



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

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

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

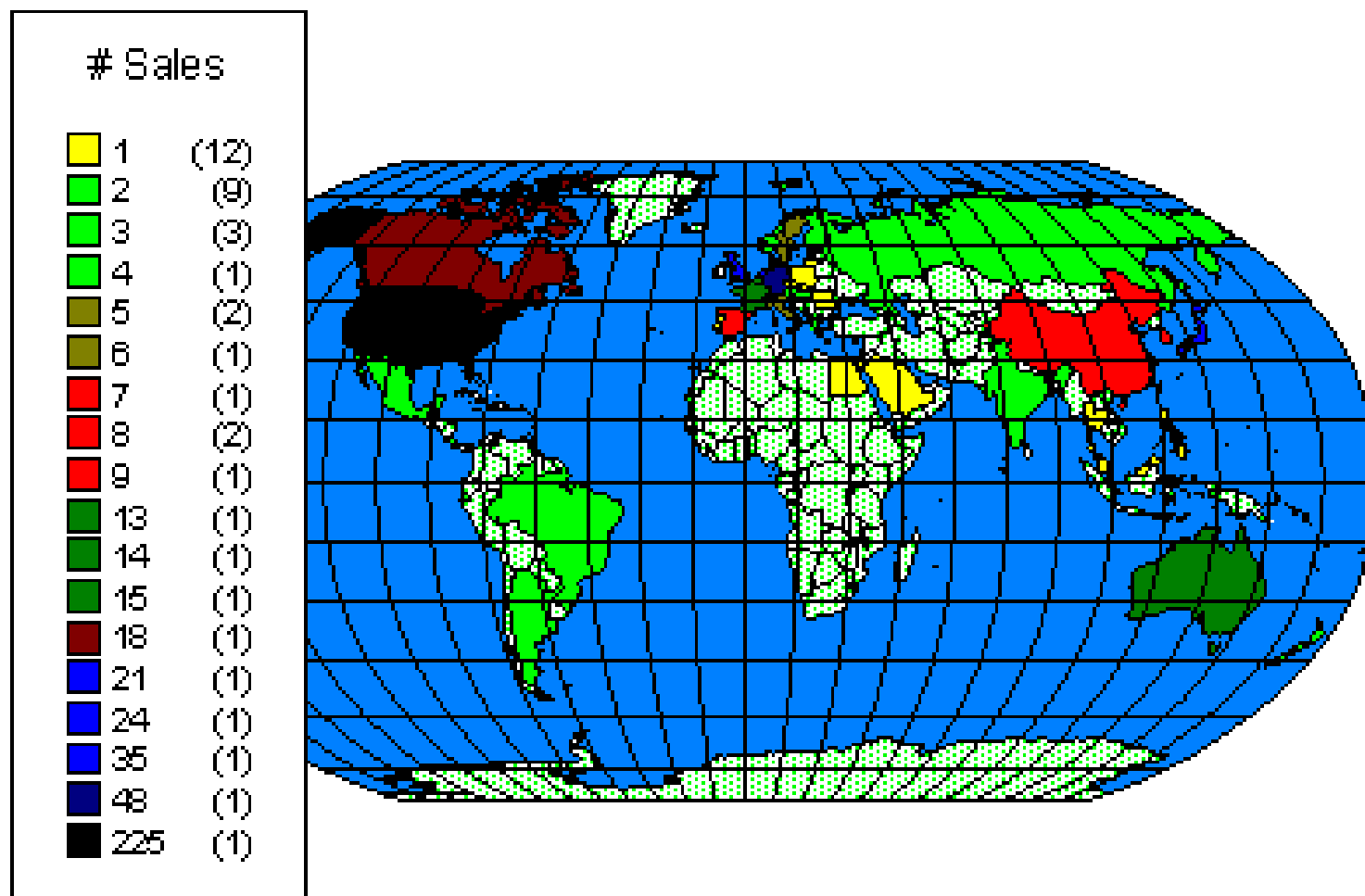
- ♦ IPR intermediary
- ♦ advice on collection, publication and IPR issues
- ♦ develop standards and tools

