Anthony Aristar, Helen Aristar-Dry
LINGUIST List, Eastern Michigan University
Jeff Good
University at Buffalo, The State University of New York
• The most important lesson you can learn in a course like this is...
• The most important lesson you can learn in a course like this is...

• The central problems are not technical but conceptual
• The most important lesson you can learn in a course like this is...

• The central problems are not technical but conceptual

• Principles are key; not implementation
• The most important lesson you can learn in a course like this is...

• The central problems are not technical but conceptual

• Principles are key; not implementation

• Implementation: The realization of a conceptual model within a software application
Nowhere is this clearer than the subject of linguistic data types. We tend to think of our data in terms of the particular implementations we use to code it. But, we should instead think of the implementation as an incidental “surface” representation of a more complex structure.
The Allegory of the Cave

- Shadows cast on wall
- Prisoners
- Fire
- Ascent to Sunlight
- Diffused Sunlight
- Roadway where puppet showmen perform

Source: <http://normanrschultz.org/Courses/graphics/Platocave.JPG>
The Allegory of the Cave

Shadowy Implementation

Structural Reality

Ascent to Sunlight

Diffused Sunlight

Roadway where puppet showmen perform

Shadows cast on wall

Prisoners

Source: <http://normanrschultz.org/Courses/graphics/Platocave.JPG>
• Shadowy implementations
  • Database record
  • Transcribed string
• Structural reality
  • Lexeme
  • Attested phrase
• Word list entry
• Word list entry

• Presentation form: *chien* n. dog
• Word list entry

• Presentation form: *chien* n. dog

• Logical structure:
• Word list entry

• Presentation form: *chien n.* dog

• Logical structure: 
<table>
<thead>
<tr>
<th>headword</th>
<th>pos</th>
<th>gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
• Word list entry

• Presentation form: *chien* **n.** dog

• Logical structure: | headword | pos | gloss |

• Or, perhaps more accurately...
Lexical Entry #43

chien

noun

dog

hasHeadword

hasPartOfSpeech

hasGloss
Headwords

Headword #43

hasTranscription

fromLanguage

French

chien

chien

French

Data Management & Archiving: Data types
chien
Transcription

Transcription #43

Orthographic

hasType

hasUnicodeString

chien
Transcription #43 has Unicode string "chien" and has type Orthographic.
Other data types

• That was just a partial analysis of a word list entry—one of the simplest data types

• Two most important data types
  • Dictionary entries
  • Annotated texts
A word list entry can be thought of as the simplest kind of dictionary entry.

A full dictionary entry will be much more complicated.

Documentation projects will rarely reach high levels of complication at early stages (if ever).

Dictionary entries

- **A word list entry can be thought of as the simplest kind of dictionary entry.**
- **A full dictionary entry will be much more complicated.**
- **Documentation projects will rarely reach high levels of complication at early stages (if ever).**
dog, n.¹

Forms: 1 docga, 3-7 dogge, (3, 6 doggue, 6 Sc. doig), 6-8 dogg, 3- dog. [late OE. docga (once in a gloss); previous history and origin unknown. (The generic name in OE., as in the Teutonic langs. generally, was hund: see HOUND.) So far as the evidence goes, the word appears first in English, as the name of a powerful breed or race of dogs, with which the name was introduced into the continental languages, usually, in early instances, with the attribute ‘English’. Thus mod.Du. dog, late 16th c. dogge (‘een dogghe, vn gros matin d'Engleterre, canis anglicus’, Plantijn Thesaur. 1573), Ger. dogge, in 16-17th c. dock, docke, dogg (‘englische Dock’, Onomast. 1582, ‘eine englische Docke’, 1653), LG. dogge, Da. dogge, Sw. dogg; F. dogue (‘le genereux dogue anglais’, Du Bellay 15..), It., Sp., Pg. dogo, Pg. also dogue; in all the languages applied to some variety or race of dog.]

