

Language Documentation

Sascha Griffiths

Canterbury, 2007

University of Kent
*Language Endangerment and
Documentation Seminar*

Convenor:

Bruce Connell

Spring Term 07

Overview

- What's That?
- Documenting a Language
 - Gathering Data
 - Archiving Data
 - Presentation of Data
- WELD
- References & Further Material (Info)

What's That?

What's That?

- Any ideas?

What's That?

- Any ideas?
- It is *NOT*:
 - Sitting *somewhere* with indigenous people writing a grammar (descriptive linguistics)
 - The construction of corpora, as such
 - A lesser task for linguists

What's That?

- Any ideas?
- It is *NOT*:
 - Sitting *somewhere* with indigenous people writing a grammar (descriptive linguistics)
 - The construction of corpora, as such
 - A lesser task for linguists
- It is a **priority task** for linguists regarding endangered languages

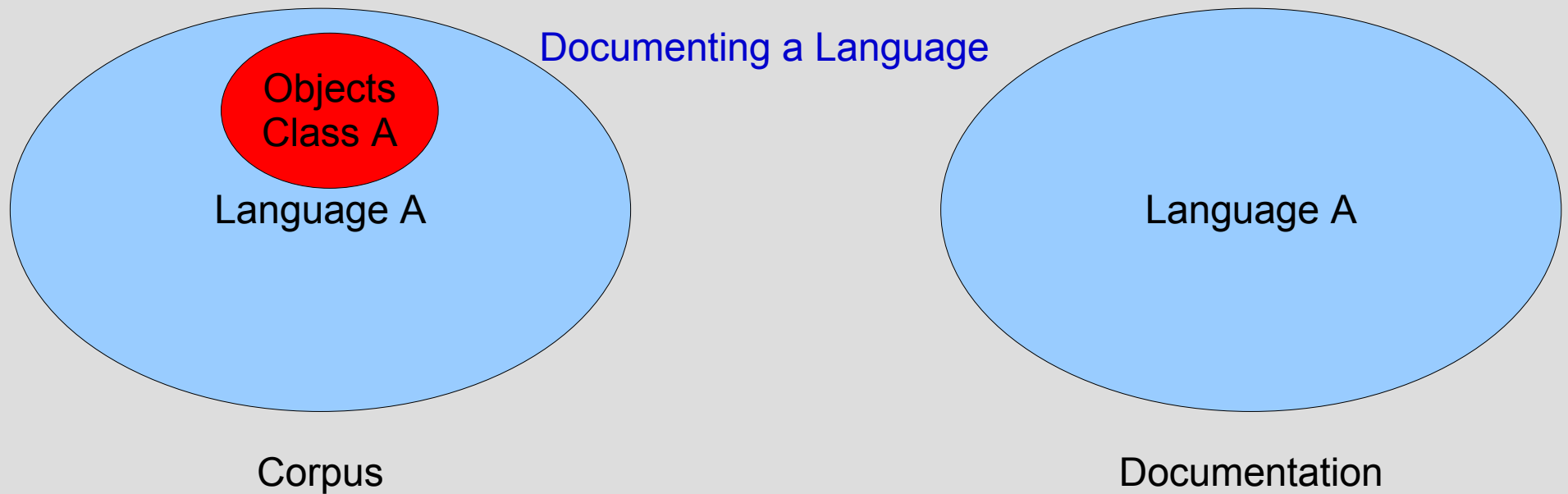
What *is* it?

Language documentation is...

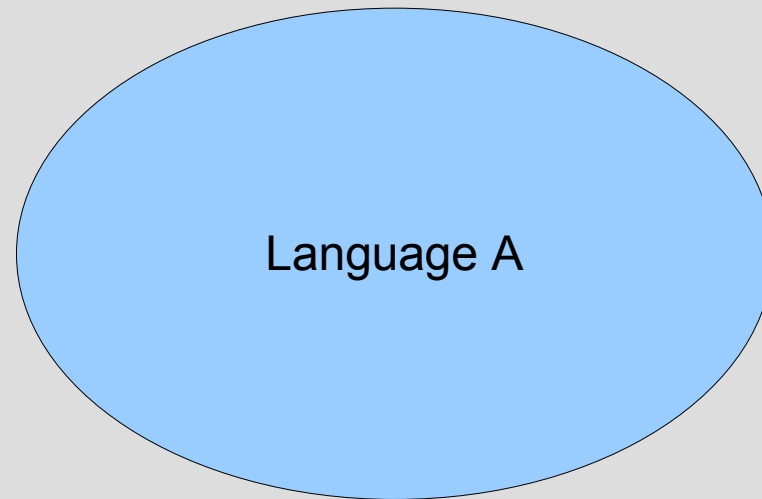
1. part of the overall documentation of a culture (Lehmann, 2001)
2. part of the comprehensive presentation of a language (Lehmann, 2002)
3. an activity (Lehmann, 2001) that serves as proof or is meant to teach about a human skill
4. an independent field (Himmelman, 1998) of research for linguistics and linguistic anthropology which focuses on the collection of primary data (language data), with special emphasis on endangered (minority) languages

The Difference to Corpus-linguistics

Is a set-theoretical one:



