Annotated Corpora in Linguistic Research

Steven Bird & Stephanie Strassel

Linguistic Data Consortium
University of Pennsylvania
Philadelphia, PA 19104
www.ldc.upenn.edu

The Linguistic Data Consortium

- a not-for-profit organization
- serving:

  researchers, educators, technology developers
- in language-related fields by:

  creating/collecting, archiving, distributing language resources

Continuing Need

Research, development and teaching

- vast amounts of data, many languages
- 'ecologically valid' data
- range of specialized skills
  - processing of newswire, closed caption video, VOA, ...
  - intellectual property rights
  - data transformation and publication: CD, Web
- collaborative research
- replication of results

LDC Principles

Publish the data that researchers need

- sponsored programs (TDT, Hub-4, OLEADA,...)
- community initiatives (ACL/DCI, Unipen...)
- non-LDC projects (CSAE, CELEX, Trains...)
- LDC-funded (Treebank, COMLEX, WordNet...)

Make data available to everyone

- consortium membership is open to all
- most databases available to non-members

Promote the idea of shared resources

- IPR intermediary
- advice on collection, publication and IPR issues
- develop standards and tools

Publications

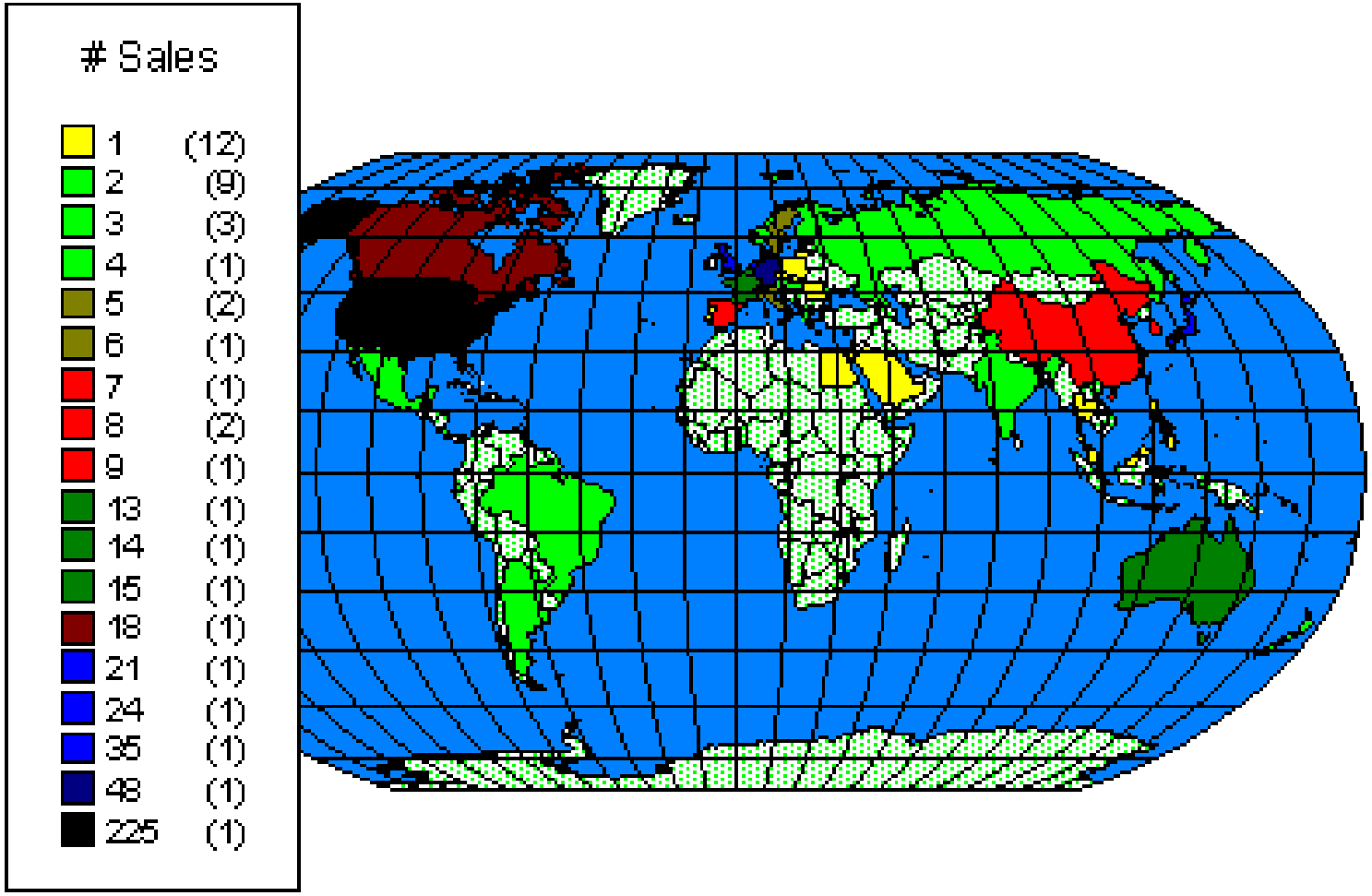
140 Publications + 12 in press

- 93 speech corpora + 7 in press
- 33 text corpora + 5 in press
- 14 lexicons
- 250 Gb