I. The simple word.

1. a. A quadruped of the genus Canis, of which wild species or forms are found in various parts of the world, and numerous races or breeds, varying greatly in size, shape, and colour, occur in a domesticated or semi-domesticated state in almost all countries. These are referred by zoologists to a species C. familiaris; but whether they have a common origin is a disputed question.

Prudentius Glosses (Recd. 148/1) [Gloss to] canum [gen. pl.] docgena. a1225 Ancr. R. 288 His [the devil's] teë beoë attrie, ase of ane wode dogge. Dauid, ine sauter, cleoë hine dogge. Ibid. 290 bet tes dogge of helle kumëë. c1290 S. Eng. Leg. I. 307/281 A teie doggue. a1300 Cursor M. 13658 (Cott. & G.) pai scott him als a dog Right vte o pair synagog. 1393 LanGl. P. Pl. C. x. 261 Thi doge dar nat berke. 1460 J. Capgrave Chron. (1858) 281 Thei seide pleynly that it was no more trst to the Pope writing thenselebl stribu" s. 68 Trec. R. Magna. P. ii. 576 who holdeth out thei...
Basic entry structure

- Core components
  - Written representation of lexeme, i.e., form (e.g., orthographic representation)
  - Morphosyntactic information, i.e., grammar (e.g., part of speech)
  - Semantic description, i.e., meaning (e.g., translation into vehicular language)
  - Plus: Metadata (ID, date edited, etc.)
Other components

- Some secondary components
- Cross references
- Etymology
- Register tags
- Obsolete
- Technical
- Vulgar
- ...

**phenon** ('fənən). [f. Gr. φαινειν to appear + -on.] a. **Biol.** A group of apparently similar plants or animals.

1943 CAMP & GILLY in Brittonia IV. 335 Phennon: a species which is phenotypically homogeneous and whose individuals are sexually reproductive, but which is composed of intersterile segments. 1969 E. MAYR Princ. Systematic Zool. i. 5 There is no generally accepted technical term for a phenotypically reasonably uniform sample, but it may be designated as a phenon. *Ibid.* 10 A phenon is not necessarily a population in the biological sense.

b. **Taxonomy.** A grouping of plants or animals established by techniques of numerical analysis.

1962 Sneath & Sokal in Nature 3 Mar. 860/1 How should we name the groups which are established by numerical taxonomy?... We call the groups simply 'phenons'. 1963 Davis & Heywood Princ. Angiosperm Taxon. iv. 136 Sneath & Sokal have introduced the concept of a phenon for the definition of groups obtained by cluster analysis... In practice a phenon defines groups by drawing lines horizontally across the dendograms. 1963 Sokal & Sneath Princ. Numerical Taxon..ix. 251 The phenon nomenclature... Phenons are groups which approach natural taxa more or less closely, and... they can be of any hierarchic rank or of indeterminate rank. 1966 New Scientist 20 Jan. 151/3 These groups are called phenons... If the original taxa are species each of the three new phenon taxa might represent a sub-genus or genus. 1973 Sneath & Sokal Numerical Taxon. v. 294 The term phenon is intended to be general, to cover the groups produced by any form of cluster analysis.

• Written representations are usually relatively short in length

• Often drawn from a different character set than other components

• Sometimes multiple forms needed (e.g., singular/plural in some languages)
### Writing implementation

**Form (plus some grammar)**

<table>
<thead>
<tr>
<th>Naki-Lexicon</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>bwɛ</strong></td>
<td><strong>bwɛ</strong></td>
</tr>
<tr>
<td><strong>M</strong></td>
<td><strong>H</strong></td>
</tr>
</tbody>
</table>

- Checked tone: [ ] Yes  [x] No  [ ] Other...
- Entry ID: Naki73191667471

<table>
<thead>
<tr>
<th>Guthrie CS</th>
<th>BLR Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>174</td>
<td>282</td>
</tr>
</tbody>
</table>

- Notes:

- Record:
  - 1
- Found:
  - 1
- Total:
  - 1769

---

**June 25, 2008**  **Data Management & Archiving: Data types**
• Prototypically involves specification of categories in “symbolic” fashion, e.g.,