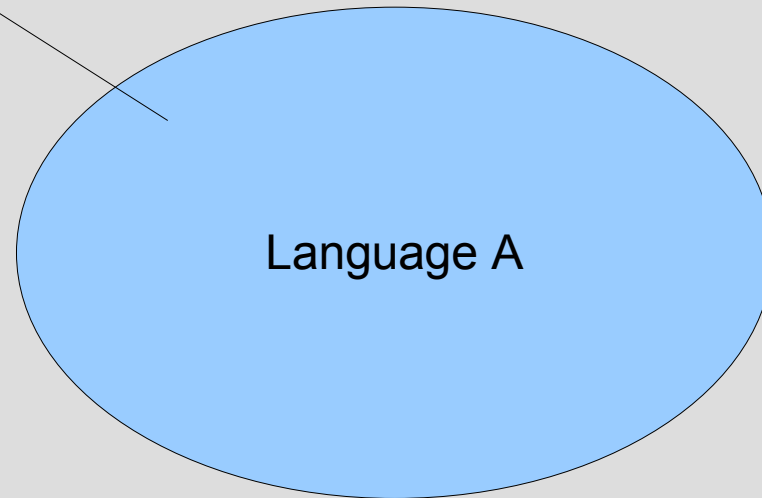
Documentation of a Semiotic System



Documentation

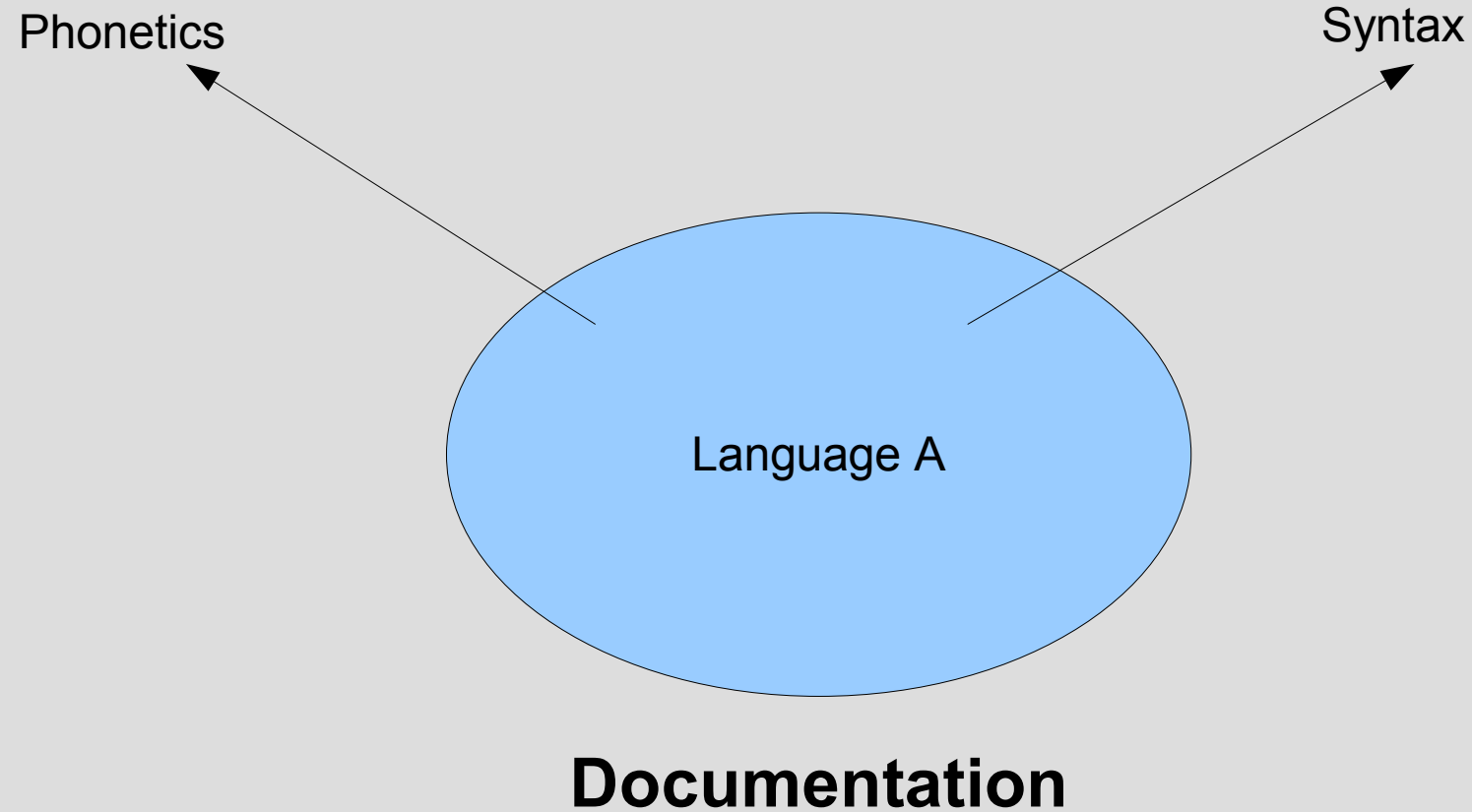
Documentation of a Semiotic System

Phonetics

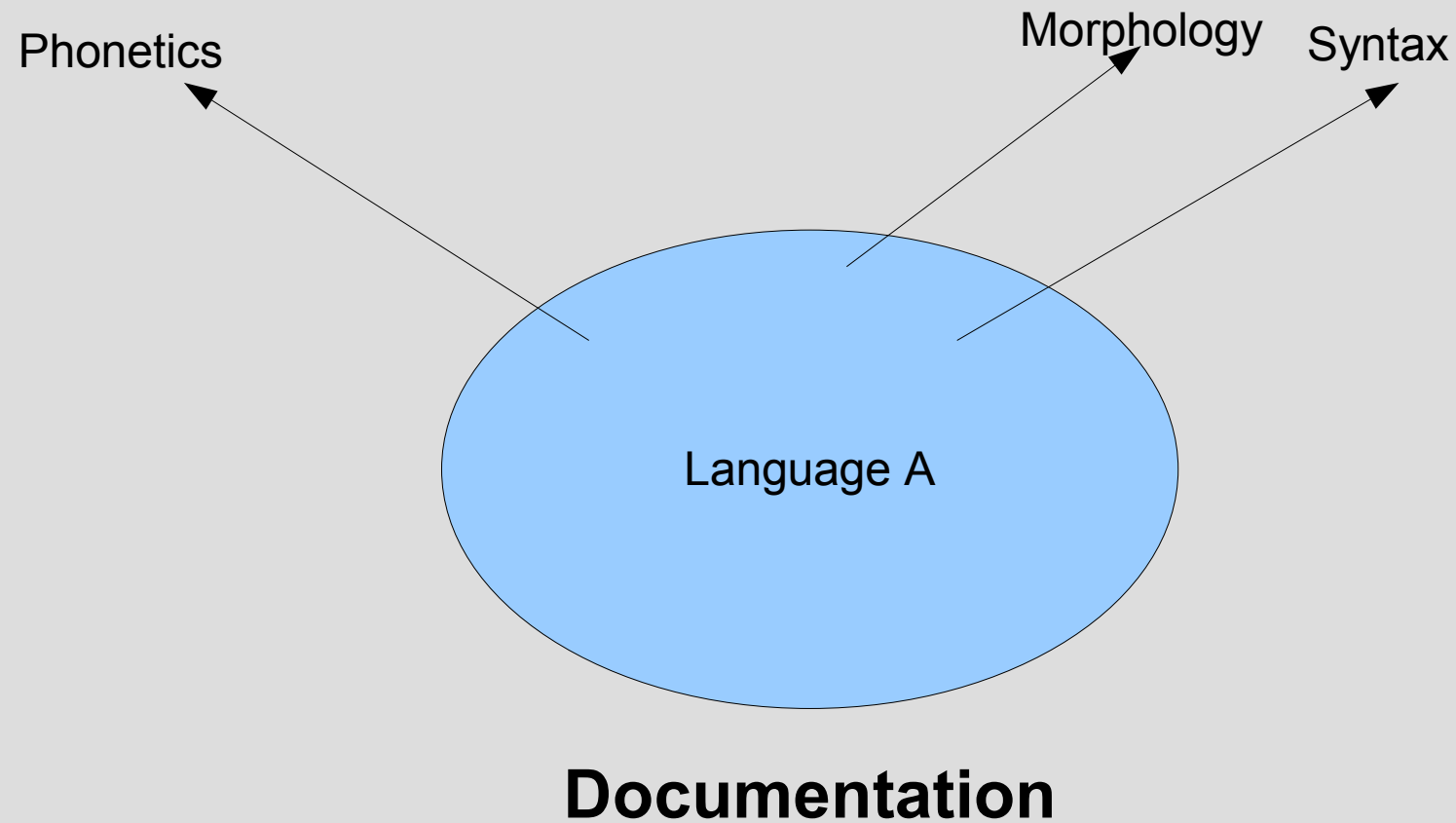


Documentation

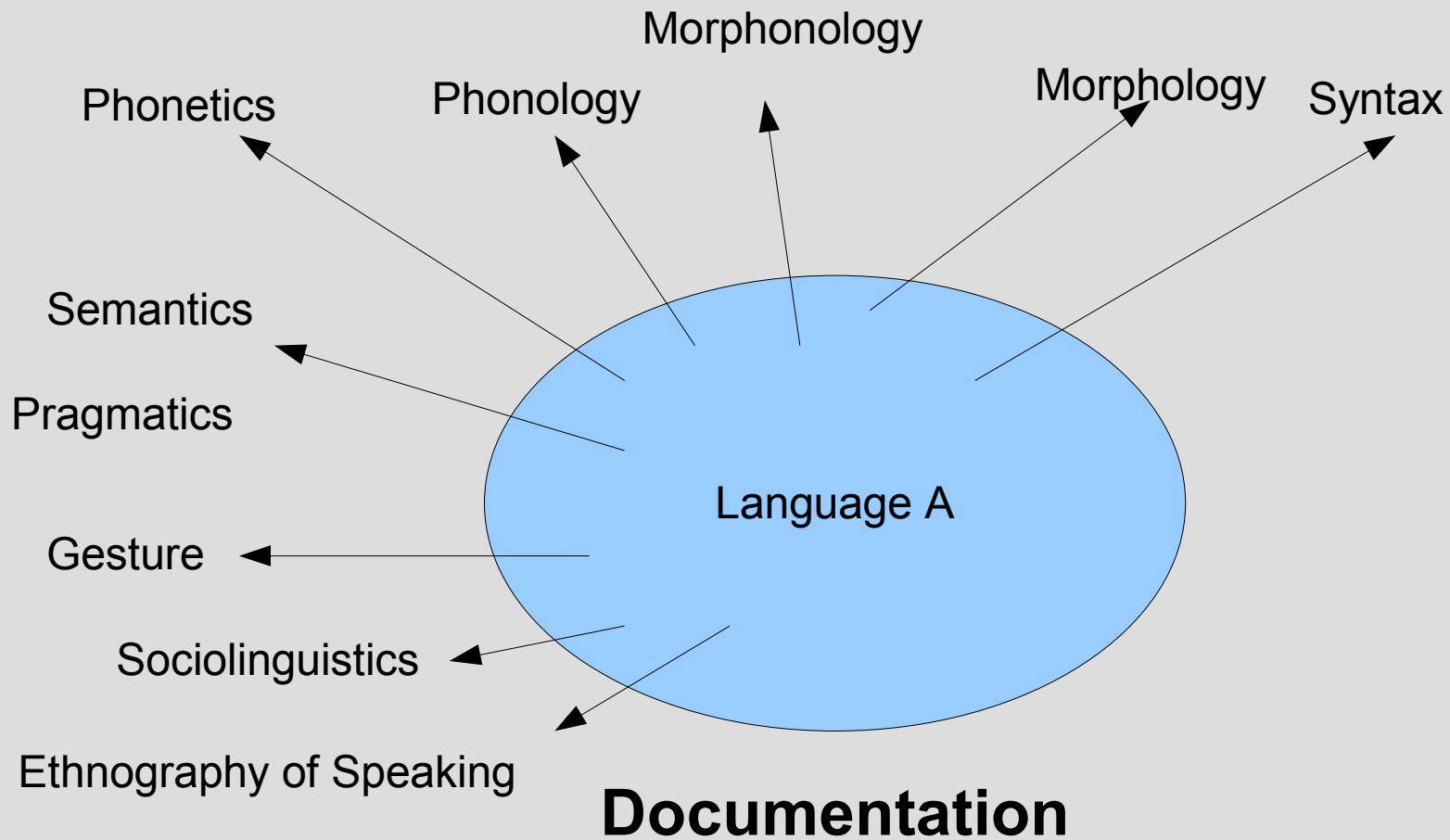
Documentation of a Semiotic System



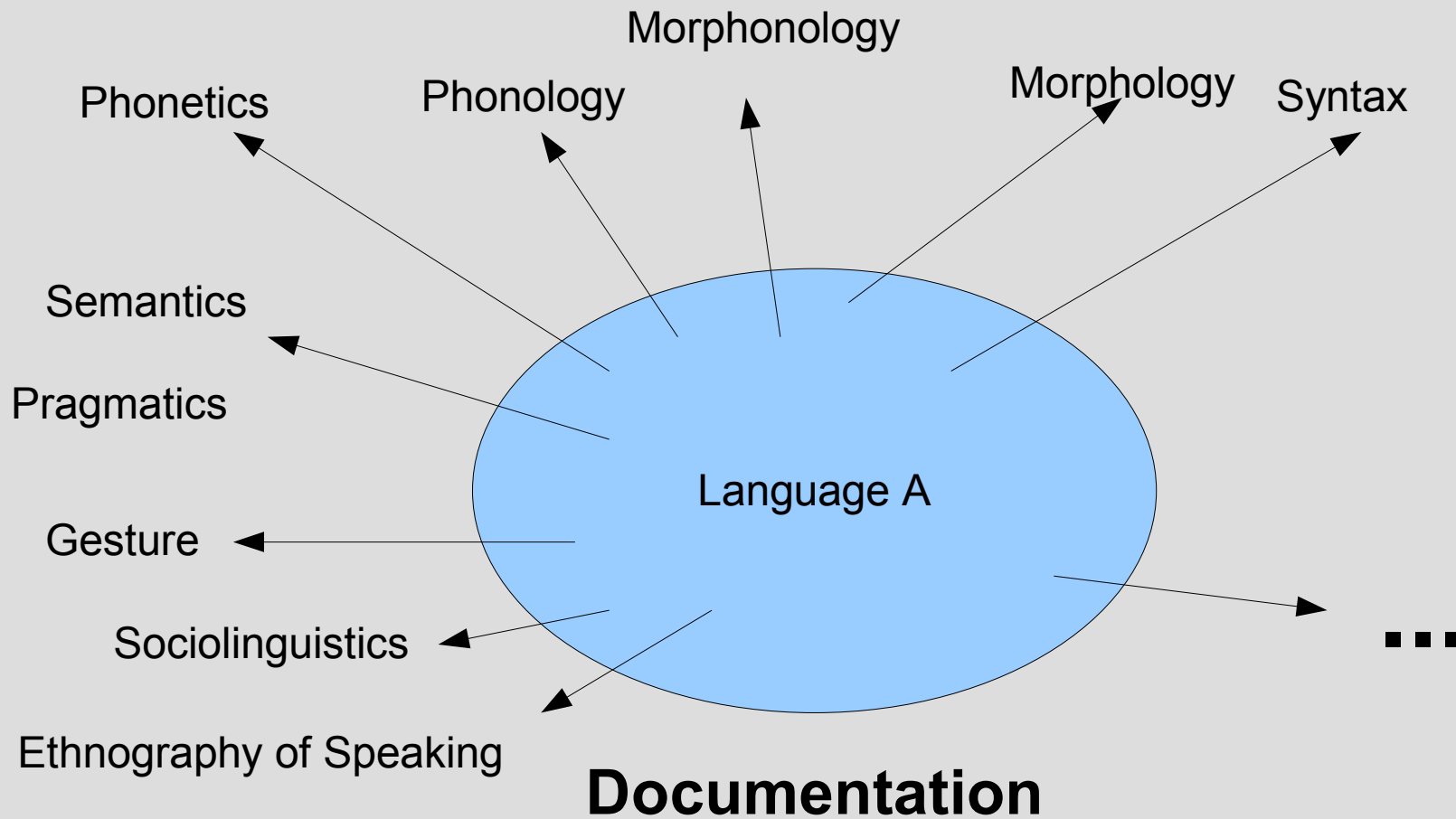
Documentation of a Semiotic System



Documentation of a Semiotic System



Documentation of a Semiotic System



Comprehensive Presentation of a Language

Language Description

Language Setting	
<i>Lexicon</i>	<i>Grammar</i>

Language Documentation

Metadata	
(Representation of) “Raw Data”	Primary Data

Comprehensive Presentation of a Language



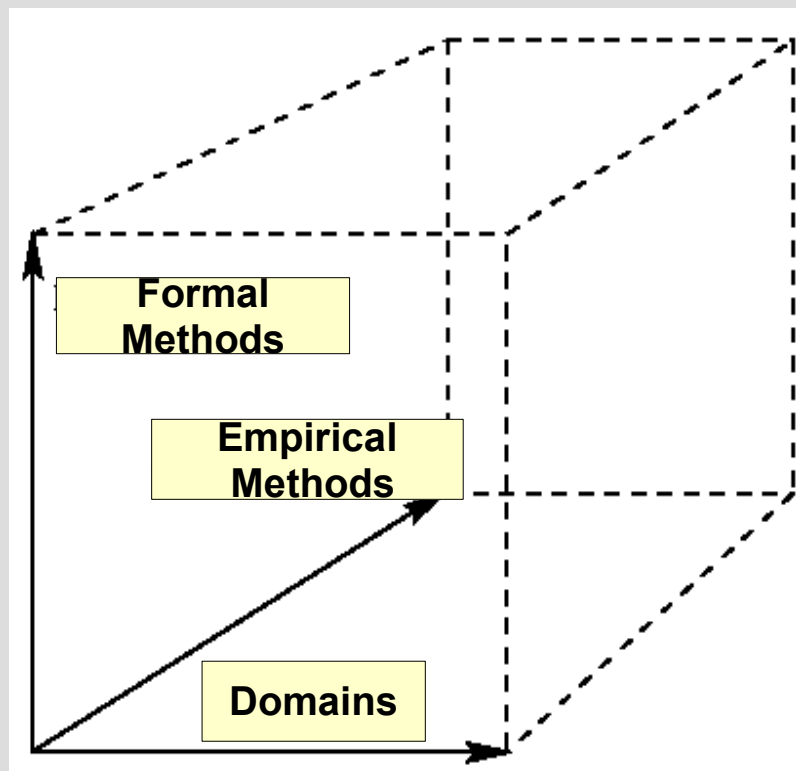
Language Description

Language Documentation

Comprehensive Presentation of a Language

The "method space" of linguistics and phonetics, like other sciences, comprises

