Archiving and linguistic databases

Jeff Good, MPI EVA
(good@eva.mpg.de)

LSA Annual Meeting
Oakland, California
January 6, 2005

Available at: http://email.eva.mpg.de/~good/databases.pdf

Goals

- Cover important conceptual issues in designing a linguistic database
- Discuss some steps to take in building a database
- Discuss practical issues in creating archivable versions of databases

What is a database?

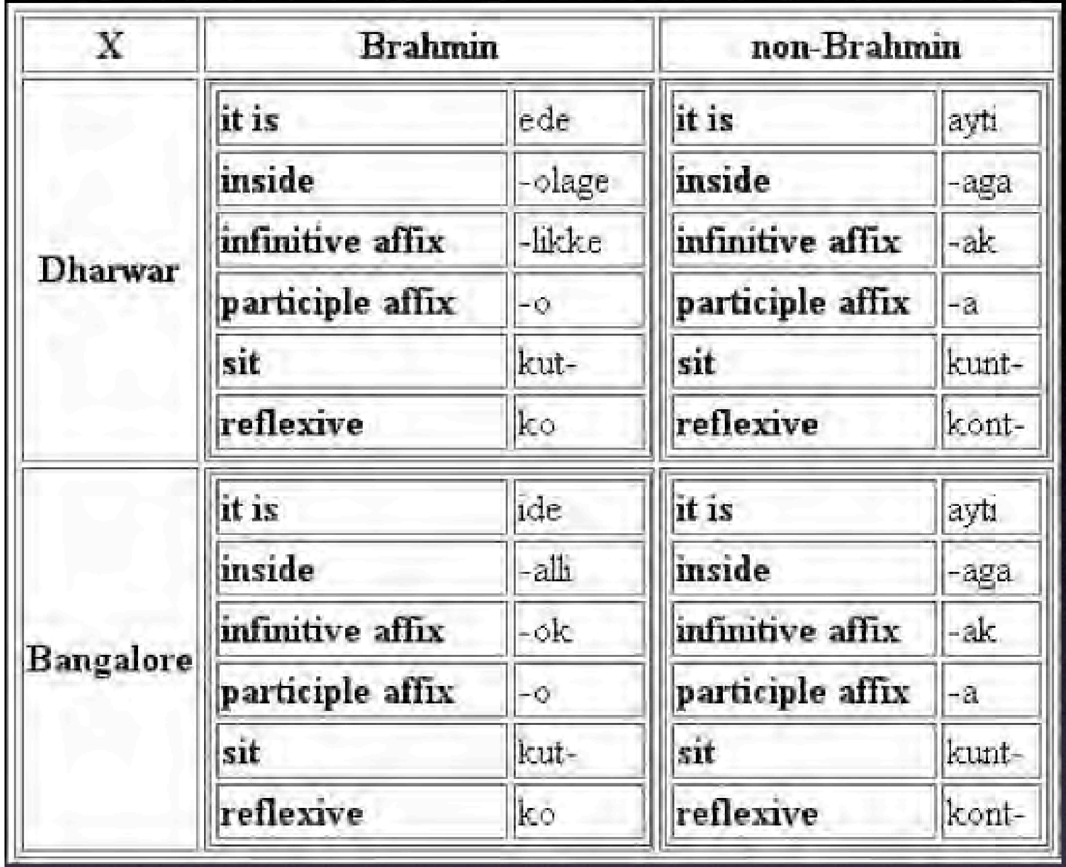
- Here, at least, I’m considering it to be any digitally-encoded data which is structured in a well-defined way
- A dictionary, a text corpus could be considered a database in this sense
- A journal article would not be a database in this sense

