Some computational aspects of natural language understanding:
A case study on Spoken Sinhala

Chandima H. de Silva
and
Neelakshi Chandrasena Premawardhena

Address for Correspondence

Department of Statistics and Computer Science
University of Kelaniya, Kelaniya, Sri Lanka.
Email: chandima@kln.ac.lk

Paper submitted for the

9th International conference on Sri Lanka Studies,
28th – 30th November 2003,
Matara, Sri Lanka
Some computational aspects of natural language understanding: A case study on Spoken Sinhala

Chandima H. de Silva
Department of Statistics and Computer Science
and
Neelakshi Chandrasena Premawardhena
Department of Modern Languages

University of Kelaniya, Sri Lanka

Abstract
Computational Linguistics (CL) is a relatively new area of research that incorporates knowledge from both computer science and linguistics. CL is mainly concerned with the computational aspects of the human language faculty. Natural language understanding plays a major role in designing effective CL systems. This paper attempts to address some issues related to transferring Spoken Sinhala into electronic form in the context of computational linguistics. Though studies have been conducted in relation to European languages, a study of this nature has not been done involving Sinhala.

Diglossia in modern Sinhala has paved for two separate areas of research today. Linguists refer to the two varieties as Literary and Spoken Sinhala. Although much research has been conducted on Literary Sinhala, the area of Spoken Sinhala has been somewhat neglected. The present study is based on data available from Spoken Sinhala, thus attempting to contribute to filling this void.

This paper concentrates only on reference devices in Spoken Sinhala. Unlike major European languages used in applied CL, Sinhala belongs to the category of languages that employs many sociolinguistic variables not signalled directly by the grammatical devices. Identifying speaker and addressee generally causes no great difficulties in the discourse. However, keeping track of third persons poses many difficulties in Spoken Sinhala, for instance, due to heavy use of zero anaphora and gender neutral pronouns. Thus, this paper also discusses the possible difficulties that would arise in using Spoken Sinhala in CL.

1. Introduction

Computational Linguistics is mainly concerned with the computational aspects of the human language faculty. It belongs to the cognitive sciences and overlaps with the field of Artificial Intelligence (AI), a branch of computer science that aims at simulating human cognition. For many students and practitioners of computational linguistics, the special attraction of the discipline is the combination of expertise from the humanities, natural and behavioural sciences, and engineering. Scientific approaches and practical techniques in CL come from linguistics, computer science, psychology, and mathematics. Theoretical CL takes up issues in theoretical linguistics. It deals with formal theories about the linguistic knowledge that a human needs for generating and understanding language. Today these theories have reached a
degree of complexity that can only be managed by employing computers. Computational linguists develop formal models simulating aspects of the human language faculty and implement them as computer programs. These programs constitute the basis for the evaluation and further development of the theories. Applied CL focuses on the practical outcome of modelling human language use. The methods, techniques, tools and applications in this area are often subsumed under the term language engineering or language technology. Although existing CL systems are far from achieving human ability, they have numerous possible applications. A major goal of applied CL is to create software products that have some knowledge of human language. Such products are going to have a tremendous impact on both computer science and linguistics. They are urgently needed for improving human-machine interaction since the main obstacle in the interaction between human and computer is a communication problem.

Interpretation of formal languages (i.e., logical or mathematical languages) can be done using formal semantics. However, until Montague put forward his theory, the common belief was that formal semantics could not be applied to natural languages mainly due to inexact syntax found in almost every natural language. Montague argued that it is possible to comprehend the syntax and semantics of both formal and natural languages within a single mathematically precise theory (see Montague 1970a, 1970b, 1974 and Cann 1993). A natural consequence of Montague’s theory is that one should be able to design a computational model to represent any grammatical relation in any language. This, however, is by no means an easy task. If one can restrict the world of discourse to a particular subset of grammatical relations, building a computational model would be a feasible task.

2. Natural Language Understanding

Natural Language Understanding (NLU) is a sub-field of artificial intelligence research devoted to making computers ‘understand’ statements written in human languages (see Allen, 1995). Though there are have been many attempts to develop NLU computer systems using European languages, there has been no research on developing similar systems involving Sinhala. The study presented in this paper attempts to fill this void and produce a NLU system that would be capable of ‘understanding’ and analysing Spoken Sinhala. The motivation for this research stems from two facts, namely, the need for NLU in most day-to-day situations and the involvement of several sociolinguistic variables in Spoken Sinhala than in Literary Sinhala.

Knowledge representation in computer programs can mainly be done using agents in the form of sentences in a knowledge representation language that are stored in a knowledge base (see Russell & Norvig, 1995 and Cawsey, 1998). Our attempt in this study is to represent Spoken Sinhala in a knowledge base, with special emphasis on reference devices.