Distribution

- 267 members
- 520 nonmembers

Distribution

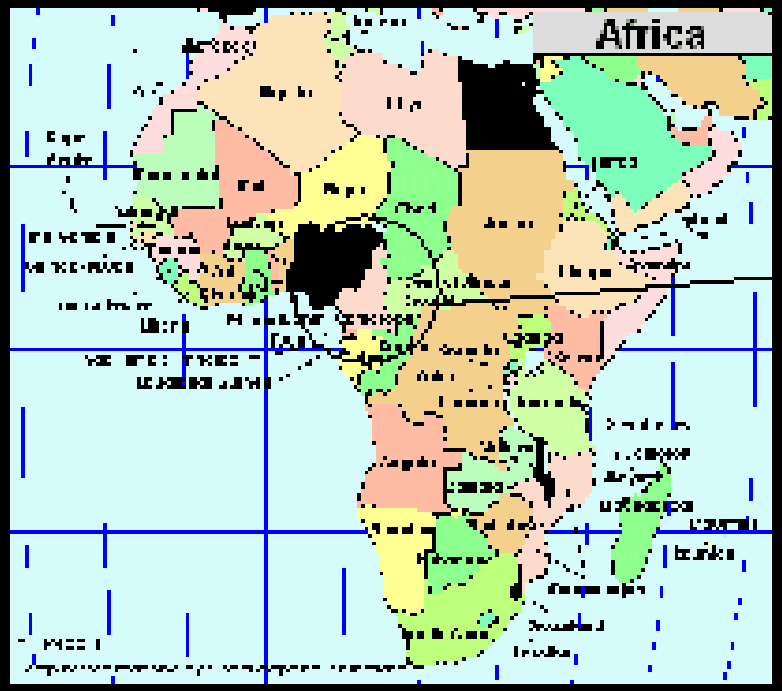


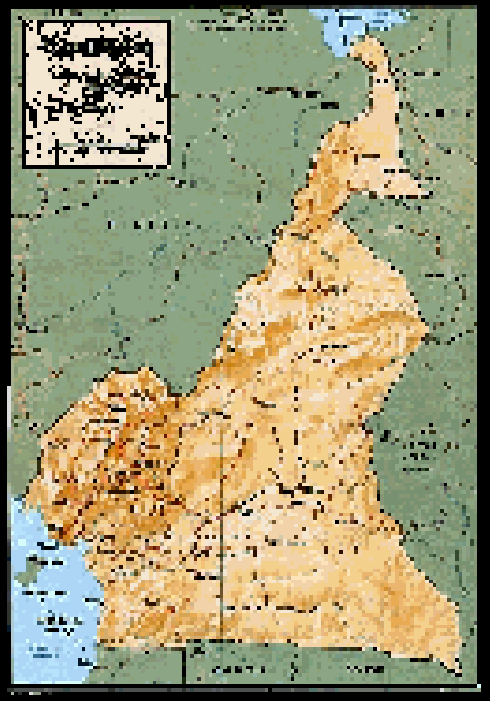
Languages

	Speech / Transcripts						
Language	Broadcast	Telephone	WideBand	Parallel Text	Newswire/ Other Text	Lexicon	Traditional Dictionary
Arabic (Egyptian)		■			■	■	
Czech	■						■
Dutch					■	■	■
English	■	■	■	■	■	■	■
French		■		■	■		■
German		■		■	■	■	■
Hindi		■					
Japanese		■	■		■	■	
Korean		■					
Mandarin	■	■		■	■	■	
Persian		■			■		
Portuguese					■		■
Russian					■		■
Serbo-Croatian					■		■
Spanish	■	■	■	■	■	■	■
Taiwanese			■				
Tamil		■			■		
Thai					■		
Turkish					■		
Vietnamese		■					

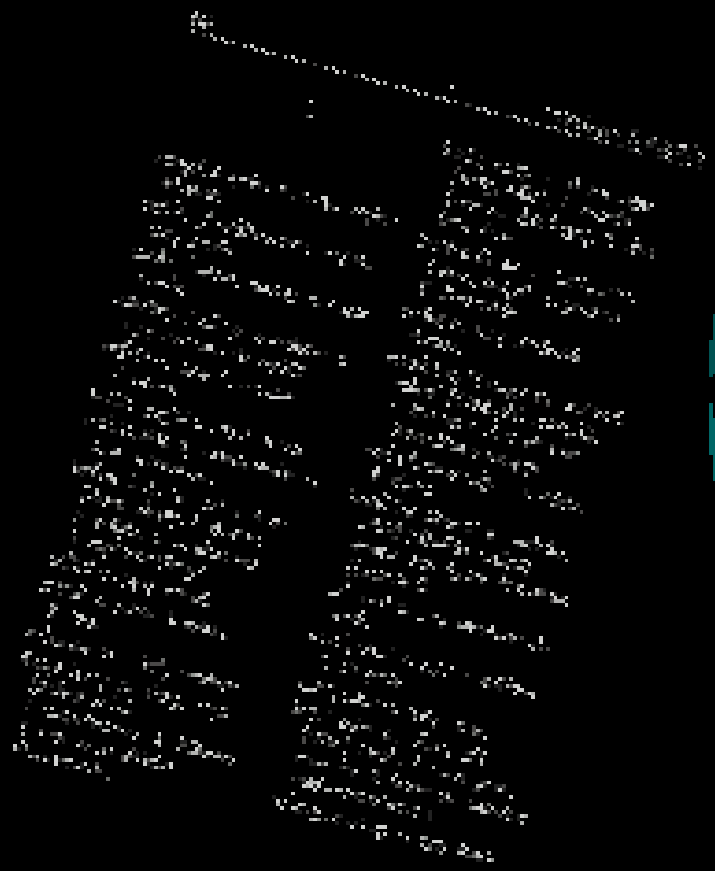
➢ **Afrikaans, Bamileke, Basque, Estonian, Hungarian, Italian, Kazakh, Kurdish, Latvian, Manding, Polish, Slovene, Ukrainian, Uzbek, Xhosa, Yoruba**

