South Asian Languages in Multilingual TTS-Related Database

Ksenia Shalonova, Roger Tucker
Mobile and Media Systems Laboratory
HP Laboratories Bristol

HPL-2003-9
January 28th , 2003*

E-mail: ksenia_shalonova@hplb.hpl.hp.com, roger_cf_tucker@yahoo.co.uk

text-to-speech technology

In this paper we overview possible problems to be overcome in building TTS systems for different languages, in particular for the languages of South Asia. We do this by an analysis of both script and language features (presented in the Multilingual TTS-Related Database), and observe that all languages (not just in South Asia) have a limited number of these features. We briefly describe possible TTS problems in accordance with the described script and language categories (several examples from South Asian languages are provided). By attributing scores to the features, we can rank the languages in order of difficulty. On this basis, Bengali is one of the easiest languages and Pashto is one of the most difficult.

* Internal Accession Date Only

Approved for External Publication

© Copyright Hewlett-Packard Company 2003

South Asian Languages in Multilingual TTS-Related Database

Ksenia Shalonova
Mobile and Media Systems
HP Labs
ksenia_shalonova@hplb.hpl.hp.com

Roger Tucker
Mobile and Media Systems
HP Labs
roger_cf_tucker@yahoo.co.uk

Abstract

In this paper we overview possible problems to be overcome in building TTS systems for different languages, in particular for the languages of South Asia. We do this by an analysis of both script and language features (presented in the Multilingual TTS-Related Database), and observe that all languages (not just in South Asia) have a limited number of these features. We briefly describe possible TTS problems in accordance with the described script and language categories (several examples from South Asian languages are provided). By attributing scores to the features, we can rank the languages in order of difficulty. On this basis, Bengali is one of the easiest languages and Pashto is one of the most difficult.

1 Introduction

Most TTS engines nowadays are mainly developed and tested for commercially profitable languages like English, French, German, Spanish, Japanese, Chinese and etc. Although there are a number of commercial companies and research laboratories that aim to base their TTS technologies on language-independent engines (Dutoit, 1997; *Multilingual Text-to-Speech Synthesis*, 1998) it seems important to obtain a formalised representation of TTS problems and their solutions for all possible languages that may require a commercial TTS development. Currently such a structured representation of languages in application to TTS development contains 105 languages. As the criterion we have chosen languages, in which official newspapers are published (see section 2 – **Multilingual TTS-related linguistic database**).

As all languages have a limited number both of linguistic and script features, there are a limited number of possible TTS problems. Their solutions can be obtained by means of re-using TTS components from one language into another one. In order to create a TTS system for a particular language it is best to understand from the beginning the language-dependent and language-independent TTS problems that have to be solved during TTS development. To avoid unnecessary effort in the development cycle it is useful to predict most of possible problems/solutions from the start. That is why it seems extremely important to work on Multilingual Transfer – re-use of modules from existing languages and also to develop techniques based on (semi-) automatic tools in order to solve the problems in an integrated way.

The current language set in the Multilingual TTS-Related Database includes major South Asian languages such as Assamese, Bengali, Gujarati, Hindi, Kannada, Kashmiri, Malayalam, Manipuri, Marathi, Nepali, Oriya, Punjabi, Pashta, Tibetan, Sindhi, Sinhala, Tamil, Telugu, Urdu – i.e. the official languages spoken in Bangladesh, Bhutan, India, Maldives, Nepal, Pakistan and Sri Lanka. From the TTS development point of view these languages combine a various number of TTS problems and technologies as they differ much both in linguistic characteristics and scripts. The language difference leads to different effort needed to create a TTS system that can be illustrated by means of our TTS-complexity language scoring system (see section 3 – **Evaluation of language complexity in application to TTS development cycle**).

2 Multilingual TTS-Related Database

2.1 Language Definition from the TTS Development Angle

TTS systems convert a limited number of graphemes into an unlimited number of speech realisations. At the same time the generation of speech realisations (grapheme-to-sound conversion rules) is based on a limited set of features. All world languages seem to have a limited number of such features to be taken into account during TTS development. These features can be subdivided into 2 main categories.