2.1 Designing a CL system for Spoken Sinhala
In designing an effective CL system for Spoken Sinhala one has to consider the linguistic forms of the language as well as sociolinguistic factors which are not directly signalled in speech, but play a most significant role. In an attempt to discuss
the significance of the 'non-grammatical' factors in discourse, this paper concentrates on reference devices (mechanisms employed by native speakers to keep track of participants in discourse, i.e. pronouns, nouns, honorific and kinship terms) in Spoken Sinhala, where heavy use of sociolinguistic variables is evident.

2.2 Literary Sinhala vs. Spoken Sinhala
Diglossia in modern Sinhala with a vast difference between the spoken and the written variety has paved the way for two separate areas of research today. Linguists refer to the two varieties as Literary and Spoken Sinhala. Spoken Sinhala is further classified as Formal and Colloquial varieties (see Karunatillake, 1992; Disanayaka, 1998; Gair, 1998; Ch. Premawardhena 2002a, 2002b). Literary Sinhala has been the main focus on research so far and the area of Spoken Sinhala has been somewhat neglected. The present study is based on data available from research conducted on Spoken Sinhala, thus attempting to contribute to filling this void.

3. Identifying and keeping track of referents in discourse
Identifying the continuity and discontinuity of referents and keeping track of them in an ongoing conversation is no easy task for the participants. During longer phases of discourse more than one referent in the third person may occur. Thus, while listening to and absorbing the facts presented in the conversation, the addressee has to concentrate a) on the (third person) referents that are newly introduced, b) continuation and discontinuation of the referents and c) coming back to a referent who may have been mentioned right at the beginning of the conversation and then not referred to for a long period.

How do the native speakers identify these referents? Native speakers are not only aware of the grammatical devices involved in reference, i.e. gender, number, case, nominal and pronominal devices, they are also well aware of the subtle social norms and customs that serve in identifying referents, i.e. honorific and kinship terms. Further, gestures, mimicry and common ground shared by the speaker and the addressee play a major role in signifying referents in discourse.

“Common ground is the foundation for all joint actions, and that makes it essential to the creation of the speaker’s meaning and addressee’s understanding as well.” (Clark, 1996:14)

Unlike major European languages used in applied CL, Sinhala belongs to the category of languages that employs many sociolinguistic variables not signalled directly by the grammatical devices. Identifying speaker (first person) and addressee (second person) generally causes no great difficulties in reference tracking. However, identifying third persons in discourse poses many difficulties in Spoken Sinhala, for instance, due to heavy use of zero anaphora, gender neutral pronouns ('eyā'), deictic terms ('ar̄ya', 'meyā', 'oyā', 'eyā'). Further, zero anaphora occurs in first, second as well as in third person reference in Spoken Sinhala (Ch. Premawardhena, 2000a: 176-179; 2002b: 74-75). Thus, the inference system (Foley & Van Valin, 1984:324) plays an important role in tracking referents in discourse. Although inference is used in all languages in some form, in languages of South, Southeast and East Asia this “has been elevated to the state of a fine art” (Foley & Van Valin, 1984)⁴.

⁴ see Head, B.F. (1978:156, 162-63)
Zero anaphora is heavily used in these languages but assignment of coreference is often determined by the subtle use of sociolinguistic variables and is not directly signalled in the linguistic form (Foley & Van Valin, 1984:324).

The following sections discuss what aspects of reference devices in Spoken Sinhala need to be paid attention in an attempt to use Sinhala in applied CL. To overcome any constraints in designing a CL system for Spoken Sinhala, grammatical as well as sociolinguistic aspects should be incorporated.

3.1 The speaker - first person referent
Apart from the first person pronouns *mam* 'I', *api* 'we', heavy use of zero anaphora is evident in Spoken Sinhala in first person reference (Ch. Premawardhena, 2002a: 176-79; 2002b: 74-75). Further, especially in denoting possession, the plural form *apē* 'our' is preferred to the singular form *mage* 'my' although in reality it is first person singular who is being referred to. This could be attributed to the nature of the Sinhala native speakers where collective thinking is generally preferred to individuality, i.e. *apē gedrē* 'our/my house', *apē amma* 'our/my mother'. Even 'my wife' or 'my husband' is referred to as *apē nōna* 'our wife' and *apē mahattēya* 'our husband' (see also Ch. Premawardhena 2002a:145-46, 2002b: 68).

3.2 The addressee - second person referent
In second person reference both nominal and pronominal devices are employed by the Sinhala native speaker. Socio-linguistic factors play a significant role in the use of second person reference in Spoken Sinhala. Social status of the speaker and the addressee, level of education, age, sex determine the choice of the second person reference devices. The pronouns *oyā* pl *oyāla* are used among friends, educated urbanites and rural folk of the same age and social status. *u-bē* pl *u-bēla*, *tō* pl *topi, tamuse* pl *tamusela* are used by not-so-educated urbanites and rural folk in a neutral sense. However, the same set of second person pronouns are considered derogatory when used by the educated urbanites and rural folk.