Cameroon

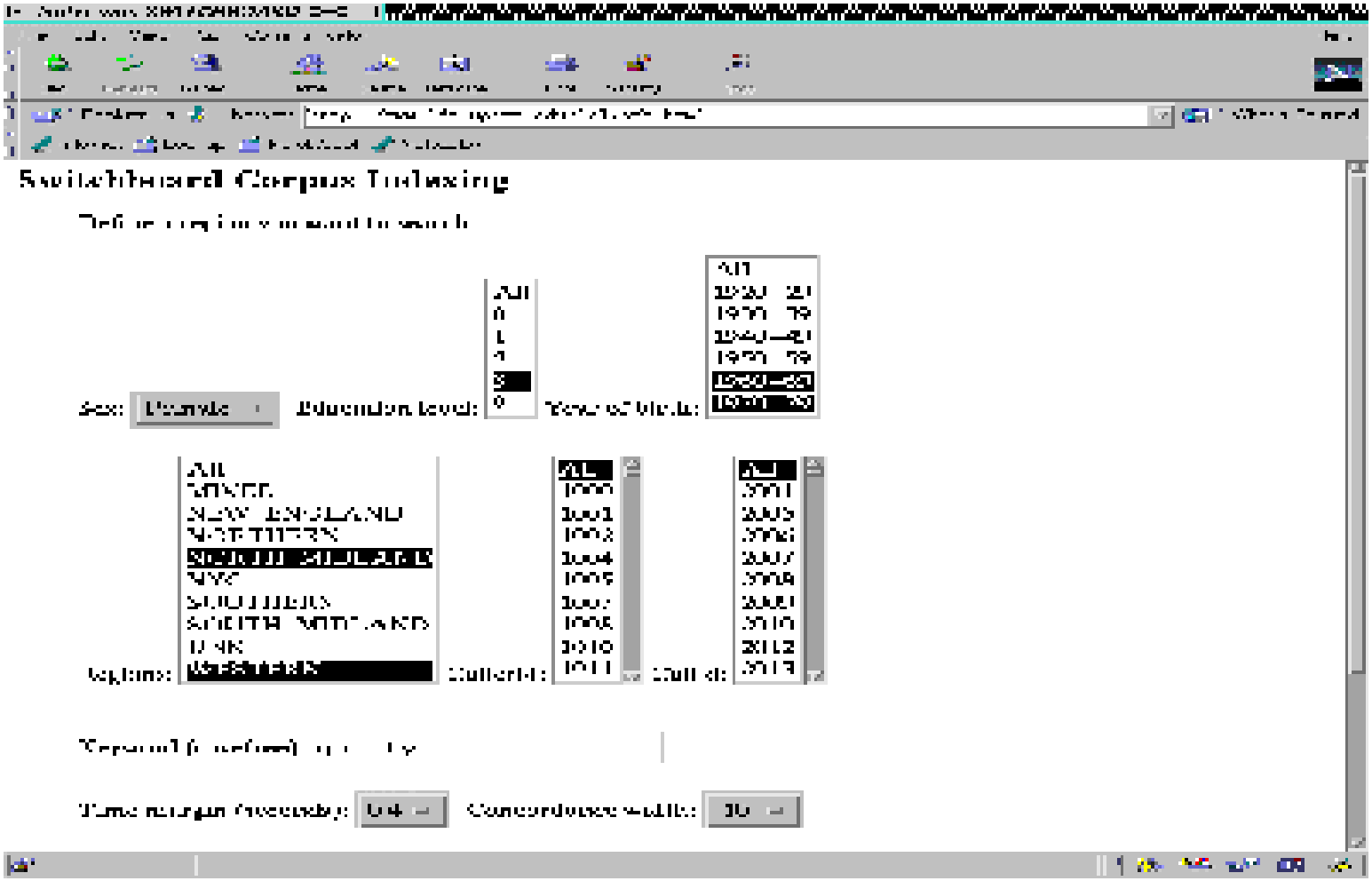




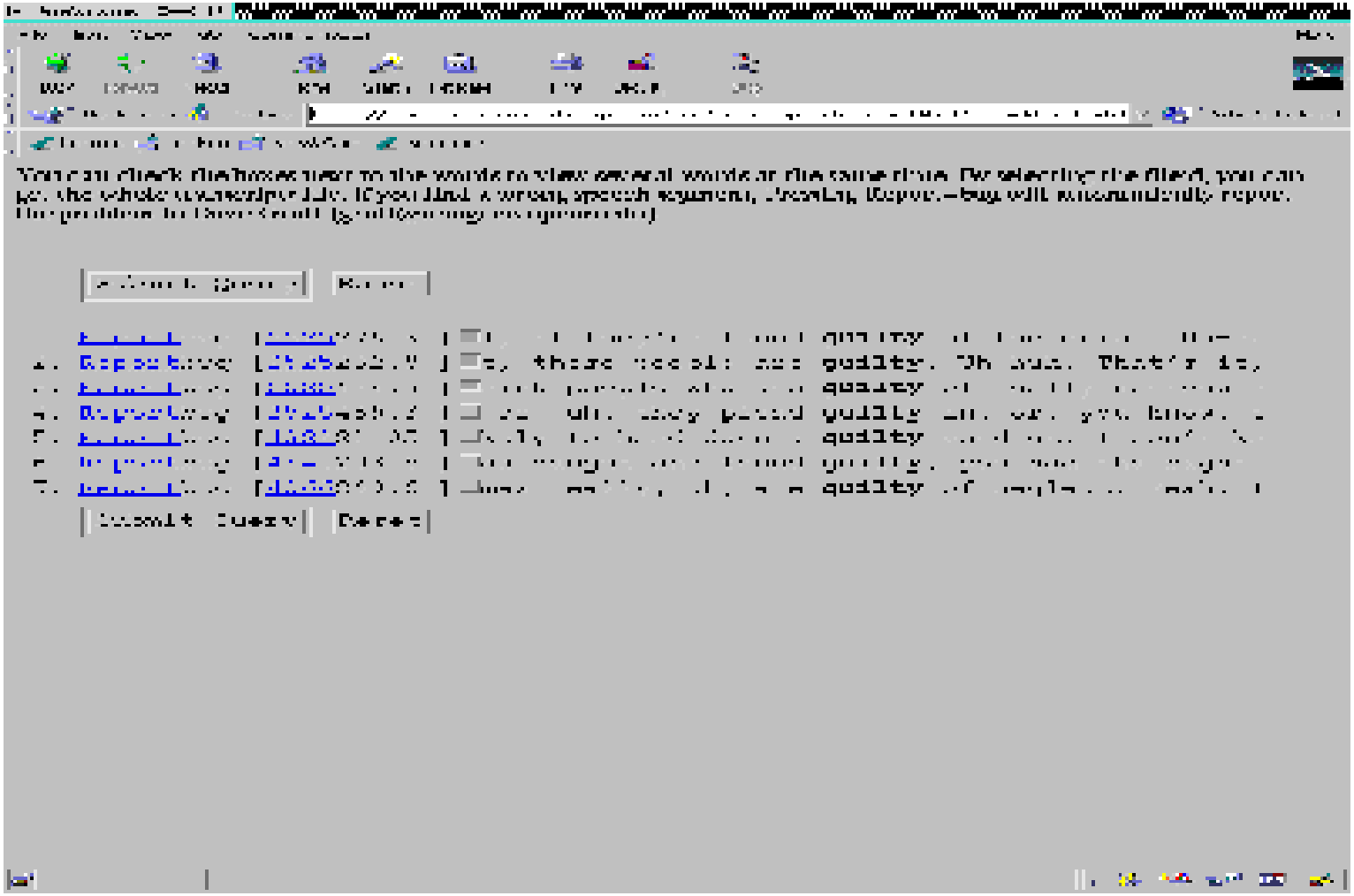