1. Linguistic features that are considered phonological/phonetic, morphological and syntactic/prosodic language peculiarities. For a number of languages some information about these linguistic levels is now available from *the Rosetta Project*.
2. Script features. There are 33 scripts in which newspapers are published (Nakanishi, 1998).

Thus, from the TTS development point of view we consider any particular **language** to be equivalent to the formalised representation of **phonological, morphological and syntactic level**[1] *plus* **script characteristics**.

Both the script and pure linguistic features for 105 languages are formalised in the Multilingual TTS-Related Database.

2.2 The Aim of the Multi-Lingual TTS Related Database

The main purpose of the Multi-lingual TTS-Related DB is to supply the following information.

1. The list of TTS problems and their possible solutions for a particular language/group of languages, for particular scripts, for languages spoken in particular countries etc. The search criteria for possible TTS problems can be composed in a query. It is also believed that on the basis of the current DB in future it will be possible for each language to prioritise the tasks to be fulfilled during the TTS development cycle.
2. Technologies that can be used in multilingual transfer, i.e. languages with similar TTS development problems and their solutions. On the basis of the structural representation of graphological/phonological, morphological and syntactical/prosodic levels for the examined languages it seems possible to obtain 2 types of information to be used in multilingual transfer:

- groups of language features and NLP techniques that can be at once applied for speech synthesis of several languages. This information can be used for adapting one TTS engine for different languages, and, therefore, for multilingual TTS transfer. For example, differentiation in GRAPHEME_TO_PHONEME _CONVERTER of stressed/unstressed and pre-stressed/post-stressed vowels in case of vowel reduction; implementation of a finite state automaton in MORPHOLOGICAL_DECOMPOSITION for highly inflective (Slavic, Baltic) and agglutinative (Turkish, Finish, Dravidian) languages; using of one algorithm to solve the problem of "consonantal" alphabets (Arabic, Hebrew), where the marking of vowels is optional etc.
- language universals – language features and NLP techniques to be applied for speech synthesis independently of language system, e.g. nasalization of vowels after nasals, algorithms for e-mails/URL processing etc[2].

2.3 The Structure of the Multi-Lingual TTS Related Database

The database includes 3 related tables with the following fields.

I. LANGUAGES
General information
LANG_NAME – the official name of the language[3] (other alternative names are given in the brackets).
LANGUAGE_HIERARCHY – the family and (sub)groups to which a language belongs.
NUMBER_OF_SPEAKERS[4] – number of speakers of a language.

[1] Semantic and discourse levels are not covered in this paper although it is an important subject for the investigation.

[2] It is important to notice that a great deal of language-independent rules are connected with phonetic/acoustic regularities, e.g. realisation of longer phonetic components in a position before a phrase boundary; realisation of higher F0 values after a voiceless consonant than after a voiced one; nasalization of vowels after nasal consonants etc.

[3] The *Ethnologue* (2000) catalogue was taken as the reference one.

COUNTRY – a list of countries where the language is spoken.
STATE – states or other geographical units within a country where the language is spoken.

Phonetic characteristics

TONES – is filled in for the tone languages (2 main types are included into the default value list: pitch-accented or tonal languages).
LEXICAL_STRESS –

- Type of the lexical stress. Three main types are included into the default value list: lexical, fixed and quantity-sensitive. For the last type of stress, where the position of primary stress is influenced by a syllable weight, the information about the identification of a heavy syllable is included. The main source of information for different stress types is the typological database – *StressTyp* developed at the University of Leiden (Goedemans et al., 1996).
- Another important feature to be introduced into this field is the acoustic realisation of stress, that for a great number of languages is very poorly described in phonetic literature[5].

SECONDARY_STRESS_OR_RHYTHM: rules for secondary or rhythm stress assignment. In most European languages not-primary (or secondary) stress can occur in clitics and compound words. Another type of not-primary stress is rhythm, which is sensitive to the place of the main stress (e.g. in Garawa, Seri etc).
INTONATION_PATTERNS – the number and description of intonation patterns.[6]
OTHER_PHONETIC_CHARACTERISTICS – phonetic phenomena not described in the previous fields, e.g. vowel reduction, palatalization, vowel harmony etc.

