SUPERB, an extension of the UCLA Phonological Inventory Database

Johan Stark
Institutionen för lingvistik
Stockholms Universitet
106 91 Stockholm, SWEDEN

ABSTRACT

UCLA Phonological Inventory Database or UPSID contains phonological segment data from a large number of the world's languages. SUPERB is a program for APOLLO workstations, which gives access to UPSID and extends it with recorded speech samples.

INTRODUCTION

UPSID was created at the UCLA Phonetics Laboratory and is described in detail in *'Patterns of Sounds'* by Ian Maddieson (1984). UPSID is also available as a data program made at UCLA for PC computers. All information described hereafter on the UPSID material has been taken from Maddieson's book, for the description of the data program and from the data base itself. UPSID was designed to give reliable answers to questions concerning the phonological segment inventories of the world's languages.

The objective of SUPERB is primarily to give access to UPSID data on APOLLO workstations and eventually to extend it to connect to the large inventory of sampled speech data, known as IRIS, which is currently beeing transferred to our local workstations from tape recordings made by Olle Engstrand at the Department of Linguistics of Uppsala University.

LANGUAGES

Referring to *'Patterns of Sounds'* the languages (currently 435) represented in UPSID were chosen 'to approximate a properly constructed quota sample on a genetic basis of the world's extant languages' (p.5). Only one language from each small family grouping has been chosen and each such small family grouping should be represented by one language. Each language is assigned a 4 digit number. The first and second digits together describe the family a language belongs to. The two last numbers are chosen to give each language a unique number.

Table 1. *The family numbering of languages.*

20xx Indo-European	27xx Other Eurasion	61xx Na-Dene
21xx Ural-Altaic	29xx Dravidian	67xx N. Amerind
23xx Austro-Asiatic	41xx Niger-Korofanian	68xx S. Amerind
24xx Austro-Tai	42xx Nilo-Saharan	69xx Eskimo-Aleut
25xx Sino-Tibetan	43xx Afro-Asiatic	83xx Australian
26xx Caucasian	49xx Khoisan	86xx Papuan

SEGMENTS and FEATURES

In UPSID every segment which is considered phonemic in a language is represented by its most characteristic allophone. Each segment type is assigned a so called character code, which is a unique ASCII string. The reason for this was to make it possible to specify any segment using an ordinary keyboard. Each segment is further specified by a combination of phonetic attributes or features. There are currently 64 features available in UPSID belonging to one of the following five groups:

1. Features indicating manner of articulation.
2. Features indicating place of articulation.
3. Features indicating secondary articulations.
4. Vowel features.
5. Features for phonation types, etc.

THE SUPERB DATABASE

The data in SUPERB is copied from the UPSID data base and reformatted in order to suit the SUPERB system. No effort has been made to reorganize or rename any UPSID data items. SUPERB is build using Sicstus Prolog which means that all data are represented by ASCII text files. The SUPERB data base is build as a relational database. Data are consequently stored in tables. The following five tables are defined:

1. Features	(from UPSID)
2. Languages	(from UPSID)
3. Character Codes	(from UPSID)
4. Segments	(from UPSID)
5. Samples	(reference to IRIS data)

Features table

The features table has five fields:
a/ Order number (1-64)
b/ Abbreviated name
c/ A description text string
d/ Group number (1-5, see section FEATURES above)

Languages table

The language table has three fields:
a/ Language name
b/ Language number (see LANGUAGES section above)
c/ Number of used segments

Character code table

The character code table has three fields:
a/ Character code (see section SEGMENTS above)
b/ A list of feature order numbers for the features included in this segment type
c/ Number of languages using this segment type

Segments table
The segments table has three fields:
a/ Language number (indicating what language the segment refers to)
b/ Character code (indicating the segment type)
c/ Anomaly number (0-5 see below)