Databases overview

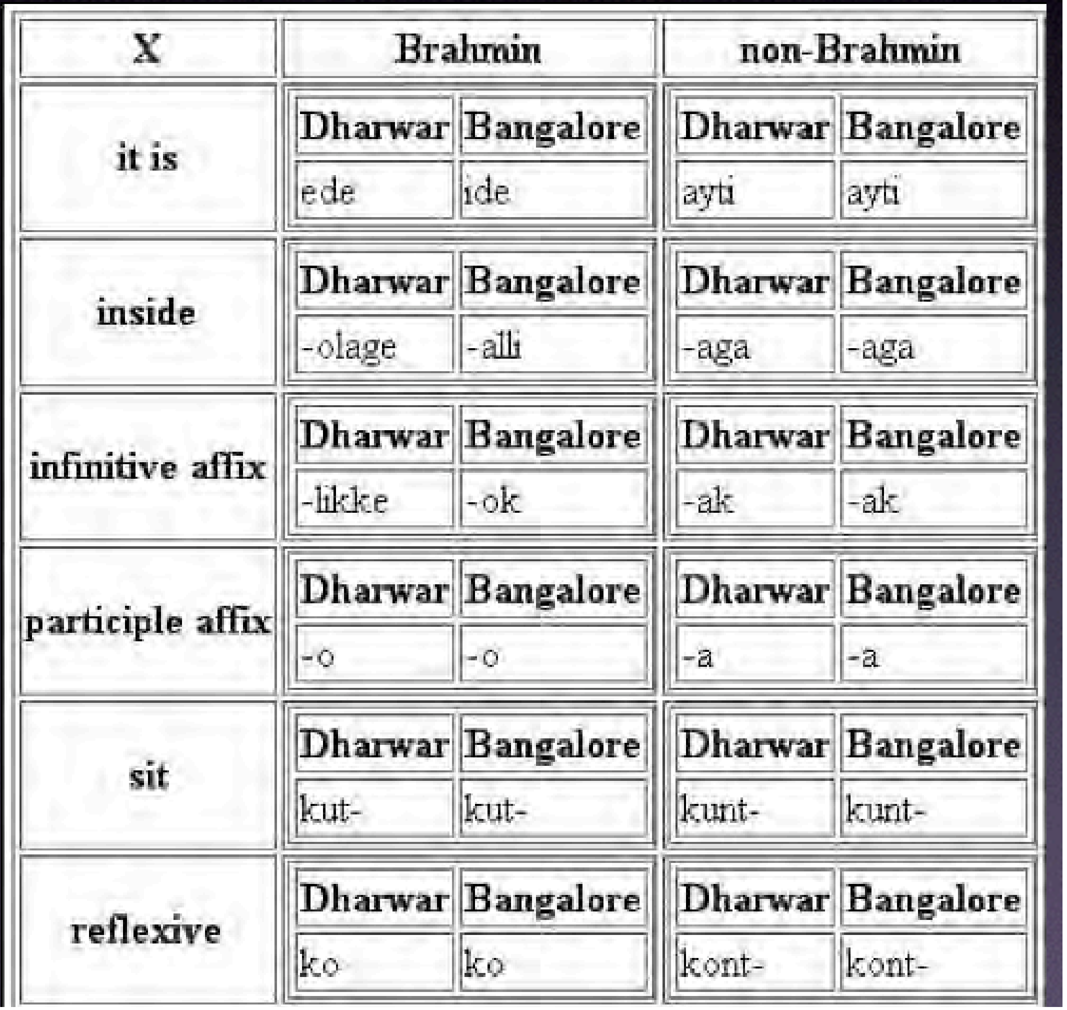
- One could, in principle, encode a database in files produced by a word processor
- However, the existence of more specialized tools like database and spreadsheet software allows one to encode the logical structure of some set of data
- By using a logical encoding, it then becomes easy to quickly generate useful different “views” of a single underlying data set

Database views

- A given underlying logical structure must be given some “surface” structure to be viewed by humans
- The following example of multiple views of a Kanarese paradigm comes from Penton et. al (2004)