Bamileke Dictionary



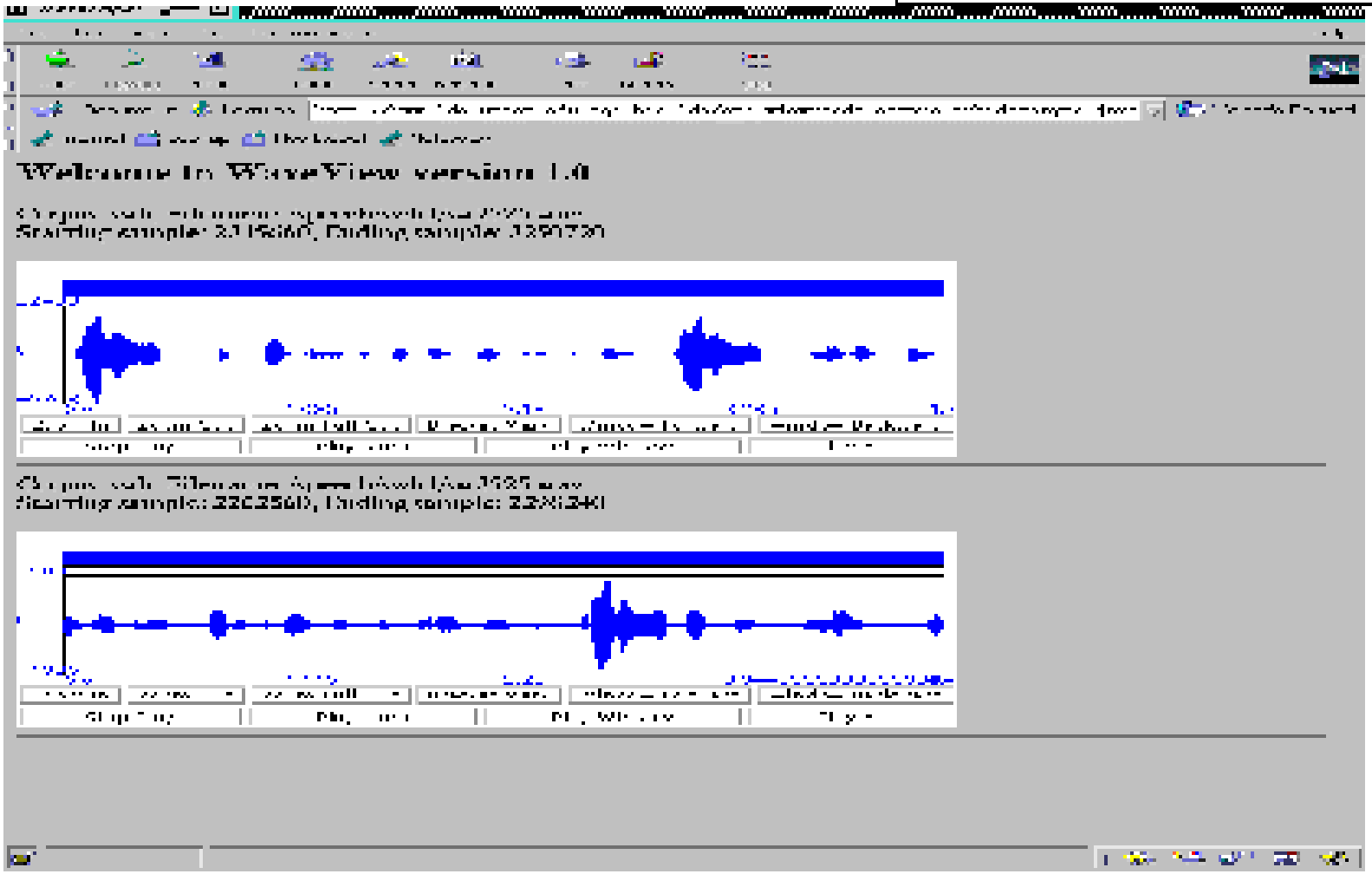
Switchboard Keyword Search