MORPHO-SYNTACTIC_ CHARACTERISTICS – Two main morphological types are distinguished: analytical and synthetic languages, where synthetic is subdivided into agglutinative and inflective.
MORPHOLOGICAL_CHARACTERISTICS (word-formation) – information about word-formation is included (affixation, reduplication etc.).
PROPER_SYNTACTIC_ CHARACTERISTICS – word order (fixed; free; grammatically significant).[7]
OTHER_CHARACTERISTICS – language peculiarities not included into previous fields.
LITERATURE_LANG – the literature sources found for a particular language[8].

II. SCRIPTS

General information

SCRIPT_NAME – name of the script
SCRIPT_TYPE – type of the script (alphabetic, alphabetic-syllabic, consonantal, ideographic).

CAPITALISATION – specific rules for capitalisation, e.g. no capital letters at all or capital letters used for all nouns (as in German).
GRAPHEME_TO_PHONEME_
CORRESPONDENCE – the correspondence between graphical or transliteration and phoneme levels: direct (e.g. Finnish, Spanish), direct with certain exceptions (e.g. Russian), not direct (e.g. English), not direct with optional vowel marking (e.g. Arabic). It is necessary to notice that G2P correspondence within one script can differ from language to language, e.g. in Latin and Cyrillic scripts. For such languages G2P correspondence is described in the field OTHER_CHARACTERISTICS in the table "Languages".
SYMBOLS_FOR_LOAN_WORDS – whether the script contains symbols used only in loan words.
SYMBOLS_FOR_STRESS – whether the script contains a special mark for lexical stress.
SYMBOLS_FOR_TONES – whether the script contains a special mark for tones.
PUNCTUATION_MARKS – description of punctuation marks (where they differ from those used in most European countries).
SPACES_BETWEEN_WORDS – whether the words are separated by spaces.
HOMOGRAPHS – whether homographs exist in a particular script.
COMMENTS_TO_THE FIELDS – contains remarks (explanations) for the script peculiarities described in previous fields.
OTHER_PECULIAR_CHARACTERISTICS – contains script information another than that presented in

[4] Due to obvious reasons the number of speakers may differ from one literature source to another. We have chosen one source item as the reference one (Dalby, 1999).

[5] Acoustic correlates of stress can be: pitch change, temporal structure and intensity. Changes in temporal structure (duration decreasing) may cause reduced vowels in unstressed syllables.

[6] As obtaining data for this kind of linguistic feature normally requires deep phonetic-acoustic research, full information is expected to be available only for certain languages – mainly for all European languages, some Asian languages – Japanese and Chinese.

[7] It is important to point out that there is a close connection between the syntactic characteristics and the realisations of different intonation patterns.

[8] This field is important for languages either with few sources or with sources containing contradictory information.

the main fields, e.g. differentiation of initial, middle final and isolated graphemes in Arabic script.

LITERATURE_SCRIPTS – the literature sources found for a particular script. The main reference material is taken from (Daniels and Bright., 1996; Nakanishi, 1998).

III. TTS PROBLEMS

Below we present the set of the TTS modules with the corresponding script and language features to be taken into account.

Fields in the table ***TTS problems*** (correspond to TTS modules)	Fields from the table ***Languages*** that may serve as cues for solving corresponding ***TTS problems***	Fields from the table ***Scripts*** that may serve as cues for corresponding ***TTS problems***

SENTENCE_EXTRACTION		PUNCTUATION_MARKS;
		CAPITALISATION;
		SPACES_BETWEEN_WORDS

TOKENISATION (Word extraction)		PUNCTUATION_MARKS;
		CAPITALISATION;
		SPACES_BETWEEN_WORDS

CLITICALISATION		SPACES_BETWEEN_WORDS
PROPER_NAMES_PROCESSING		CAPITALISATION
ACRONYM_PROCESSING		CAPITALISATION

ABBREVIATION_PROCESSING		PUNCTUATION_MARKS;
		CAPITALISATION;
		SPACES_BETWEEN_WORDS;
		OTHER_CHARACTERISTICS

SPECIAL_SYMBOLS_PROCESSING		OTHER_CHARACTERISTICS
E-MAILS /URLs_PROCESSING		OTHER_CHARACTERISTICS
DIGITS_PROCESSING	MORPHO-SYNTACTIC_CHARACTERISTICS	OTHER_CHARACTERISTICS
LOAN_WORDS_PROCESSING		SYMBOLS_FOR_LOAN_WORDS