140 Publications + 12 in press

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

Distribution

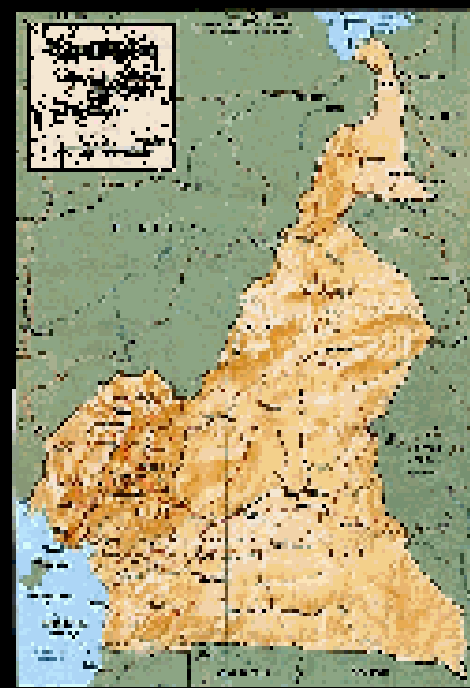
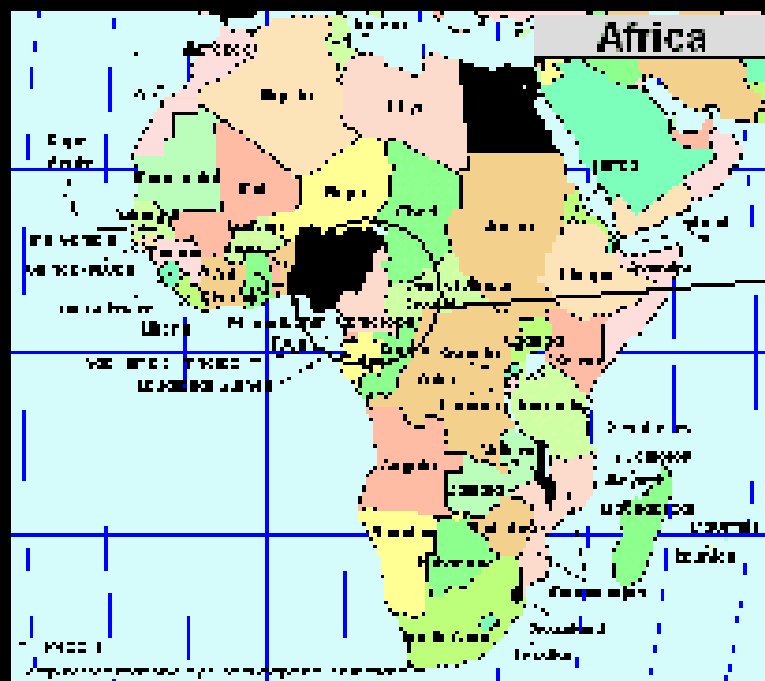
- ◆ 267 members
- ◆ 520 nonmembers



	Speech / Transcripts						
Language	Broadcast	Telephone	Wideband	Parallel Text	News wire / Other Text	Lesson	Traditional Dictionary
Arabic (Egyptian)							
Czech							
Dutch							
English							
French							
German							
Hindi							
Japanese							
Korean							
Mandarin							
Peruvian							
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





0500

Switchboard Corpus Indexing

Define a query in order to search

Sex: Education level: Year of birth:

Options:

Collection:

Collection:

Keywords (if defined)

Time range (if needed): Concordance width:

Switchboard Keyword Search Hit List

Switchboard Search

File Edit View Misc Switchboard

WSP LEXPOS HOS KTH SIMS LEXMH LTV JCLP QED

File Edit View Misc Switchboard

22

File Edit View Misc Switchboard

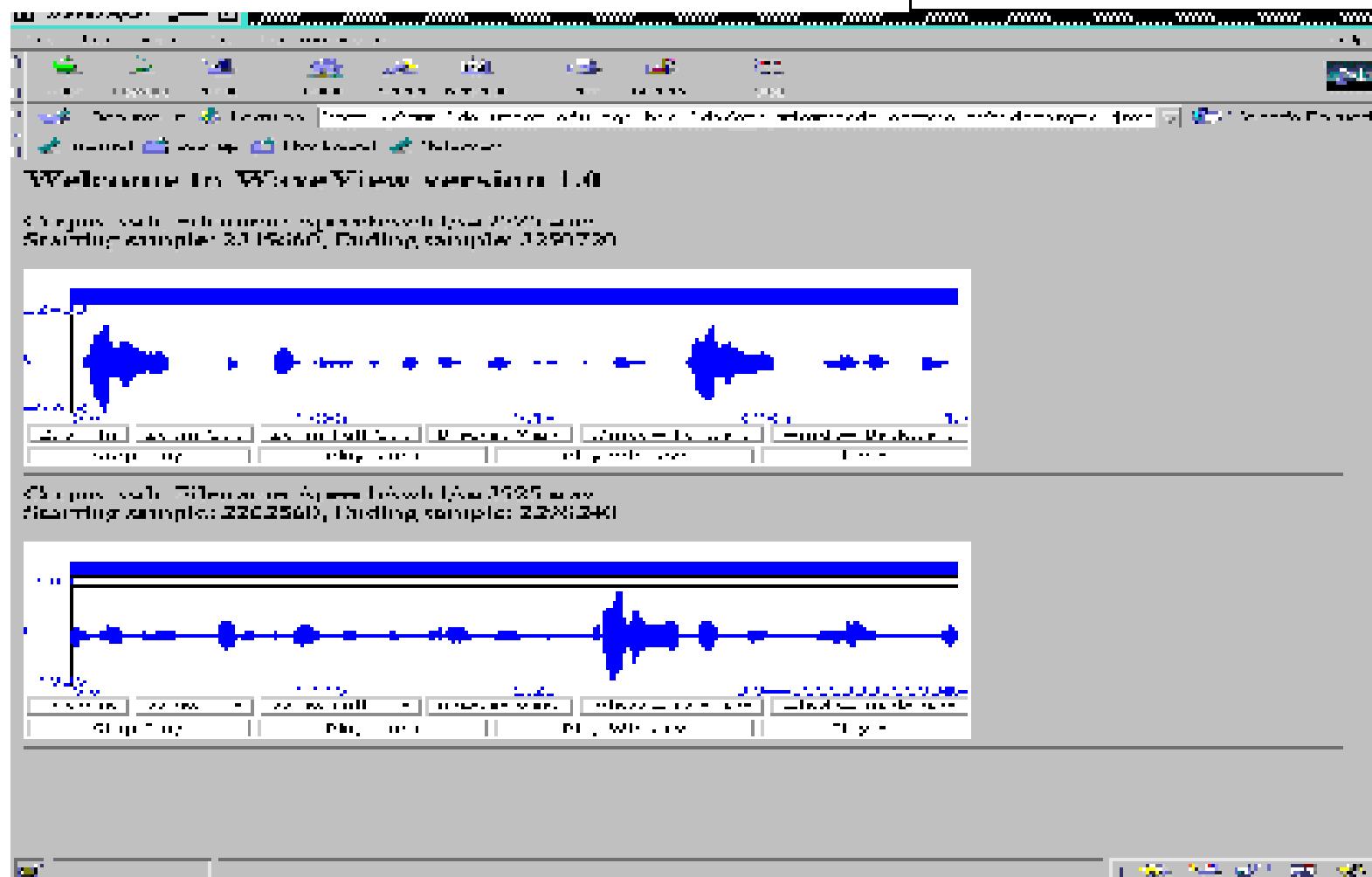
You can check the boxes next to the words to view several words at the same time. By selecting the filter, you can get the whole transcript file. If you find a wrong speech segment, Pressing Report-bug will automatically report the problem to David Smith (dsmith@cs.cmu.edu).

Submit Query Report

Word	Frequency	Filter	Text
quality	1	☐	quality of the work. That's
quality	1	☐	quality. Oh yes. That's it,
quality	1	☐	quality of the work. That's
quality	1	☐	quality and, or, you know, a
quality	1	☐	quality and, or, you know, a
quality	1	☐	quality, you know, the higher
quality	1	☐	quality of people. That's it

Submit Query Report

Switchboard Keyword Search Results







Switchboard Tagged Corpus Search Results

[illegible]