- Part of speech
- Gender
- Conjugation

Categories are ideally associated with different input methods than simple “keyboarding”
Grammar implementation

The image shows a screenshot of a software interface labeled "Naki-Lexicon." The interface is used to manage data, specifically lexicon entries. The screenshot highlights a section labeled "Grammar," which includes columns for entries such as "bwé," "bwé," and "H." Other columns include "Checked tone," with options for Yes, No, and Other, and fields for "Guthrie CS," "BLR Index," and "Notes." Additionally, there are fields for "Entry ID" with an example entry "Naki73191667471." The layout is designed to facilitate data management and archiving in the context of linguistics, focusing on the grammar of a specific language.
• For a documentation project, word senses are usually described in a language different from the one being documented

• More “free form” than other parts of the lexeme entry/description

• Multiple senses likely to be needed frequently
The image displays a computer interface titled "Naki-Lexicon" with various fields for inputting data. The interface includes fields for "bwε", "bwε", "M", "H", "n", "dog", "Guthrie CS", "BLR Index", and "Entry ID". The "Entry ID" field contains the text "Naki73191667471".
Sense implementation

Naki-Lexicon

bwε
bwε
M
H

Meaning
dog

Guthrie CS  BLR Index
174       282

Entry ID
Naki73191667471

Browse

Data Management & Archiving: Data types

June 25, 2008
• This is just a rough, very general sketch

• Each language will present its own problems

• And each research project will, too—for example, comparative lexicons require further levels of structure

• In the early stages, your analyses—and, therefore, your implementations—will change
Data structure

Example data

headword  POS  gloss

puella  n  girl
**Data structure**

- **puella**: n, girl
- puellae
- puellae
- puellam

**Example data**

**headword** | **POS** | **gloss**
---|---|---
genitive

dative

accusative
Data Management & Archiving: Data types

Data structure

headword POS gloss

genitive
dative
accusative

puella n girl
puellae
puellae
puellam

Example data
Data management & archiving: Data types

Data structure

- Headword
- POS
- DEC
- Gloss

Example data

- puella
- puellae
- puellam

- Genitive: puellae
- Dative: puellae
- Accusative: puellam

Codification
Data structure

Example data
• Once you have an initial understanding of your data structure, you can implement it
• You will always lose important data when implementing
• But at least you’ll know you did this
• Documentation—e.g., of the meaning of the fields in a database—will alleviate the problem
My implementation

Form (plus some grammar)
### Naki-Lexicon

<table>
<thead>
<tr>
<th>bwɛ</th>
<th>bwɛ</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>H</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>dog</td>
<td></td>
</tr>
</tbody>
</table>

**Guthrie CS**  
174

**BLR Index**  
282

**Notes**

**Entry ID**  
Naki73191667471

---

**Record:** 1

**Found:** 1

**Total:** 1769

**Unsorted:** 100
Grammar
My implementation

Naki-Lexicon

Meaning

dog

Guthrie CS  BLR Index
174          282

Notes

Entry ID
Naki73191667471

June 25, 2008
My implementation
### Naki-Lexicon

<table>
<thead>
<tr>
<th>gwə</th>
<th>bwe</th>
<th>Checked tone</th>
<th>Yes</th>
<th>No</th>
<th>Other...</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>H</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Guthrie CS**
- 174

**BLR Index**
- 282

**Notes**
- Entry ID
- Naki73191667471

---

**Browse**
- Layout: Full
- Record: 1
- Found: 1
- Total: 1769
- Unsorted: 100

---

**My implementation**

**Data Management & Archiving: Data types**

---

**June 25, 2008**

---
Metadata
My implementation

Naki-Lexicon

bwɛ
M
h

dog

Guthrie CS
174
BLR Index
282

Notes

Entry ID
Naki73191667471

Record:
1

Layout: Full

Found: 1

Total: 1769

Unsorted 100

Browse

Checked tone

Yes ☐ No ☒ Other...
• Translation of *regarder*: look at

• Gloss of *regarder*: ‘look’

• Definition of *regarder*: to direct one’s sight upon
The most prominent annotated text structure is interlinear glossed text (IGT).