STRESS (TONE) ASSIGNMENT	LEXICAL STRESS;
	SECONDARY_STRESS_OR_RHYTHM;
	MORPHO-SYNTACTIC _CHARACTERISTICS;
	MORPHOLOGICAL CHARACTERISTICS
		SYMBOLS_FOR_STRESS;
		SYMBOLS_FOR_TONES

MORPHOLOGICAL_DECOMPOSITION	MORPHO-SYNTACTIC _CHARACTERISTICS;
	MORPHOLOGICAL CHARACTERISTICS
	LEXICAL STRESS;
	OTHER_CHARACTERISTICS	

HOMOGRAPH_DISAMBIGUATION	MORPHO-SYNTACTIC _CHARACTERISTICS; LEXICAL_STRESS	HOMOGRAPHS

PHRASING	MORPHO-SYNTACTIC _CHARACTERISTICS;
	PROPER_SYNTACTIC_CHARACTERISTICS;
	INTONATION_PATTERNS
		PUNCTUATION_MARKS;
		CAPITALISATION;
		SPACES_BETWEEN_WORDS;
		OTHER_PECULIAR_CHARACTERISTICS

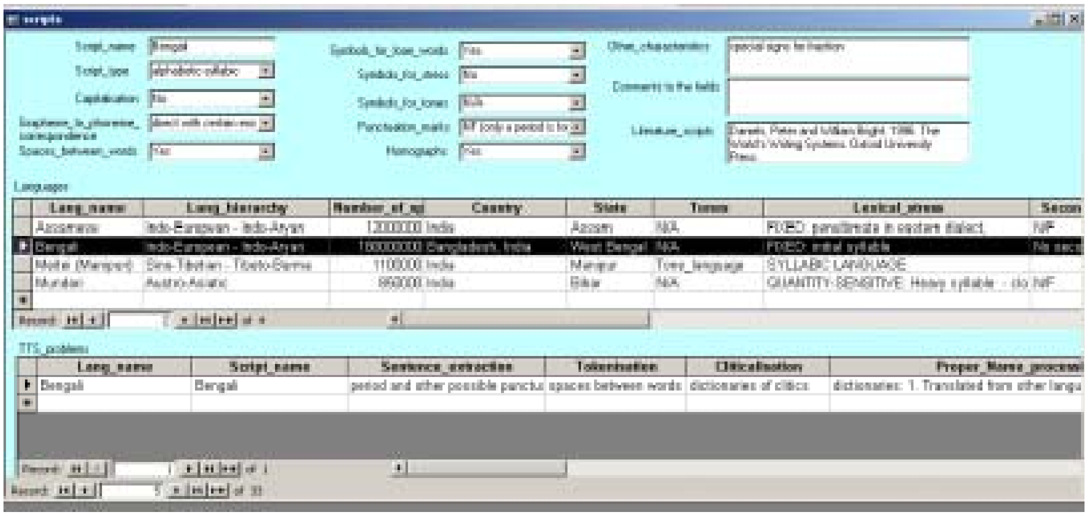
G2P_CONVERTION	MORPHO-SYNTACTIC _CHARACTERISTICS;
	MORPHOLOGICAL CHARACTERISTICS
	LEXICAL_ STRESS;
	OTHER_PHONETIC_CHARACTERISTICS
		GRAPHEME_TO_PHONEME_CORRESPONDENCE

LITERATURE_TTS		

The field values in the Database Table "TTS problems" contain the information about ways for TTS modules to be developed (or different problems to be overcome).

The DB is organised in such a way that each script can be used for several language and each language at the same time can be written in several scripts – relation many-to-many between the tables "Languages" and "Scripts" (see Appendix A)[9].

The database structure allows for different queries using any desired parameters. We plan to make the database available on the Web. Example of the current Database form representation is given in Figure 1.



