

# MULTIDIMENSIONAL EXPLORATION OF LINGUISTIC DATABASES

**Steven Bird**

University of Pennsylvania

<http://www.ldc.upenn.edu/sb>

# NSF PROJECT

---

This project aims to foster a new mode of fundamental research in linguistics, namely “web-based exploration of linguistic field data.”

## **Objectives**

1. to develop tools for manipulating linguistic databases,
2. to store and disseminate large datasets using the model,
3. to exploit the tools and datasets in teaching and research,
4. to explore new methods for representing and analyzing multimodal linguistic data.

Collaborators: supply datasets, test tools

LDC website for dissemination

# LINGUISTIC DATABASES

---

1. field notes
2. balanced corpora for experimental work
3. reference corpora

## **Field Linguistics**

- diverse motivations and types of data collections
- partial datasets reported in inaccessible publications

# COMPUTATIONAL SUPPORT

---

- should make few assumptions about the physical location of the elicitation, the choice of language, and the purpose
- should embody no prior commitment to particular linguistic structures or theoretical models

Commitment to:

1. large-scale collection; opportunistic sampling from a large space of possible combinations of elements and structures
2. query and transformation facilities to support wide-ranging exploration and multiple visualizations and projections of data
3. dissemination of the primary data along with all of the commentaries and links created by the investigator.

# DISSEMINATING LINGUISTIC DATABASES

---

## Complexity of Data

<i>low</i>	<i>high</i>
<b>CATALOGUE SITES</b>	<b>WEB-BASED INFORMATION SYSTEMS</b>
<b>WEB-PRESENCE SITES</b>	<b>SERVICE-ORIENTED SITES</b>
<i>low</i>	<i>high</i>

## Complexity of Applications

# PHASE 1: HYPERLEX VERSION 2

---

- generalization of software to facilitate reuse in other projects
- existing datasets and recordings
- new speech corpora
- browser plugin version
- high-level query interface
- embedding tabulations and queries in documents
- use in field methods classes

## PHASE 2: NEW MODELS AND PROTOTYPES

---

- data models
- prototypes
- data evolution
- new speech recordings
- fonts

## **Phase 3: Analysis, Dissemination, Teaching**

# CALL FOR PARTICIPATION

---

1. document key functions of existing or desired software
2. provide multi-modal datasets
3. road-test new prototypes