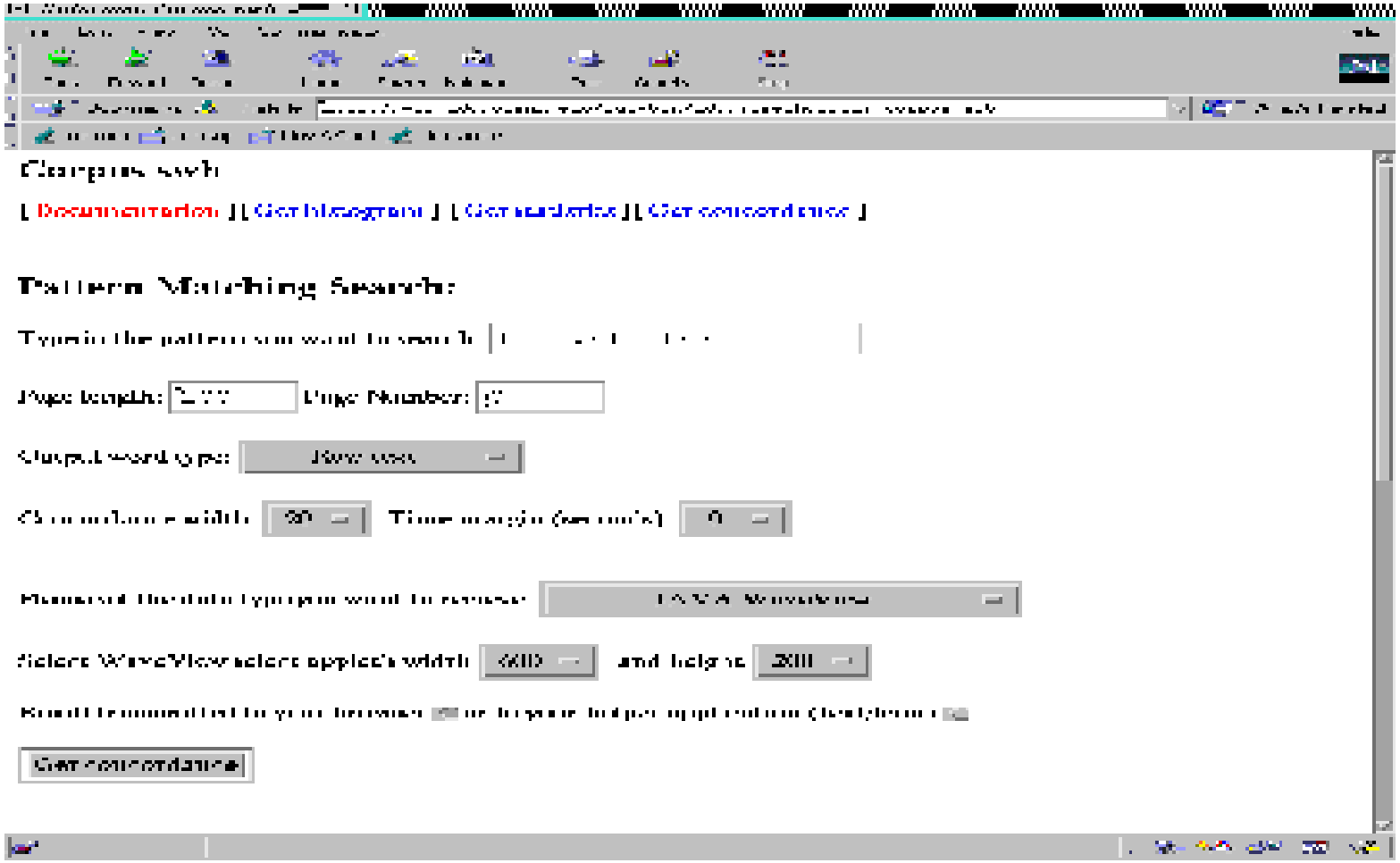
Switchboard Keyword Search: Hit List



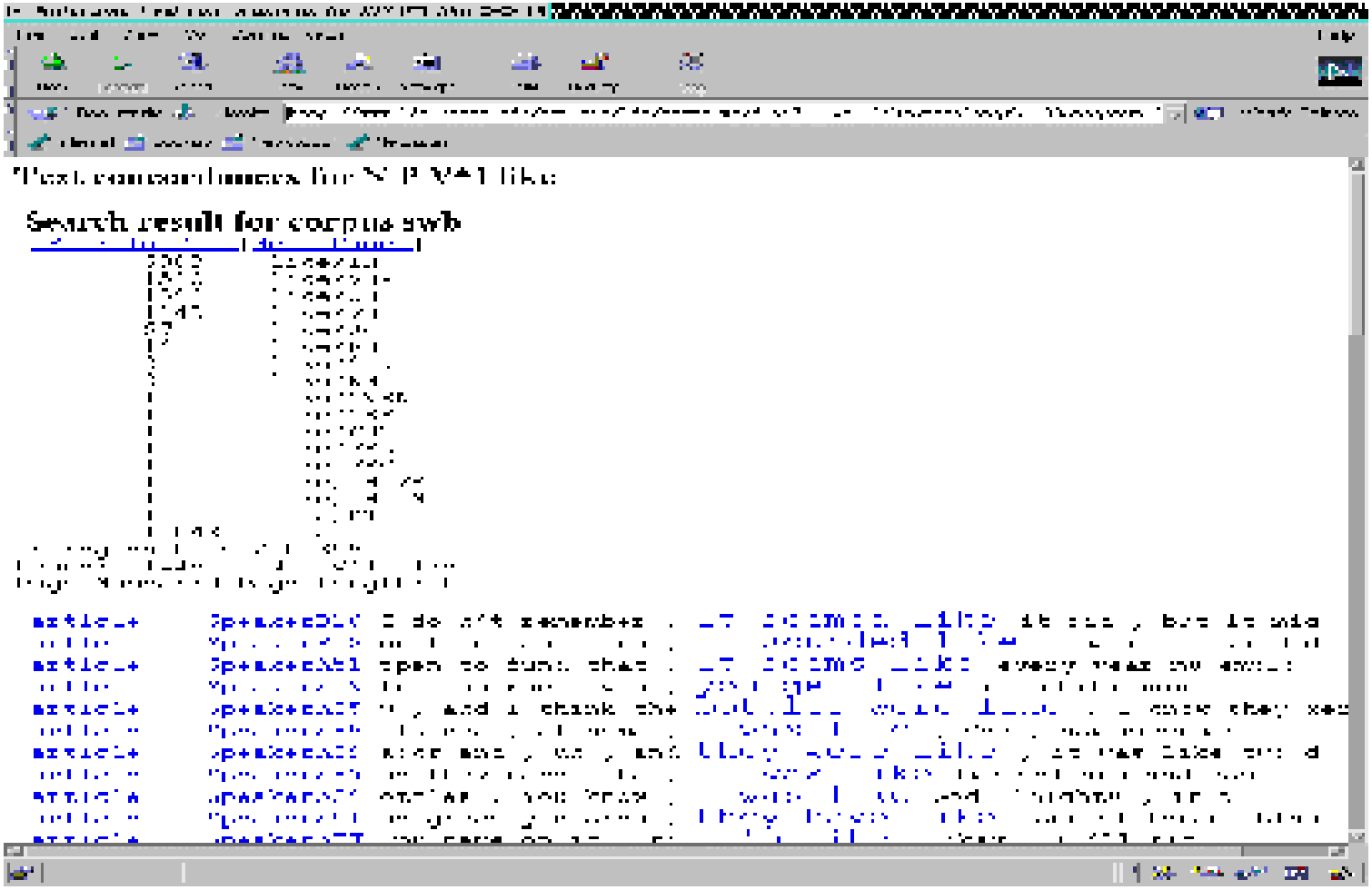