Script_name: Bengali

Script_type: alphabetic-syllabic

Capitalisation: No

Graphemes_to_phoneme_correspondence: direct with certain exc

Spaces_between_words: Yes

Symbols_for_loan_words: Yes

Symbols_for_stress: No

Symbols_for_tones: N/A

Punctuation_marks: NF (only a period is to

Homographs: Yes

Other_characteristics: special signs for fraction

Comments to the fields:

Literature_scripts: Daniels, Peter and William Bright. 1996. The World's Writing Systems. Oxford University Press.

Languages

Lang_name	Lang_hierarchy	Number_of_sp	Country	State	Tones	Lexical_stress	Secon
Assamese	Indo-European - Indo-Aryan	12000000	India	Assam	N/A	FIXED: penultimate in eastern dialect,	NF
Bengali	Indo-European - Indo-Aryan	180000000	Bangladesh, India	West Bengal	N/A	FIXED: initial syllable	No seco
Meitei (Manipuri)	Sino-Tibetan - Tibeto-Burma	1100000	India	Manipur	Tone_language	SYLLABIC LANGUAGE	
Mundari	Austro-Asiatic	860000	India	Bihar	N/A	QUANTITY-SENSITIVE: Heavy syllable - clo	NF

TTS_problems

Lang_name	Script_name	Sentence_extraction	Tokenisation	Cliticalisation	Proper_Name_processi
Bengali	Bengali	period and other possible punctu	spaces between words	dictionaries of clitics	dictionaries: 1. Translated from other langu

Figure 1. The form representation of the Multilingual TTS-Related Database (languages written in Bengali script).

This representation is based on the main form "Scripts" and two sub-forms: "Languages" and "TTS_problems"

3 Evaluation of Language Complexity in Application to TTS Development Cycle

Evaluation of the complexity score in the TTS development cycle can be considered subjective to a certain extent (as well as evaluation of the performance of the TTS system as a whole). Two main parameters are normally used for TTS evaluation: intelligibility and naturalness. We tried to formalise these parameters in terms of required structural knowledge about script and language peculiarities to be used in required TTS development stages[10].

All script and language features in the Multilingual TTS-Related Database were scored in terms of the complexity of the first two stages of TTS development.

1. Creating a basic intelligible system
2. Creating a fully intelligible system
3. Creating a natural sounding system

In this paper we present a tentative scoring system for evaluation of TTS-related language complexity.

Table 1 contains such information including five South Asian languages, written in Italic.

A difficulty in TTS development for South Asian languages arises from the contradictory information about the place of lexical stress for automatic stress assignment (but in all cases, most of references indicate either fixed or quantity-based stress that is easy to predict automatically in comparison to free stress). There are several South Asian languages using tones that can be either predicted on the basis of rules (Punjabi) or can be indicated only on the basis of a lexicon (Manipuri, Tibetan). Major problems are caused by the languages with "complex" writing systems, such

[9] One language can be written in 2 scripts and therefore have two different types of TTS problems and their solutions, e.g. Malay can be written both in Arabic and Latin.

[10] Language-independent parameters such as voice-quality, signal processing tools etc. are not taken into account in this score system.

as the Arabic script (used for Pashto and Urdu) with optional vowel symbols and the Tibetan script with no spaces between words (these two languages have high complexity scores in Table 1). In order to solve the problems with such scripts machine learning techniques both for vowel insertion and for word extraction (Hackett and Douglas., 2000) have been developed.

Language	**Intelligibility (Basic)**	**Intelligibility (Full)**
Pashto	9	9.5
Russian	6	9
Arabic (Classical)	7	8.5
Tibetan	6	7.5
Thai	5	8
English	4	6
Hindi	3	5
Punjabi	2	4
Bengali	2	2.5

Table 1. Examples of the TTS-related complexity scoring for several languages (including 5 South Asian languages).

Below we present the main complex (from the TTS development point of view) script and language features for five South Asian languages.