- **formal methods**, i.e. theories, models, formalisms, notations and computer implementations (Gibbon, 2000). Gibbon (2005) defines the formal methods as systematic textual description or mathematical formalisation.
- **empirical methods**, i.e. techniques and tools for relating reality to the models



Comprehensive Presentation of a Language

- De facto there is no such thing as a “generative grammar”
- Lexicons are expandable and grammars never capture the whole system of a language
- Documentation is needed to have an empirical basis for analytic abstraction from the data

Data in Language Documentation

Data in Language Documentation

- Primary Data
 - As audio files (mono-modal)
 - As video files (multi-modal)
 - As text (technical medium)
 - As representation (IPA, ToBI, Syntactic Information, Morpheme Gloss, ...)
- Meta-data
 - Who?
 - Whom?
 - What?
 - When?
 - Where?
 - Why?
 - ...?

Data in Language Documentation

Collection – Archiving - Presenting

Data in Language Documentation

Austin, 2006:

- Data in *working* context - the way the data is stored during on-going research work of annotation and analysis
- Data in *archiving* context – how the materials are to be stored for **long-term** preservation
- Data in *presentation* context – the form of the data in distribution and publication

Data in Language Documentation

	<i>Working</i>	<i>Archiving</i>	<i>Presentation</i>
Text	<i>Word, XLS, FMPro, Toolbox/Shoebox</i>	XML	PDF, HTML
Audio	WAV	WAV, BMF	MP3, <i>WMA</i> , RA
Video	MPEG2	MPEG2, MPEG4	<i>QuickTime, AVI, WMV</i>

Collecting Data: Fieldwork

Language documentation is geared towards gathering an amount of data that is representative of the linguistic system as a whole. For this purpose the researcher has to overcome the observation paradox (Senft, 2002); means have to be found to *systematically observe language use by speakers when they are not systematically observed* (Senft, 2002). Such methods need to be found in order to find the data the empirical researcher desires to obtain.

