私／版权
Privacy/Copyright

语料库往后会公开／出版 *corpus will become publically accessible/ published*

- 准许你录音／录影 ≠ 准许你公开／出版 *permission for audio/video recording ≠ permission for publication*
- 内容的控制权在话语社团的手里 *content of corpus is controlled by the speech community*
- 会不会侵犯到别人的隐私／令人受损 *will it breach people's privacy/ cause damage*
- 出版物的版权 *copyright of publications*
    - 批准使用要有明确的证明 *clear evidence of permission of usage*

资料及原数据的处理 - 档案格式
Treatment of Data and metadata – file format

典藏格式，显示格式，工作格式

archival format, presentation format, working format

典藏格式 *archival format*:

- 非专有 *non-proprietary*
  - 公共领域 *public domain*
  - 没有版权／专利问题 *no copyright/patent problem*
  - 兼容度高 *high compatibility*
  - 将来转成别的格式容易 *easy to export to other formats in the future*
- 可携，可再用，多用途 *portable, reusable, repurposeable*
- 原本／品质最好的副本 *original, or best possible reproduction*

资料及原数据的处理 - 档案格式
Treatment of Data and metadata – file format

典藏格式，显示格式，工作格式

archival format, presentation format, working format

	文字档 *text files*	录音 *audio recordings*	图片 *images*	录影 *audio recordings*
典藏格式 *archival format*	*XML, PDF/A, plain-text UTF-8*	*.wav (48Khz/16 bits)*		
显示格式 *presentation format*	*pdf, html*			
工作格式 *working format*	文字处理器／排版系统的格式 *format of word processor/type setting program*			*mp4*

资料及原数据的处理
Treatment of Data and metadata

文字资料及元数据的格式 *Representation of data and metadata*	*XML-based (e.g. IMDI)* *plain-text (e.g. toolbox)*
影音资料的处理 *Treatment of audiovisual materials*	与时同步标注 *time-aligned transcription of audiovisual material (e.g. ELAN, Transcriber)*
语言材料发放 *Distribution of language materials*	语言典藏 *language archives (e.g. ELAR, PARADISEC)* 独立网页 *individual websites,* 纸 / 光盘 等 *paper/DVD etc.*
注解工具 *Annotation tools*	*(Corpus linguistics has semi-automated annotation tools for tagging, lemmatisation etc. Generally not available for language documentation.)*

档案的命名，与原数据
Naming of files, and metadata

格式要 **一致!!!**

format must be ***CONSISTENT!!!***

最好按照一个典藏组织的规定，

It is best to follow the requirements of an archival organisation

e.g. ELAR 的档案命名格式:

www.soas.ac.uk/elar/preparing-materials/file-names-and-folders/

uruan001.mp4
uruan001.wav
uruan001.eaf
uruan002.wav
uruan003-1.jpg
uruan003-2.jpg
uruan003-3.jpg

原数据
Metadata

最好按照一个典藏组织的规定，
It is best to follow the requirements of an archival organisation
e.g. ELAR 用 CMDI maker 或者 Arbil

最少有： *At least:*

- 录音／录影人 *creator*
- 录音／录影对象 *participants*
- 语言名称 *name of language*
    - 尽量具体！ *As specific as possible!*
        - ❌ 汉语 *Chinese*
        - ✅ (汉语)普通话 *Standard Mandarin*
        - ✅ 南宁普通话 *Nanning Mandarin*
        - ✅ 河口粤语 *Hekou Yue/ Hekou Cantonese*
- 年月日时，地点 *date/time, place*
- 资料索取条件 *condition for data access*
    - e.g. 要有密码 *password protected*
- 资料种类 *genre*
    - e.g. 叙述 *narrative*， 词表 word list

文献
Bibliography

Cox, Christopher. Corpus linguistics and language documentation: challenges for collaboration. In Newman, John & Baayen, R. Harald & Rice, Sally (eds.). *Corpus-based studies in language use, language learning, and language documentation*, 239–264. Amsterdam: Rodopi.

Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36(1): 161–195.

Johnson, Heidi. 2004. Language documentation and archiving, or how to build a better corpus. *Language documentation and description* 2: 140–153.

备份!!! 备份!!! 备份!!!

back up!!!