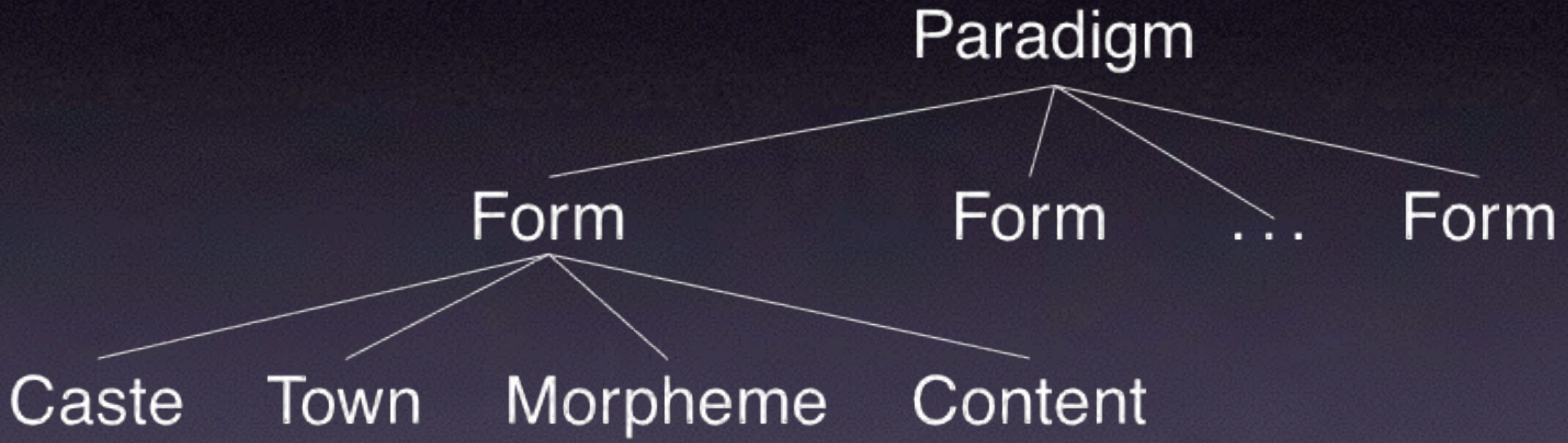


X	Brahmin		non-Brahmin	
Dharwar	**it is**	ede	**it is**	ayti
	inside	-olage	**inside**	-aga
	infinitive affix	-likke	**infinitive affix**	-ak
	participle affix	-o	**participle affix**	-a
	sit	kut-	**sit**	kunt-
	reflexive	ko	**reflexive**	kont-
Bangalore	**it is**	ide	**it is**	ayti
	inside	-alli	**inside**	-aga
	infinitive affix	-ok	**infinitive affix**	-ak
	participle affix	-o	**participle affix**	-a
	sit	kut-	**sit**	kunt-
	reflexive	ko	**reflexive**	kont-



X	Brahmin		non-Brahmin	
it is	**Dharwar**	**Bangalore**	**Dharwar**	**Bangalore**
	ede	ide	ayti	ayti
inside	**Dharwar**	**Bangalore**	**Dharwar**	**Bangalore**
	-olage	-alli	-aga	-aga
infinitive affix	**Dharwar**	**Bangalore**	**Dharwar**	**Bangalore**
	-likke	-ok	-ak	-ak
participle affix	**Dharwar**	**Bangalore**	**Dharwar**	**Bangalore**
	-o	-o	-a	-a
sit	**Dharwar**	**Bangalore**	**Dharwar**	**Bangalore**
	kut-	kut-	kunt-	kunt-
reflexive	**Dharwar**	**Bangalore**	**Dharwar**	**Bangalore**
	ko	ko	kont-	kont-

Logical structure



Paradigm

Form Form . . . Form

Caste Town Morpheme Content

The logical structure of the Kanarese paradigm

Logical structure

- Linguists do not generally think explicitly about the logical structure of the types of data they work with
- However, we do frequently work with data formats for which there are standardized ways of presenting their logical structure
- For example, a word list entry
    - Example entry: *chien* **n.** dog
    - Logical structure: *headword* **pos.** gloss

Building a database

- Things to consider when building a database
  - What is the logical structure of my data?
  - What kinds of views (or products) do I intend to produce with the database?
  - Do I have special computing needs limiting my software choices (e.g., need special character support, primarily working online/offline, only have limited computing power)?

Building a database

- There are many tools which can produce linguistic databases, though not all are suited for encoding all kinds of logical structures
  - For complex logical structures specialized database software, e.g. FileMaker Pro, SQL database, may be required
  - For simple databases, software which is good at producing tables, e.g., Microsoft Excel or Microsoft Word
  - XML editor for producing XML databases

Archiving

- Your choice of a tool will also be influenced by the products you wish to produce
- The one product which needs to be considered at the outset by any project is the archival format of the database

Archiving

- For now, the only electronic archival formats for databases are text files formatted with a machine-readable encoding of the logical structure of the data in the database
- The overarching goal of an archive format: Self-documenting, machine-readable encoding of logical structure
- In theory, best practice is to use XML
- In practice, the necessary tool support isn't sufficient for the needs of the "ordinary working linguist"

Archiving

- Self-documenting, machine-readable word-list record in XML

```
<entry>
        <headword>chien</headword>
        <pos>n.</pos>
        <gloss>dog</gloss>
</entry>
```

Archiving

- Same kind of data, not best practice, but still good practice, in tab-delimited text with carriage returns separating records

```
headword        pos     gloss
chien           noun    dog
chat            noun    cat
...
```

Archiving

- Some common bad practices
  - Not regularly producing an archive format for your database (e.g., working solely with a FileMaker or Excel file)
  - Not documenting the structure of your database and notational conventions used within it

Summary

- Come to an understanding of the logical structure of your data before building a database
- Consider the kinds of views you will need of your data when choosing a tool for building a database
- From the outset, develop a plan for regularly producing a version of your database in an archive format

Reference

Penton, David, Catherine Bow, Steven Bird, and Baden Hughes. 2004. Towards a general model for linguistic paradigms. Proceedings of the E-MELD 2004 Workshop on Linguistic Databases and Best Practice, Detroit, Michigan. Available at: http://emeld.org/workshop/2004/bird-paper.pdf

Acknowledgements

I would like to thank all the presenters and participants at the 2004 E-MELD workshop on Linguistic Databases and Best Practice. The bulk of the content of this talk consists of my own interpretation of the discussion at that workshop.

Available at: http://email.eva.mpg.de/~good/databases.pdf