Bengali:
SCRIPT FEATURES: No capitalisation
LANGUAGE FEATURES: grapheme-to-phoneme correspondence direct with certain exceptions (e.g. not rule-based pronunciation variants of the inherent vowel);
Hindi:
SCRIPT FEATURES: No capitalisation
LANGUAGE FEATURES: direct G2P correspondence with exceptions (e.g. not rule-based shwa-deletion);
Pashto:
SCRIPT FEATURES: No capitalisation; optional vowels
LANGUAGE FEATURES: Free stress; highly inflective morphology;
Punjabi:
SCRIPT FEATURES: No capitalisation
(Punjabi uses tones, but in contrast to Tibetan, symbol combinations can serve as cues for automatic tone assignment)
LANGUAGE FEATURES: rule-based stress assignment;
Tibetan:
SCRIPT FEATURES:
No capitalisation; no cues for tones; no spaces between words; no punctuation marks.

4 Conclusions and Further Work

On the basis of the developed Multi-Lingual Database it seems possible to make some conclusions about the TTS development for different South Asian languages. Most Indian languages (all Dravidian - Kannada, Malayalam, Tamil, Telugu; Indo-Aryan – Hindi, Nepali, Oriya etc.) are not difficult from the TTS development point of view – they use alphabetic-syllabic alphabet with direct (or almost direct) grapheme-to-sound correspondence and are characterised with relatively simple language characteristics (from the TTS development point of view).

The next stage of the current work is to test both available and new TTS techniques on several South Asian languages.

References

Dalby, Andrew. 1999. *Dictionary of Languages* (The definitive Reference to More than 400 Languages). Bloomsbury.

Daniels, Peter and William Bright. 1996. *The World's Writing Systems.* Oxford University Press.

Dutoit, Thierry. 1997. *An Introduction to Text-to-Speech Synthesis.* Kluwer Academic Publishers, Dordrecht.

Ethnologue (Volume 1). *Languages of the World.* 2000. SIL International, Dallas.

Goedemans, Rob, Harry van der Hulst and Ellis Visch. 1996. *Stress Pattern of the World (Part 1).* Holland Academic Graphics, The Hague.

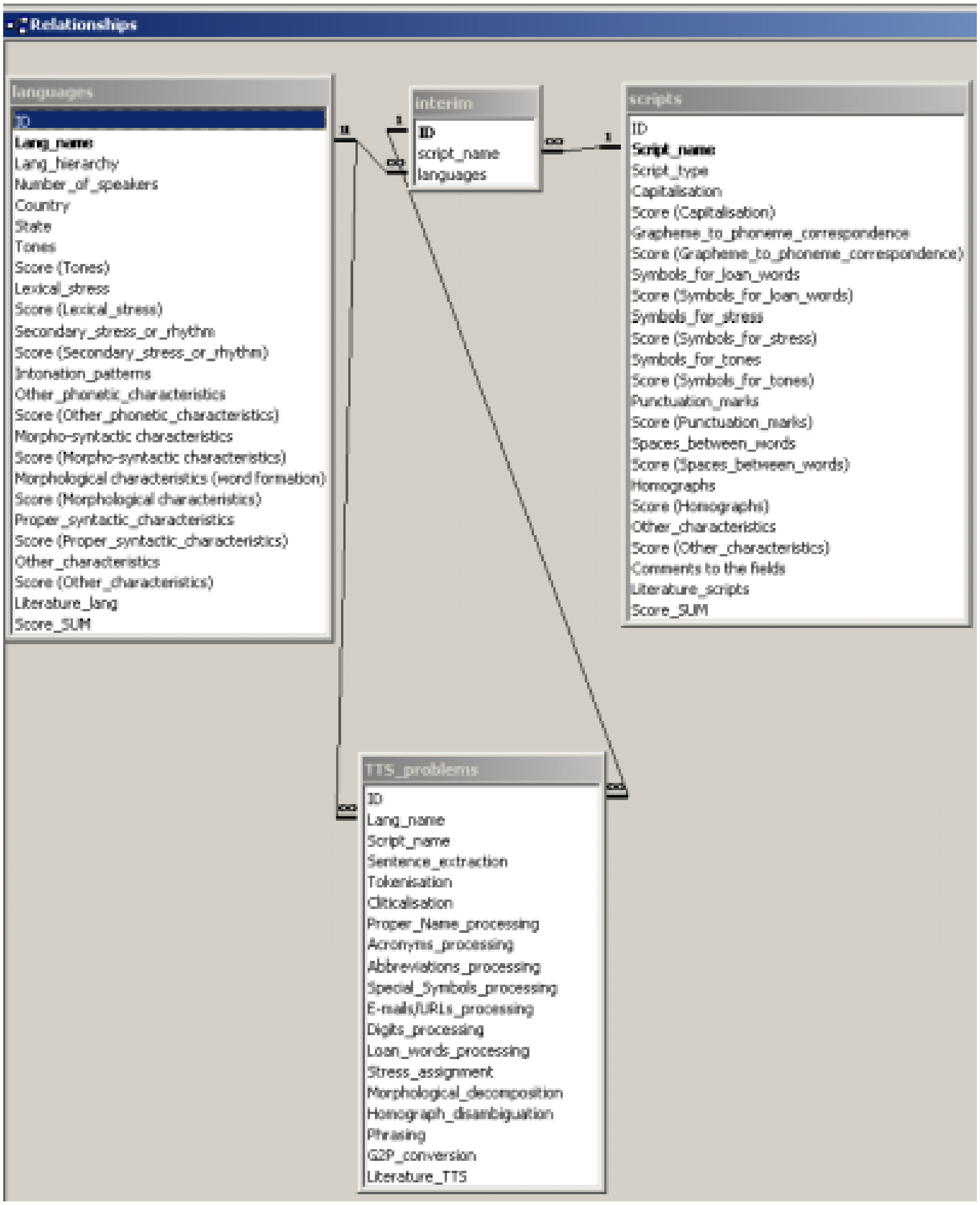
Hackett, Paul and Douglas W. Oard. *Comparison of Word-Based and Syllable-Based Retrieval for Tibetan.* Proceedings of the 5-th international workshop on Information retrieval with Asian languages, November 2000.