Collecting Data: Fieldwork

The “sub-space” of empirical methods includes many dimensions, among these are:

- Introspection
- Experimental methods
- Corpus methods

An important issue in this is whether the methods used are quantitative or not.

(Gibbon, 2002)

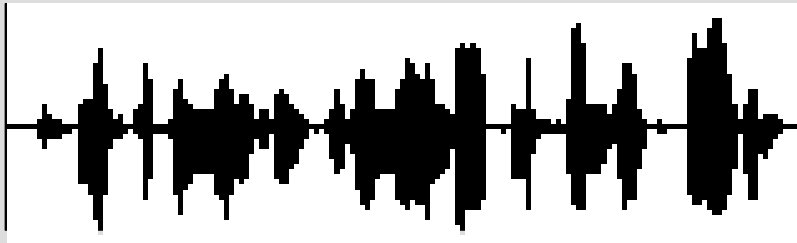
Collecting Data: Fieldwork

Data collection in the field involves a specific kind of complete prior planning:

- Expedition planning: mode of finance, international & local travel, local contacts, permissions (visas, regional permits)
- Personal planning: medical prophylaxis, standard medicines, region-specific accessories
- Research planning: research cooperation and permits, preparation of elicitation materials, recording and computing equipment , power supplies

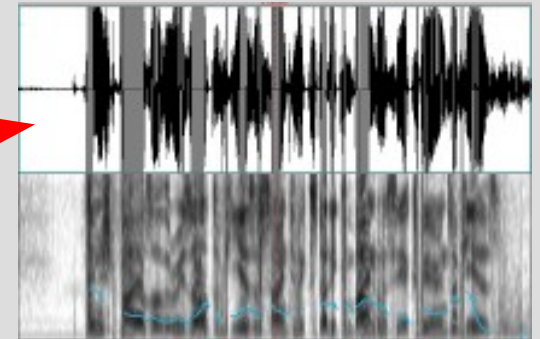
(Gibbon, 2002)

Collecting Data: Fieldwork

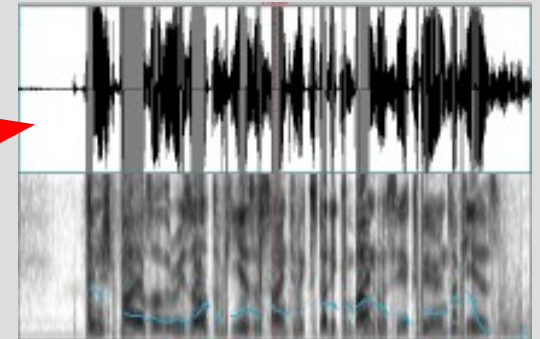


- Audio recordings
- Video recordings
- Text Collection
- ...