Switchboard Keyword Search Results



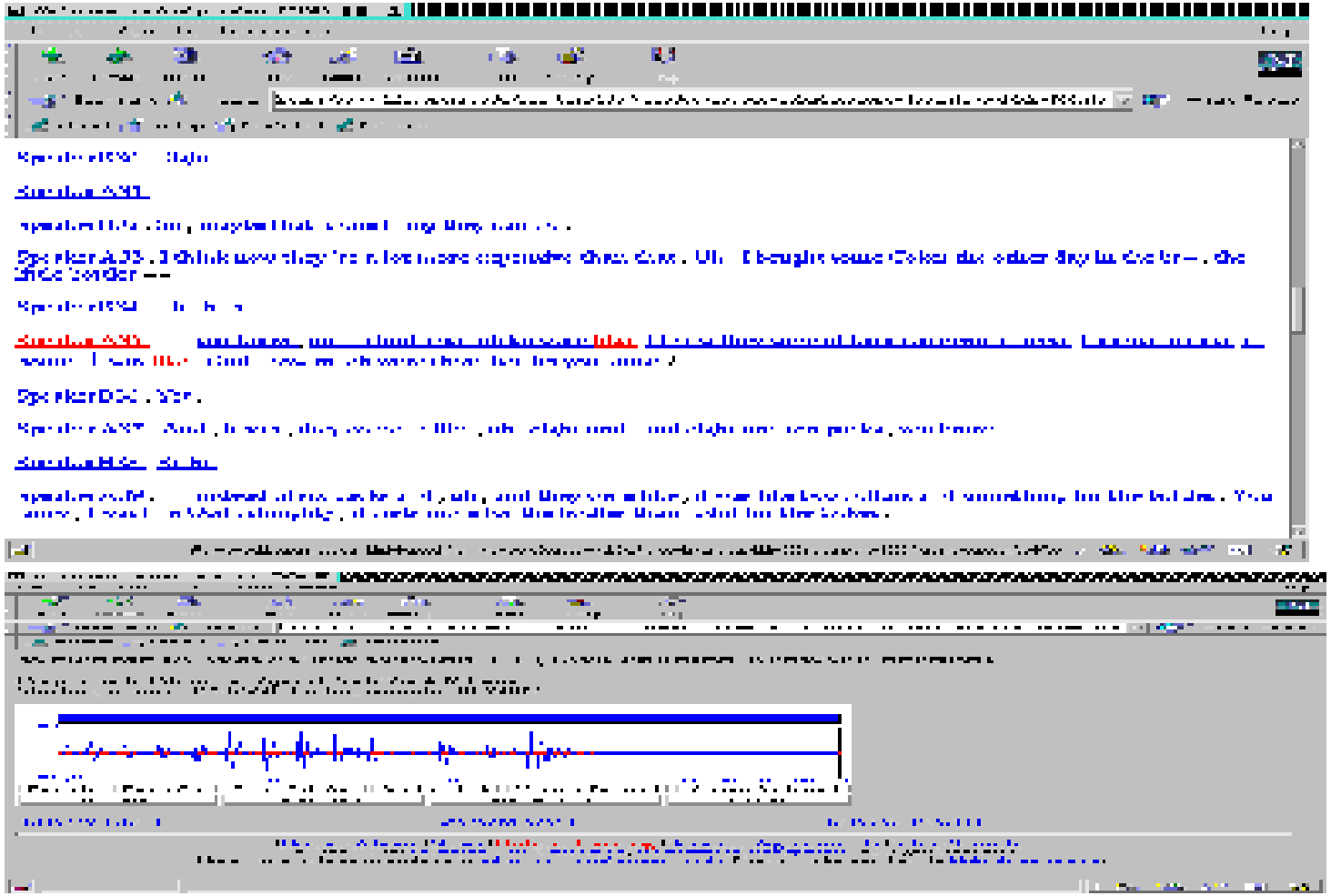
Switchboard Tagged Corpus Search