Cicko, [ch’aara ’a goj,] ’i bu’u.
cat.ERG fish & see.CVPAN 3S.ABS B.eat.PRS
‘The cat sees a fish and eats it.’
The most prominent annotated text structure is interlinear glossed text (IGT).

Subject Language

Cicko, [ch’aara ’a goj,] ’i bu’u.
cat.ERG fish & see.CVPAN 3S.ABS B.eat.PRS
‘The cat sees a fish and eats it.’
• The most prominent annotated text structure is **interlinear glossed text** (IGT)

Cicko, [ch’aara ’a goj,] ’i bu’u.
cat.ERG fish & see.CVPAN 3S.ABS B.eat.PRS
‘The cat sees a fish and eats it.’
The most prominent annotated text structure is interlinear glossed text (IGT).

Cicko, [ch’aara ’a goj,] ’i bu’u.

cat.ERG fish & see.CVPAN 3S.ABS B.eat.PRS

‘The cat sees a fish and eats it.’
The most prominent annotated text structure is interlinear glossed text (IGT)

Cicko, [ch’aara ’a goj,] ‘i bu’u.
cat.ERG fish & see.CVPAN 3S.ABS B.eat.PRS
‘The cat sees a fish and eats it.’
The most prominent annotated text structure is **interlinear glossed text (IGT)**

Cicko,  
[ch’aara ’a goj,]  
‘i  
bu’u.  
cat.ERG fish & see.CVPAN 3S.ABS B.eat.PRS

‘The cat sees a fish and eats it.’
The most prominent annotated text structure is interlinear glossed text (IGT)

Cicko, [ch’aara ’a goj,] ’i bu’u.
cat.ERG fish & see.CVPAN 3S.ABS B.eat.PRS
‘The cat sees a fish and eats it.’
• The most prominent annotated text structure is interlinear glossed text (IGT)

Cicko, \([ch’aara ’a goj,] \quad ’i \quad bu’u.
\text{cat.ERG} \quad \text{fish} \quad \& \quad \text{see.CVPAN} \quad 3\text{S.ABS} \quad \text{B.eat.PRS}

‘The cat sees a fish and eats it.’
The most prominent annotated text structure is interlinear glossed text (IGT).

Cicko, [ch’aara ’a goj,] ’i bu’u.
cat.ERG fish & see.CVPAN 3S.ABS B.eat.PRS
‘The cat sees a fish and eats it.’
• The most prominent annotated text structure is interlinear glossed text (IGT)

Alignment

\[
\begin{array}{cccccc}
\text{Cicko,} & [\text{ch’ara} & ’a & \text{goj,}] & ’i & \text{bu’u.} \\
\text{cat.ERG} & \text{fish} & \& & \text{see.CVPAN} & 3\text{S.ABS} & \text{B.eat.PRS} \\
\end{array}
\]

‘The cat sees a fish and eats it.’
There is a range of relationships that can hold among annotations.

In addition to the traditional ones depicted above, the possibility of time-aligned annotations adds a number of new possibilities.

Also, annotations may also be associated with metadata.
Time alignment
Time alignment

[Image of a software interface showing a transcription and audio waveform]
Time alignment
Time alignment
• Annotation content: Literal vs. symbolic

• Annotation anchoring: External vs. internal

• Annotation relationships
  • One-to-one
  • Subdivision
    • Externally-anchored
    • Internally-anchored
Annotation content

- Literal content
- Transcription string
- Translation string
- Symbolic content
  - Category label (e.g., morphosyntactic category like “conditional construction”)
  - Bookkeeping/metadata fields (e.g., “checked tone?”)
String as symbol

Naki-Lexicon

bwε

bwε

M

H

dog

Guthrie CS 174

BLR Index 282

Notes

Entry ID

Naki73191667471

June 25, 2008

Data Management & Archiving: Data types
### String as Symbol

#### Naki-Lexicon

<table>
<thead>
<tr>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>dog</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Guthrie CS</th>
<th>BLR Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>174</td>
<td>282</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Entry ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naki73191667471</td>
</tr>
</tbody>
</table>
• **External anchoring**
  • Time-aligned to an audio file
  • Aligned to a location on a scanned image