Collecting Data: Fieldwork



Collecting Data: Fieldwork



Collecting Data: Fieldwork

West African Language Data Sheets

Mary Esther Kropp Dakubu

1980 (typed 28 February 2001, Dafydd Gibbon; version: October 18, 2004)

Key to the Data Sheets
Clé des Fiches
Name of language / nom de langue

A. Demographic Data / Données Démographiques

B. Present state of classification of the language
Comment la langue est actuellement classée

C. Bibliography / Références bibliographiques

C.1. Basic grammars, dictionaries and major linguistic articles
Les grammaires, les dictionnaires, et les principaux articles ou cette langue a été étudiée dans une perspective linguistique

C.2. Publications in the language. Usually included only when there is little or nothing to mention in C.1.
Les publications dans cette langue. Indiquées seulement dans le cas où il n'y a rien à mentionner dans C.1.

D. Linguistic Data / Données Linguistiques

Name of dialect from which the data are drawn.
Nom du dialecte dont les données sont tirées.

D.1. Wordlist / Liste des mots

D.1.1. Nouns / Noms

1. woman/femme
2. man/homme

Collecting Data: Fieldwork

West African Language Data Sheets

Mary Esther Kropp Dakubu

1980 (typed 28 February 2001, Dafydd Gibben; version: October 18, 2004)

Key to the Data Sheets
Clé des Fiches
Name of language / nom de langue

A. Demographic Data / Données Démographiques

B. Present state of classification of the language Comment la langue est actuellement classée

C. Bibliography / Références bibliographiques

C.1. Basic grammars, dictionaries and major linguistic articles

Les grammaires, les dictionnaires, et les principaux articles ou cette langue a été étudiée dans une perspective linguistique

C.2. Publications in the language. Usually included only when there is little or nothing to mention in C.1.

Les publications dans cette langue. Indiquées seulement dans le cas où il n'y a rien à mentionner dans C.1.

D. Linguistic Data / Données Linguistiques

Name of dialect from which the data are drawn.
Nom du dialecte dont les données sont tirées.

D.1. Wordlist / Liste des mots

D.1.1. Nouns / Noms

1. woman/femme
2. man/homme

The screenshot shows a computer interface with three windows:

- Texts.txt**: A text editor window containing a template for a text record. The fields and their values are:

\id Text identification	Sample text
\ref Reference	Text.001
\tx Text	Sample
\mb Morphemes	sample
\ge Gloss	example
\ps Part of Speech	n
\ft Free Translation	Free translation.
\nt Notes	You should replace the contents of this text record with your first text.
\nt Notes	This is a startup kit for using the toolbox to make a dictionary and a set of interlinearized texts. Detailed instructions on how to use this startup kit are in the file "Instructions for Starting Your Own Project.doc". Please open it in Word.
- Dictionary.txt**: A text editor window showing a single entry:

\lx Lexeme	sample
\ps Part of speech	n
\ge English Gloss	example
\nt Notes	You should replace the contents of this entry with your first entry.
\dt Date edited	25/May/2003
- wordlist.db**: A database window showing a table with the following data:

Word	Count	References
sample	1	Text.001

Collecting Data: Fieldwork

3. head/tête
4. ear/oreille
5. eye/œil
6. nose/nez
7. mouth/bouche
8. tooth/dent
9. tongue/langue
10. foot/pied

A.
B.
C.
C.1
Les
etc

C.2
or 1
Les particularités dans cette langue. quelques éléments sont à citer ou il y a rien à mentionner dans C.1.

D. Linguistic Data / Données Linguistiques

Name of dialect from which the data are drawn.
Nom du dialecte dont les données sont tirées.

D.1. Wordlist / Liste des mots

D.1.1. Nouns / Noms

1. woman/femme
2. man/homme

The screenshot shows a software window titled 'Texts.txt' with the following content:

```

\id Text identification Sample text
\ref Reference Text.001
\tx Text Sample
\mb Morphemes sample
\ge Gloss example
\ps Part of Speech n
  
```

Below this is a menu bar (File, Edit, Database, Project, Tools, View, Window, Help) and a toolbar. A search box contains 'do not'. Below the toolbar is a window titled 'Nawtulaikli.dic' containing a table:

\lx Lexeme	\ps Part of Speech	\gl Gloss
moomoo	n	sound of cow
moomoor	n	cow
nuttin	n	nothing
pigaut	v	eat heavily
piggy	n	toe
pikkat	v	eat
pinky	n	finger
shrubby	n	tree
stinkinrat	n	mouse
t'	prep	to
tudeweerz	adv	until very late
tweetr	n	bird
wuf	v	eat
wufr	n	dog
va	nron	wou

A red arrow points from the 'mouth/bouche' entry in the list on the left to the 'shrubby' entry in the dictionary table.

Collecting Data: Fieldwork

The Leipzig Glossing Rules:

Rule 1: Word-by-word alignment

Interlinear glosses are left-aligned vertically, word by word, with the example. E.g.

(1) Indonesian (Sneddon 1996: 237)

Mereka di Jakarta sekarang.

they in Jakarta now

'They are in Jakarta now.'

Collecting Data: Fieldwork

The Leipzig Glossing Rules:

Rule 2: Morpheme-by-morpheme correspondence

Segmentable morphemes are separated by hyphens, both in the example and in the gloss. There must be exactly the same number of hyphens in the example and in the gloss. E.g.

(2) Lezgian (Haspelmath 1993: 207)

<i>Gila abur-u-n</i>	<i>ferma</i>	<i>hamišaluğ</i>	<i>giğiina</i>	<i>amuq'-da-c</i>
now they-OBL-GEN	farm	forever	behind	stay-FUT-NEG
'Now their farm will not stay behind forever.'				

Collecting Data: Fieldwork

*The Leipzig
Glossing Rules:*

...