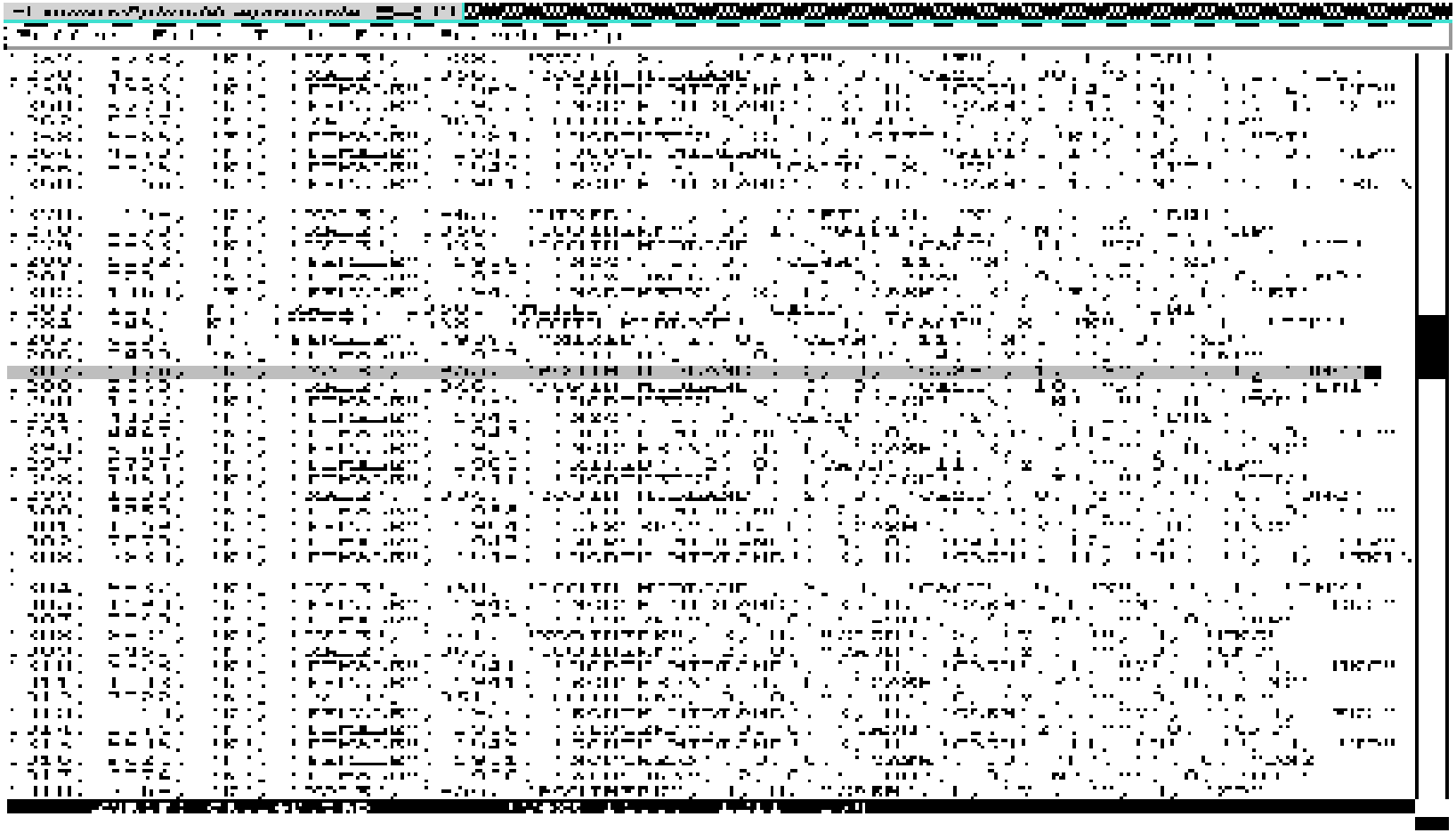
Switchboard Tagged Corpus Search: Concordance