Multilingual Text-to-Speech Synthesis. The Bell Labs Approach. 1998. Editor R.Sproat. Kluwer Academic Publishers.

Nakanishi, Akira. 1998. *Writing Systems of the World.* Charles E. Tuttle Company.

http://www.rosettaproject.org:8080/live - *The Rosetta Project database.*

A. Appendix A. Relationships between the Tables in the Multilingual TTS-Related Database



languages

- ID
- **Lang_name**
- Lang_hierarchy
- Number_of_speakers
- Country
- State
- Tones
- Score (Tones)
- Lexical_stress
- Score (Lexical_stress)
- Secondary_stress_or_rhythm
- Score (Secondary_stress_or_rhythm)
- Intonation_patterns
- Other_phonetic_characteristics
- Score (Other_phonetic_characteristics)
- Morpho-syntactic characteristics
- Score (Morpho-syntactic characteristics)
- Morphological characteristics (word formation)
- Score (Morphological characteristics)
- Proper_syntactic_characteristics
- Score (Proper_syntactic_characteristics)
- Other_characteristics
- Score (Other_characteristics)
- Literature_lang
- Score_SUM

interim

- **ID**
- script_name
- languages

scripts

- ID
- **Script_name**
- Script_type
- Capitalisation
- Score (Capitalisation)
- Grapheme_to_phoneme_correspondence
- Score (Grapheme_to_phoneme_correspondence)
- Symbols_for_loan_words
- Score (Symbols_for_loan_words)
- Symbols_for_stress
- Score (Symbols_for_stress)
- Symbols_for_tones
- Score (Symbols_for_tones)
- Punctuation_marks
- Score (Punctuation_marks)
- Spaces_between_words
- Score (Spaces_between_words)
- Homographs
- Score (Homographs)
- Other_characteristics
- Score (Other_characteristics)
- Comments to the fields
- Literature_scripts
- Score_SUM

TTS_problems

- ID
- Lang_name
- Script_name
- Sentence_extraction
- Tokenisation
- Cliticalisation
- Proper_Name_processing
- Acronyms_processing
- Abbreviations_processing
- Special_Symbols_processing
- E-mails/URLs_processing
- Digits_processing
- Loan_words_processing
- Stress_assignment
- Morphological_decomposition
- Homograph_disambiguation
- Phrasing
- G2P_conversion
- Literature_TTS