Collecting Data: Fieldwork - Metadata

Metadata are:

- informally: headers, front matter, speaker details, permissions, ...
- more precisely: data about the form (genre, format, ...) and context (locale, participants, responsables, ...) of corpora
- more formally: a representation of an instantiation of the modelling conventions for a specific fieldwork situation
- thus for fieldwork: a variety of participants, genres, formats, locales, etc.
- but specifiable only in general terms because of the variety

→ “These are like library cards for the language material”
(D. Gibbon)

([Gibbon, 2002](#))

Collecting Data: Fieldwork - Metadata

Airstair Dry, 2004:

“Efficient search and retrieval of language resources requires the use of metadata”

- Title: **Biao Min Data**
- Creator (depositor): **David Solnit**
- Subject (linguistic field): **Language Description**
- Subject (language): **Biao Min**
- Date created: **April 5, 1982**
- Description: **The Biao Min data on the E-MELD site includes over 3,000 lexical items.**

Collecting Data: Fieldwork - Metadata

Airstair Dry, 2004:

“Efficient search and retrieval of language resources requires the use of metadata”

```
<title> Biao Min Data </title>
```

```
<creator xsi:type="olac:role" olac:code="depositor"> David Solnit  
</creator>
```

```
<subject xsi:type="linguistic-field" olac:code="language_description"/>
```

```
<subject xsi:type="olac:language" olac:code="x-sil-BJE"> Biao Min  
</subject>
```


→ ***XML – Metadata in XML-format (“standard”)***

Collecting Data: Fieldwork - Metadata

Airstair Dry, 2004:

OLAC
recommends 5 extensions:

- **Contributor**
 - Role
- **Language**
 - OLAC language
- **Creator**
 - Role
- **Subject**
 - OLAC Language
 - Linguistic Field
- **Type**
 - Linguistic Data Type
 - Discourse Type

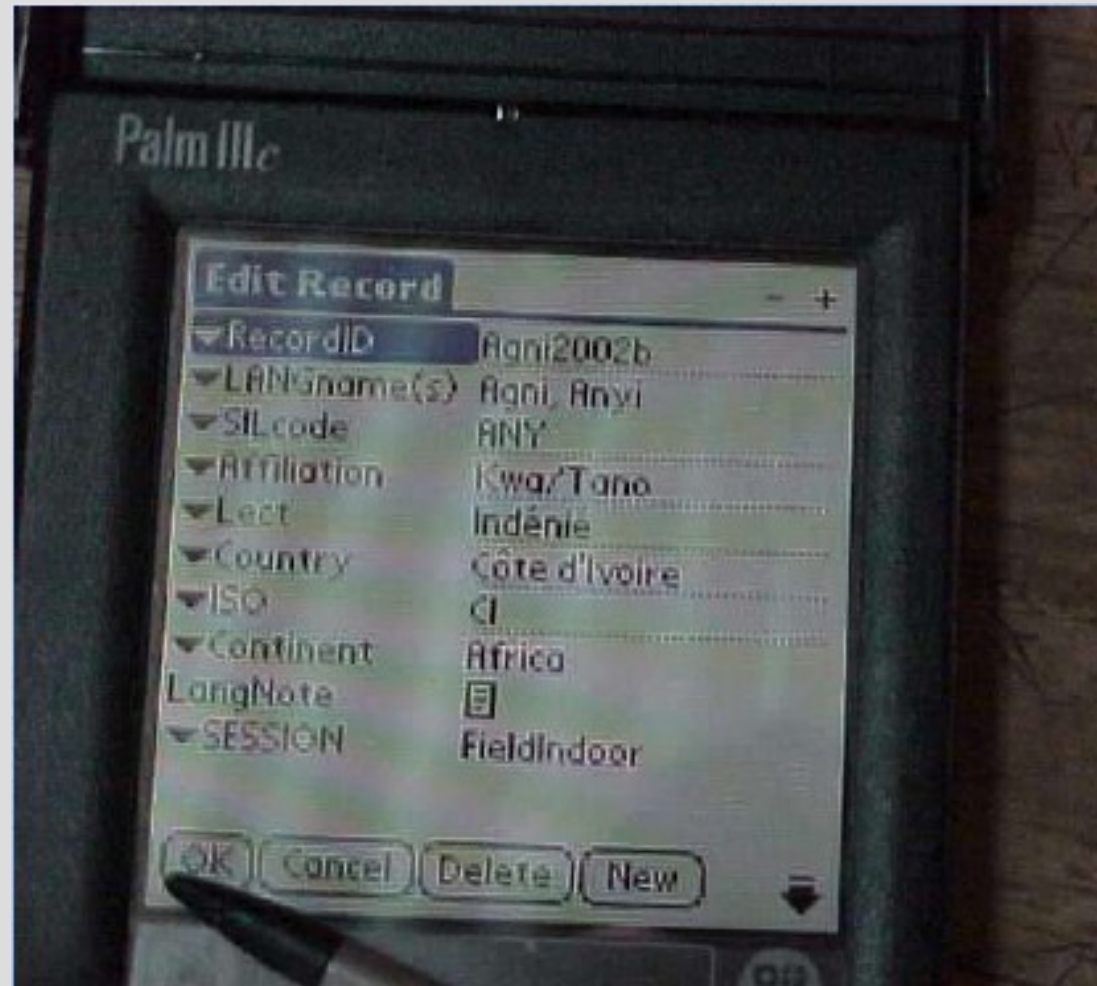

Jan 9, 2004

Symposium on Best Practice
LSA, Boston, MA

14

Collecting Data: Fieldwork - Metadata

Gibbon, 2002



Archiving Data

- Language archives are not “back-ups”
- Language archiving is not publication
- A language archive is a trusted repository created and maintained by an institution with a commitment to permanence and the long-term preservation of archived resources
- A language archive will have clear policies for materials acquisition, cataloguing, dissemination, assurance, forward migration formats, disaster recovery

(Austin, 2005)

Archiving Data

Archives are e.g.