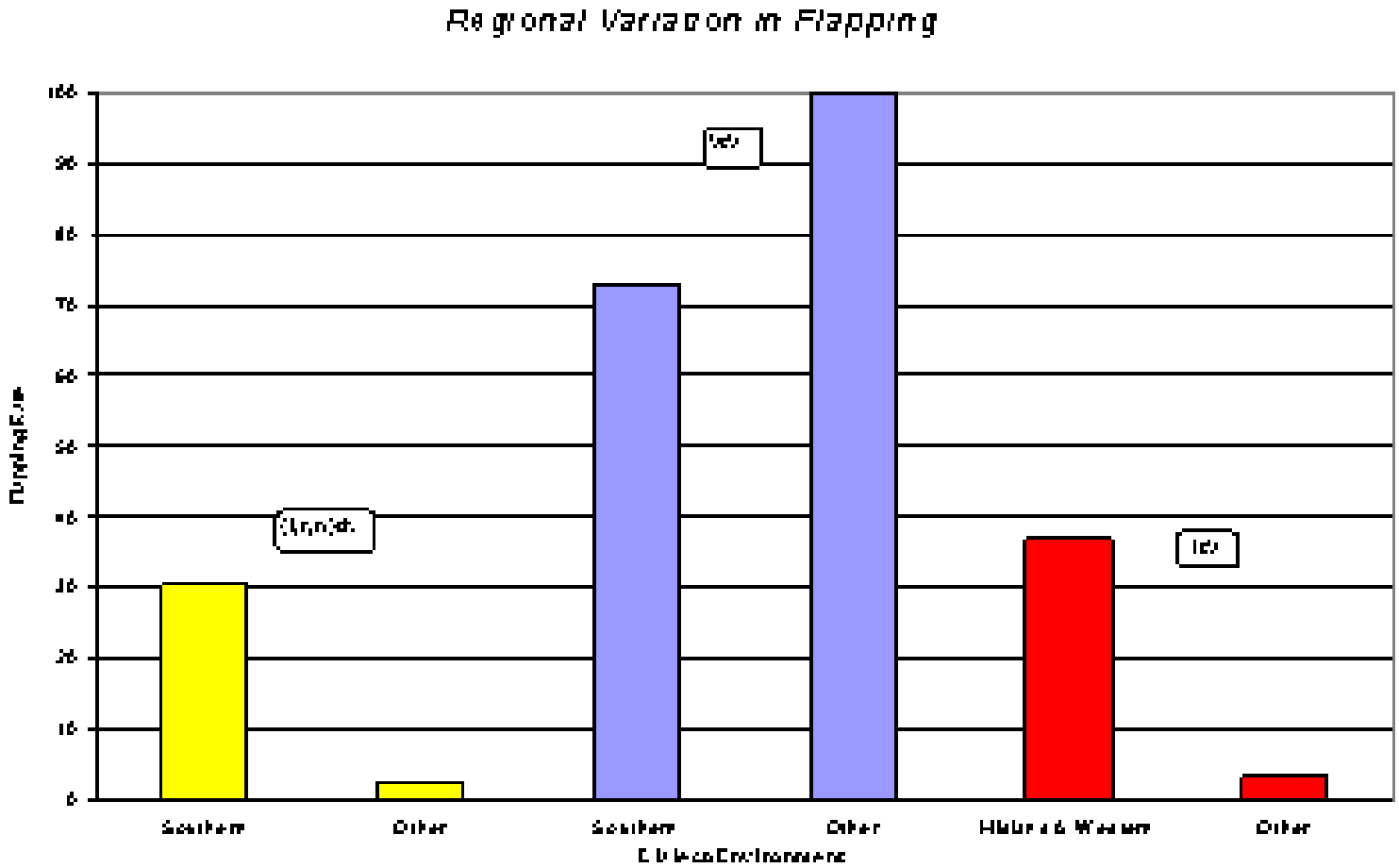


Switchboard Tagged Corpus Search: Results



Switchboard Tagged Corpus Search: Demographic Information





(from Suzuki 1998)