• **Internal anchoring**
  • A translation aligned with a transcription
  • A note aligned with another annotation
Annotation relationships

- One-to-one
- A free translation of a transcription
- A gloss of a single word
- Subdivision
  - Division of a phrase into words
  - Division of a word into morphemes
• Subdivisions, too, can be externally-aligned or internally-aligned

• Externally-anchored: Division of sentence transcription into time-aligned words

• Internally-anchored: Division of word into morphemes with no indication of the particular times associated with the pronunciation of each morpheme
The boy loves the girl.
Sentence-level transcription time-aligned (external anchoring) with waveform.

Schematic example of tier:

```
puer  puellam  amat
puer  puell   am   am   a   t
boy   girl    ACC  love  PRS  3s
```

The boy loves the girl.
The boy loves the girl.

<table>
<thead>
<tr>
<th>puer</th>
<th>puellam</th>
<th>amat</th>
</tr>
</thead>
<tbody>
<tr>
<td>puer</td>
<td>puell</td>
<td>am</td>
</tr>
<tr>
<td>boy</td>
<td>girl</td>
<td>ACC</td>
</tr>
</tbody>
</table>

Time-aligned (external anchoring) subdivision of sentence into words.

Puer puellam amat.

The boy loves the girl.
Internally-anchored subdivision of words into morphemes

<table>
<thead>
<tr>
<th>puer</th>
<th>puellam</th>
<th>amat</th>
</tr>
</thead>
<tbody>
<tr>
<td>puer</td>
<td>puell</td>
<td>am</td>
</tr>
<tr>
<td></td>
<td></td>
<td>am</td>
</tr>
<tr>
<td></td>
<td></td>
<td>a</td>
</tr>
<tr>
<td></td>
<td></td>
<td>t</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>boy</th>
<th>girl</th>
<th>ACC</th>
<th>love</th>
<th>PRS</th>
<th>3s</th>
</tr>
</thead>
</table>

The boy loves the girl.
Internally-anchored one-to-one association of morphemes with glosses

Schematic example of tier

Puer puellam amat.

<table>
<thead>
<tr>
<th>puer</th>
<th>puellam</th>
<th>amat</th>
</tr>
</thead>
<tbody>
<tr>
<td>puer</td>
<td>puell</td>
<td>am</td>
</tr>
<tr>
<td>boy</td>
<td>girl</td>
<td>ACC</td>
</tr>
</tbody>
</table>

The boy loves the girl.
Internally-anchored one-to-one association of sentence transcription with free translation.

<table>
<thead>
<tr>
<th>puer</th>
<th>puellam</th>
<th>amat</th>
</tr>
</thead>
<tbody>
<tr>
<td>boy</td>
<td>girl</td>
<td>ACC</td>
</tr>
</tbody>
</table>

The boy loves the girl.
**Strings representing symbols**

<table>
<thead>
<tr>
<th>puer</th>
<th>puellam</th>
<th>amat</th>
</tr>
</thead>
<tbody>
<tr>
<td>puer</td>
<td>puell</td>
<td>am</td>
</tr>
<tr>
<td></td>
<td></td>
<td>am</td>
</tr>
<tr>
<td></td>
<td></td>
<td>a</td>
</tr>
<tr>
<td></td>
<td></td>
<td>t</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>boy</th>
<th>girl</th>
<th>ACC</th>
<th>love</th>
<th>PRS</th>
<th>3s</th>
</tr>
</thead>
</table>

The boy loves the girl.
Subdivision and relationships only implicit!
• Implementation implies conceptual analysis
• Good implementation implies good conceptual analysis
• Conceptual analysis is ongoing—implementation will need adjustment
• There are other data types: Grammatical notes, paradigms, metadata
• Each requires its own conceptual analysis