- **OLAC** (Open Language Archives Community)
- **The Rosetta Project**
- **DoBeS**
- **LDC** (Linguistic Data Consortium)
- ...

Archiving Data (e.g. **EMELD**)

Field Data Input Tool - A web based tool for field linguists to insert their language data into a fully searchable online database.

You are working in
Wawa

[Change Language](#)

[Update User Profile](#)



FIELD

Field Input Environment
for Linguistic Data -
Lexicon Version

Archiving Data (e.g. **EMELD**)

Field Data
field lin
into a f

Make the appropriate changes to the grammatical categories for Wawa :
Add or delete grammatical categories using the checkboxes.
Roll the mouse over a term for a definition.
Click  to expand a category.

- (1)  [Adjective](#)
- (2)  [Adposition](#)
- (3)  [Adverb](#)
- (4)  [Clitic](#)
- (5)  [Connective](#)
- (6) [Deictic formative](#)
- (7)  [Determiner](#)
- (8) [Interjection](#)
- (9)  [Noun](#)
- (10)  [Particle](#)
- (11)  [ProForm](#)
- (12) [Verb](#)

Continue

Reset

used tool for
language data
database.



FIELD

Field Input Environment
for Linguistic Data -
Lexicon Version

Archiving Data

- Database
 - Relational database?
 - Other possibilities?
- Where is data stored? - What physical medium?
- Who can one “trust” with the data, if one submits it to larger archives for long-term preservation?

Data Presentation

- Presentation, Publication, Dissemination,...
- **Not** just *articles and papers...*
- *First:* Giving something back to the community! But what?
- *Second:* How, can other linguists and researchers obtain the data?
- “Normal Outputs”: PhD - , MA – dissertations, Lexicons, Grammars, Books, Technical Reports, etc.
- Is there a way of/need/desire/the possibility to make (some of) the material available for the general public? (cf. [Ega Web Archive](#))

The WELD-Paradigm

The WELD-Paradigm

Workable Efficient Language Documentation
(Gibbon, 2002):

- 1) Language documentation must be comprehensive
- 2) Language documentation must be efficient
- 3) Language documentation must be state-of-the-art
- 4) Language documentation must be affordable
- 5) Language documentation must be fair

References

- 1) Austin, P. K. (2006). Data and Language Documentation. In Gippert, J., Himmelmann, N. P., and Mosel, U., editors, *Essentials of Language Documentation*, chapter 4, pages 87 – 112. Mouton de Gruyter, Berlin/New York.
- 2) Gibbon, D. (2000). [Computational Lexicography](#). In Van Eynde, F. and Gibbon, D., editors, *Lexicon Development for Speech and Language Processing*, chapter 1, pages 1–42. Kluwer Academic Publishing, Dordrecht.
- 3) Gibbon, D. (2002b). [Workable Efficient Language Documentation: a Report and a Vision](#). *ELSNews*, 11.3.
- 4) Gibbon, D. (2005a). [Forms and Formalisations. A Note on Formal Linguistics](#). In *Proceedings of the first Student Conference on Formal Linguistics*, Poznan.
- 5) Himmelmann, N. (1998). [Documentary and Descriptive Linguistics](#). *Linguistics*, 36:161–195.
- 6) Lehmann, C. (1999). [Documentation of Endangered Languages: A Priority Task for Linguistics](#). *Arbeitspapiere des Seminars für Sprachwissenschaft der Universität Erfurt*.
- 7) Lehmann, C. (2001). [Language Documentation A Program](#). In Bislang, W., editor, *Aspects of Typology and Universals*. Akademie Verlag, Berlin.
- 8) Lehmann, C. (2002). [Structure of a Comprehensive Presentation of a Language](#). In Tsunoda, T., editor, *Basic materials in minority languages 2002*, pages 5–33. ELRP Publication Series B003, Osaka: Osaka Gakuin University.

Further Material

1) Data

- Bernard, H. R., Pelto, P. J., Werner, O., Boster, J., Romney, A. K., Johnson, A., Ember, C. R., and Kasanoff, A. (1986). The construction of primary data in cultural anthropology. *Current Anthropology*, 27(4):382 – 396.
- Lehmann, C. (2004a). Data in linguistics. *The Linguistic Review*, 21(3/4):275 – 310.
- Lehmann, C. (2006). Daten - korpora - dokumentation. In Kallmeyer, W. and Zifonun, G., editors, *Sprachkorpora – Datenmengen und Erkenntnisfortschritt*, pages 61–74, Berlin and New York. W. de Gruyter.

2) Further Documentation (specific)

- Lehmann, C. (2004b). Documentation of grammar. In *Lectures on endangered languages: 4. From Kyoto Conference 2001*, pages 61–74, Osaka. CLIPP. [PDF](#)
- Gibbon, D. (2005b). Spoken language lexicography: an integrative framework. In Zybatow, L., editor, *Translatologie - neue Ideen und Ansätze. Innsbrucker Ringvorlesungen zur Translationswissenschaft IV. Forum Translationswissenschaft*, pages 247–289. [PDF](#)

Further Material

1) Interlinear Text

- Bow, C., Hughes, B. and Bird, S. (2003). Towards a General Model of Interlinear Text, Proceedings of the EMELD Conference 2003. HTML PDF
- Toolbox
- Leipzig Glossing Rules HTML PDF

2) OLAC

- Bird, S. and Simons, G. (2004). Building an open language archives community. In Hillmann and Westbrook, editors, *Metadata in Practice: A Work in Progress*, pages 203–222. ALA Editions. PDF
- Bird, S. and Simons, G. (2001). The OLAC metadata set and controlled vocabularies. In *Proceedings of ACL/EACL Workshop on Sharing Tools and Resources for Research and Education*. PDF
- OLAC
- OLAC Metadata Set