CHAPTER 1. INTRODUCTION

1.1 Background

As the usage of Information Technology (IT) increases amongst the Sri Lankan community, it is essential to build local language IT infrastructure. This includes system software that is usable in local language, local language support in application software and software tools that enable local language computing.

IT has been used in Sri Lanka for a long period, but this use has been limited to those with knowledge of English (Dias, et al., 2004), which concentrates the IT usage to urban areas. However now the situation is changing and the need for IT usage among non-English speaking community is growing. Several initiatives has been carried out and completed to develop local language computing in Sri Lanka (Dias, et al., 2004). Nevertheless still, there is a need for supporting software tools/technology such as searching, collation, storage, etc in local language to enable practical use of fully localized IT, including localized storage and retrieval of information/data.

Sinhala is the majority language in Sri Lanka (about 82% of the population of Sri Lanka is Sinhalese). However according to the last census data, the percentage out of Sinhala population who are able to speak English is about 13% (Dept. of Census and Statistics, 2001). Therefore, development of Sinhala computing infrastructure and software tools will greatly benefit a majority of the Sri Lankan community.

1.2 Data and Information in Local Language

With the growing use of IT in a multi-ethnic country like Sri Lanka where English is not the official language, the need to store and retrieve information and data in the local language has become a growing necessity. Therefore, most databases in public & large organizations will have to be multi-lingual. A vast majority of this information will include personal data of people—names and addresses. Various government organizations such as the Department of Registration of Persons, Department of Immigration & Emigration will need to maintain large personal information databases.

Ability to search for data/information is a necessary aspect of databases. Therefore, all these organizations will greatly benefit from being able to search for information using the local language.

Personal information consists of proper nouns or out of vocabulary (OOV) words. Due to the spelling variations that are found in these OOV words searching, retrieving them is a challenging problem. The difficulty of the problem increases when it comes to cross-language information retrieval as the variation of spelling is very high (AbdulJaleel, et al., 2003).

1.3 Complications due to Spelling Variations of Proper Nouns

Names of people and places can be spelt in different ways by different people. Especially if a particular proper noun or an out of vocabulary (OOV) word has a foreign origin, the target language spelling variability could be very high. As AbdulJaleel cites (AbdulJaleel, et al., 2003) an article identifies 32 different English spellings for the name of the Libyan leader Muammar Gaddafi (Whitaker, 2005).

Sri Lankans use names originated from several languages. Major ones are Sinhala, Tamil, Arabic, English, Dutch, Portuguese, Pali and Malay (Weerasinghe, 2006). Therefore, the above problem persists in Sri Lankan OOV words. It is not only due to cross-ethnic name conversion but also within the same language. For example, the pure Sinhala name විශාඛා can be spelt in different ways such as විසාකා, විශාකා or විසාඛා by different people. One cannot say a particular way of spelling is incorrect because it is a proper name. Similarly, cross-ethnic names are spelt in different ways in Sinhala due to inconsistent transliterated form of usage since old times. For example, the Arabic name Mohommad (which may be spelt in number of different ways in English) can be spelt in different ways in Sinhala such as: මොහොමඩ්, මොහම්මඩ්, මොහමඩ්, මහමඩ්, මොහම්මදු, මුහම්මද්, මුහම්මදු, මහම්මද්.

When it comes to Tamil words written in Sinhala, there are notable spelling variations due to certain phonetic differences of the two languages. For example, ප්‍රදීපන් can be spelt as පිරිතීපන් and similarly ගුණදර්ශන්, කුනදර්ශන්, ගුණතර්ශන්, etc could be different spellings of the same name. If observed carefully, it can be realized that the variation patterns of these spelling variations differs based on the language origin of the name. For example for a Tamil name it can be identified that the sound of letter 'ක' is getting replaced with the sound of letter 'ග' and similarly 'ද' with 'ත'. However, this type of replacement will not occur to a name with a Sinhala origin or a Portuguese origin.

On the other hand, due to the Western influence and due to being transliterated through English (double transliteration) a person may search for a purely Sinhala word using different spelling. For example, a person intending to find 'දිලිනි' may search for 'ඩිලිනි'. Similarly, to find 'චානක' one could search 'ශානක' or 'ශනක' due to the same reason.

Due to this variation of spelling, a conventional search on a set of proper names data using a single search string may not match with an intended target record. In order to get the intended matches the data set may have to be searched using a set of words with possible combinations of different spellings.

1.4 Research Objectives, Scope and Deliverables

This research focuses on building a solution for the problem mentioned in the above sections. The objective is to build a rule based search engine to search a database consisting names of people and place names using a Sinhala search string as the input. The engine should determine different ways of spelling to the given name and search the database with all those strings so that it could match a record in the database even though the corresponding name is spelt in a slightly different way. The engine should match records even if the search key is spelt incorrectly due to cross-ethnic conversion issues or due to personal pronunciation/spelling differences as mentioned in preceding sections.

This research is scoped out to only process a Sinhala string as the input search key. The research involved in identifying the unit(s) of replacement (i.e. parts of word) to derive similar key words with different spelling and building string replacement rule base. For example, for the input string විශාඛා, a set of words such as විශාඛා, විසාකා, විශාකා and විසාඛා is generated, through replacing characters/substrings by executing the rules identified. Then finally, a database is queried using the generated set of words and the matching records are displayed. In the above example, a person searching for ‘විශාඛා’ may find a match with a record spelt as ‘විසාකා’.

Since there was no existing database with Sinhala Unicode names, creating the necessary database schemas and test data was also part of this effort.

Apart from this dissertation the working application, documented string replacement rules and the test database with 1000+ Sri Lankan names (in Sinhala) were outcomes of this research effort.

The remaining chapters of this dissertation will provide more insight to the methodology and approach taken in designing the systems together with the issues and challenges involved. Chapter 2 covers the literature review done in conjunction with this research. Chapter 3 describes the methodology followed in completing this research, how the problem was decomposed and solution was approached. The Chapter 4 covers the design and implementation of the system in detail, where as chapter 5 gives details of testing and evaluates results from the resulting system. Finally, chapter 6 has the conclusion and the future work relating to this system.