Linguistic
Development
Corpora

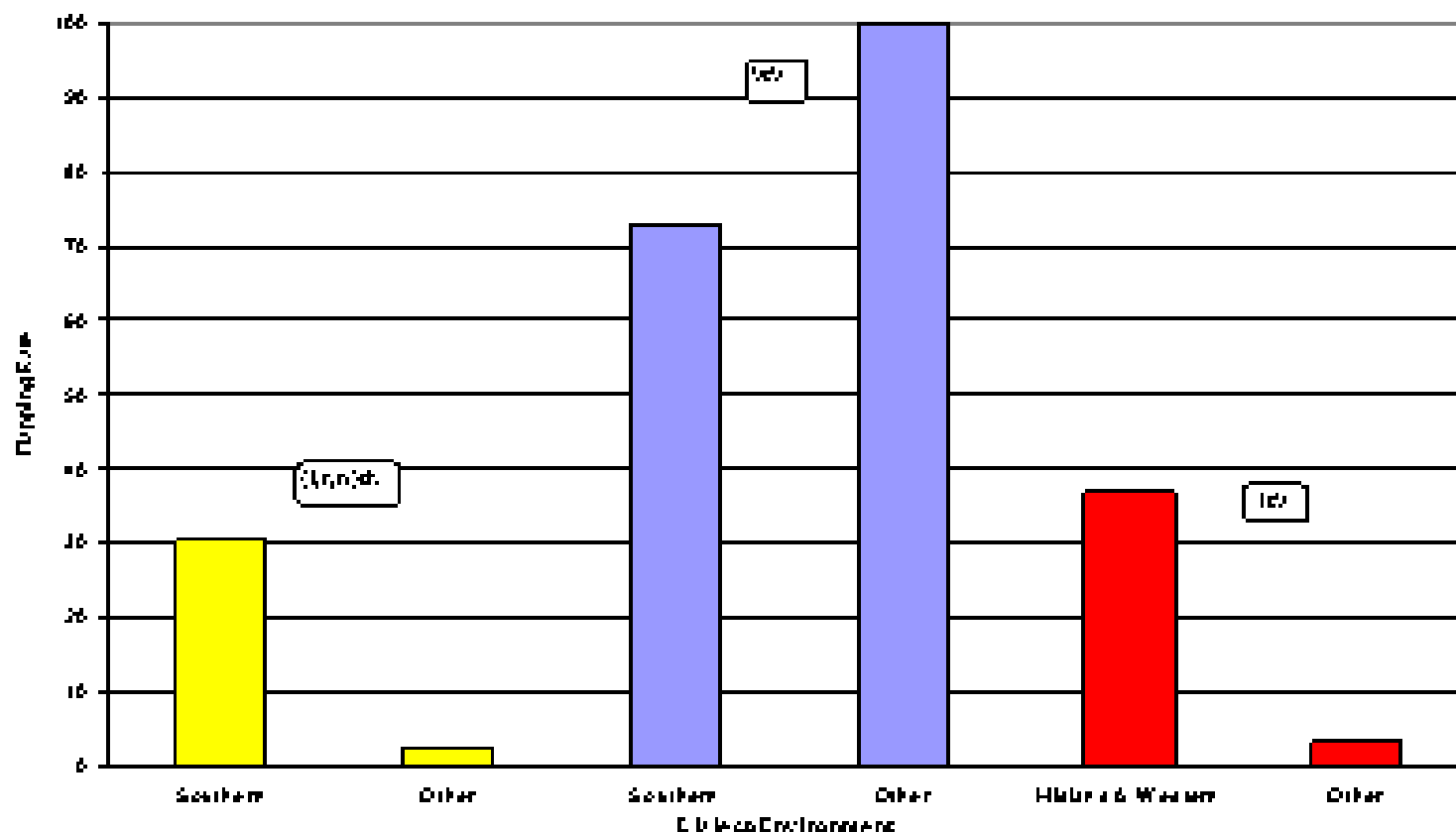
Acoustic
Phonetic
Annotations

Lexical
Syntactic
Annotations

Switchboard Tagged Corpus Search: Demographic Information

- LDC Switchboard Tagged Corpus Search: Demographic Information									
File #	File Name	Speaker	Gender	Age	Education	Occupation	Religion	Marital Status	Address
1001	100101	1001	M	30-39	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	100101
1002	100201	1002	F	40-49	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	100201
1003	100301	1003	M	50-59	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	100301
1004	100401	1004	F	60-69	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	100401
1005	100501	1005	M	70-79	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	100501
1006	100601	1006	F	80-89	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	100601
1007	100701	1007	M	90-99	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	100701
1008	100801	1008	F	100-109	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	100801
1009	100901	1009	M	110-119	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	100901
1010	101001	1010	F	120-129	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	101001
1011	101101	1011	M	130-139	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	101101
1012	101201	1012	F	140-149	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	101201
1013	101301	1013	M	150-159	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	101301
1014	101401	1014	F	160-169	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	101401
1015	101501	1015	M	170-179	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	101501
1016	101601	1016	F	180-189	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	101601
1017	101701	1017	M	190-199	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	101701
1018	101801	1018	F	200-209	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	101801
1019	101901	1019	M	210-219	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	101901
1020	102001	1020	F	220-229	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	102001
1021	102101	1021	M	230-239	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	102101
1022	102201	1022	F	240-249	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	102201
1023	102301	1023	M	250-259	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	102301
1024	102401	1024	F	260-269	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	102401
1025	102501	1025	M	270-279	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	102501
1026	102601	1026	F	280-289	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	102601
1027	102701	1027	M	290-299	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	102701
1028	102801	1028	F	300-309	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	102801
1029	102901	1029	M	310-319	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	102901
1030	103001	1030	F	320-329	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	103001
1031	103101	1031	M	330-339	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	103101
1032	103201	1032	F	340-349	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	103201
1033	103301	1033	M	350-359	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	103301
1034	103401	1034	F	360-369	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	103401
1035	103501	1035	M	370-379	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	103501
1036	103601	1036	F	380-389	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	103601
1037	103701	1037	M	390-399	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	103701
1038	103801	1038	F	400-409	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	103801
1039	103901	1039	M	410-419	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	103901
1040	104001	1040	F	420-429	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	104001
1041	104101	1041	M	430-439	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	104101
1042	104201	1042	F	440-449	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	104201
1043	104301	1043	M	450-459	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	104301
1044	104401	1044	F	460-469	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	104401
1045	104501	1045	M	470-479	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	104501
1046	104601	1046	F	480-489	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	104601
1047	104701	1047	M	490-499	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	104701
1048	104801	1048	F	500-509	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	104801