The anomaly number is interpreted as follows:
0 normal segment
1 extremely rare
2 only in foreign words
3 only in underlying forms to account for some phonological patternings
4 phonemic in UPSID but derived from other underlying segments
5 described in a obscure or contradictory fashion

Samples table
The only non UPSID table is the samples table having 6 fields:
a/ Language number (indicating what language the segment refers to)
b/ Character code (indicating the segment type)
c/ File name of sample data file
d/ Start position (in file) in seconds
e/ End position in seconds
f/ File name of optional additional data

The relations
The other characteristic of a relational database is its defined relations between tables. The relations in SUPERB are:
1. The 2nd field of 'languages' is referenced by the 1st field of 'segments'.
2. The 1st field of 'charcter code' is referenced by the 2nd field of 'segments'.
3. The 2nd field of 'character code' refererences one or several features by its 1st field.
4. The 1st and 2nd fields of samples refererence the corresponding fields of 'segments'.

Relations 1 and 2 are of the type One-to-Many, but together there is a Many-to-Many relation between segment types (character codes) and languages. Relation 3 is a true Many-to-Many relation since the 2nd field of 'charcter code' is a list of feature numbers.

The queries
Several queries for subsets from one or several tables' data may be defined. The output is controlled by a 'method' which is a definition of how to proceed with the desired data. The output may be plain text but it can also be some processing of sampled data from a corresponding sample file: audio play back, spectrograms etc.

The queries can be made using several techniques. The simplest level is to choose from a menu of predefined query templates and only change the arguments to what ever is needed to complete the correct query. Since SUPERB is a Prolog program it is possible, while running the program, to define new queries or even extend it with new query predicates and save in a file. This requires the user to be familiar with prolog syntax.

Table 2. *Predefined query templates.*

COUNTING PREDICATES	INSPECTION PREDICATES
number_of_charcods.	segments_of_language(Language).
number_of_segments.	anomaly_of_language(Language).
number_of_segments(Language).	language_info(Language).
number_of_occurences(Charcod).	features_of_segment(Charcod).

SELECTION PREDICATES
segments_by_features(Feature_condition).
languages_by_features(Feature_condition).
languages_by_features(Range,Value,Feature_condition).
languages_by_segments(Charcod_condition).
classes_by_size(Class_size).

DATABASE PREDICATES	STATISTICAL ANALYSIS
charcod(Charcod,Features,Occurences).	use_of_features(Feature_condition,Size).
feature(No,Feature,Description,Group).	total_of_features(Feature_condition,Size).
language(Language,L_number,Size).	
segment(L_number,Charcod,Anomaly).	

Editing
Editing the data tables is very easy since all data is ASCII text. This means that any text editor can be used, and that transferring the data to other formats is very easy.

The user interface
The user interface is text oriented which means that all commands, queries results etc. are presented as text. However, since SUPERB is implemented on an APOLLO workstation a more user friendly interface is available just by setting up the user's environment properly. This gives the user the possibility to pick templates of queries by using a mouse and then edit them with the exact arguments wanted. If the same argument will be used several times the whole query, including arguments, may be saved in the menu window and chosen by just a click of the mouse next time. However, since it is text oriented, the program may also be run remotely from a PC, terminal or workstation using a telephone modem, internet connection or the like.

ACKNOWLEDGEMENTS
This work was supported by grants from the National Science Foundation (BNS-9011894) and the HSFR in Sweden (HSFR, grant F 149/91).

REFERENCES
Carlsson, M. et.al. 1991. *SICStus Prolog User's Manual.* Stockholm: SICS.
Lindblom, B., Krull, D. & Stark J. 1993. *Phonetic Systems and Phonological Development,* in B. de Boysson-Bardies et al. (eds), Developmental Neurocognition: Speech and Face Processing in the First Year of Life, pp. 399-409
Maddieson, I. 1984 *Patterns of Sounds.* Cambridge: Cambridge University Press.
Precoda, K. 1989. *Upsid source code.* UCLA.