1049	104901	1049	M	510-519	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	104901
1050	105001	1050	F	520-529	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	105001
1051	105101	1051	M	530-539	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	105101
1052	105201	1052	F	540-549	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	105201
1053	105301	1053	M	550-559	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	105301
1054	105401	1054	F	560-569	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	105401
1055	105501	1055	M	570-579	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	105501
1056	105601	1056	F	580-589	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	105601
1057	105701	1057	M	590-599	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	105701
1058	105801	1058	F	600-609	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	105801
1059	105901	1059	M	610-619	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	105901
1060	106001	1060	F	620-629	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	106001
1061	106101	1061	M	630-639	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	106101
1062	106201	1062	F	640-649	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	106201
1063	106301	1063	M	650-659	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	106301
1064	106401	1064	F	660-669	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	106401
1065	106501	1065	M	670-679	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	106501
1066	106601	1066	F	680-689	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	106601
1067	106701	1067	M	690-699	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	106701
1068	106801	1068	F	700-709	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	106801
1069	106901	1069	M	710-719	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	106901
1070	107001	1070	F	720-729	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	107001
1071	107101	1071	M	730-739	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	107101
1072	107201	1072	F	740-749	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	107201
1073	107301	1073	M	750-759	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	107301
1074	107401	1074	F	760-769	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	107401
1075	107501	1075	M	770-779	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	107501
1076	107601	1076	F	780-789	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	107601
1077	107701	1077	M	790-799	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	107701
1078	107801	1078	F	800-809	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	107801
1079	107901	1079	M	810-819	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	107901
1080	108001	1080	F	820-829	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	108001
1081	108101	1081	M	830-839	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	108101
1082	108201	1082	F	840-849	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	108201
1083	108301	1083	M	850-859	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	108301
1084	108401	1084	F	860-869	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	108401
1085	108501	1085	M	870-879	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	108501
1086	108601	1086	F	880-889	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	108601
1087	108701	1087	M	890-899	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	108701
1088	108801	1088	F	900-909	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	108801
1089	108901	1089	M	910-919	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	108901
1090	109001	1090	F	920-929	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	109001
1091	109101	1091	M	930-939	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	109101
1092	109201	1092	F	940-949	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	109201
1093	109301	1093	M	950-959	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	109301
1094	109401	1094	F	960-969	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	109401
1095	109501	1095	M	970-979	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	109501
1096	109601	1096	F	980-989	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	109601
1097	109701	1097	M	990-999	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	109701
1098	109801	1098	F	1000-1009	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	109801
1099	109901	1099	M	1010-1019	HS	TELEPHONE OPERATOR	PROTESTANT	MARRIED	109901
1100	110001	1100	F	1020-1029	HS	TELEPHONE OPERATOR	CATHOLIC	MARRIED	110001

Regional Variation in Flapping



(From Sussman 1998)