The most common second person reference devices are honorific *mahattēya* pl *sē*: 'mister, master, gentleman', *nōna/mis* 'lady' and kinship terms *tāttā* 'father', *amma* 'mother', *mēma* 'uncle', profession + honorific term *dostērē* *mahattēya* 'doctor'. Proper nouns are usually accompanied by an honorific term *mis* 'miss/lady', *misē* 'mister'. Zero anaphora occurs too, in place of second person referent (see also Ch. Premawardhena 2002a:146-150, 2002b: 67).

3.3 The third person referent
As in second person reference, social status, age, sex and educational level of the referent determines the use of third person reference devices. The proximity (1 PROX, 2PROX, distal -DIST and anaphoric -APH) to the referent also plays a significant role in identifying the referent (see Gair, 1998:112, Gair/Karunatillake, 2000:721,729). Ch. Premawardhena (2002a:150-57, 2002b:67-74) discusses third person reference in detail. Apart from the two sets of pronominal devices used by the educated and not-so-educated native speakers, proper nouns, honorific and kinship

---

2 see Head, B.F: (1978:166)
terms, profession + honorific and zero anaphora belong to the third person reference devices.

Table 1: Reference devices in Spoken Sinhala

<table>
<thead>
<tr>
<th></th>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>educated urbanites, educated rural folk</td>
<td>not-so-educated urbanites, not-so educated rural folk, men among peers</td>
</tr>
<tr>
<td>m/f</td>
<td>m     f</td>
<td>m     f</td>
</tr>
<tr>
<td>animate</td>
<td>Ø, proper nouns, honorific and kinship terms</td>
<td>undæ, undæ</td>
</tr>
<tr>
<td>human</td>
<td>miniha</td>
<td>ge:ni</td>
</tr>
<tr>
<td>non-human</td>
<td>ü, méka, óka, aržka, éka, ü</td>
<td>měki, óki, aržki, éki</td>
</tr>
<tr>
<td>inanimate</td>
<td>měku, óku, aržku, éku</td>
<td>měku, óku, aržku, éku</td>
</tr>
</tbody>
</table>

(Ch. Premawardhena, 2002b:80)

The reference devices of Group 2 is used more often than not by Group 1 derogatorily to denote disrespect towards the referent. Further, in Spoken Sinhala, the degree of respect towards the referent is denoted by the use of each reference device, i.e. unnāñse, hāmuduravo 'reverend' for Buddhist monks or kinship terms in both second and third person reference for elders.

Since the same reference devices, i.e. zero anaphora, honorific and kinship terms, proper nouns and other nominal devices, are used by the Sinhala native speakers to identify both second and third person referents, there may be difficulties in tracking referents, i.e. 3SG animate, non-human ü is used by Group 2 to denote both animate human and non-human referents. Further, Group 1 uses the same pronoun for animate non-human, whereas it can also be used for animate human derogatorily (see Ch. Premawardhena, 2002a:.187-194, 2002b: 77-79).

4. Conclusion

To keep pace with the advancement of computer science related tools available to linguists, developing a NLU system for Sinhala is of utmost importance. Such a system would be invaluable to researchers in the field of computational linguistics. This paper attempted to address the need for developing such a system for Spoken Sinhala and the importance of incorporating the formal grammatical rules as well as sociolinguistic variables. With examples cited from reference tracking mechanisms employed by Sinhala native speakers, what aspects of Spoken Sinhala should be taken into consideration in NLU apart from the formal grammatical rules was discussed in detail. The more evident reference devices common to many languages of the world as gender, number, case, position of subject and object were not elaborated in this paper since the focus was mainly on constraints that may occur in developing a NLU system for Spoken Sinhala.
References


Glossary

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1st person</td>
</tr>
<tr>
<td>2</td>
<td>2nd person</td>
</tr>
<tr>
<td>3</td>
<td>3rd person</td>
</tr>
<tr>
<td>APH</td>
<td>anaphoric</td>
</tr>
<tr>
<td>DIST</td>
<td>distal</td>
</tr>
<tr>
<td>m</td>
<td>masculine</td>
</tr>
<tr>
<td>f</td>
<td>feminine</td>
</tr>
<tr>
<td>pl</td>
<td>plural</td>
</tr>
<tr>
<td>PROX</td>
<td>proximal</td>
</tr>
<tr>
<td>SG</td>
<td>singular</td>
</tr>
</tbody>
</table>

Symbols

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>- (ã), : (æ:)</td>
<td>long vowel</td>
</tr>
<tr>
<td>ø</td>
<td>central unrounded vowel</td>
</tr>
<tr>
<td>Ø</td>
<td>zero anaphora</td>
</tr>
<tr>
<td>æ</td>
<td>higher-low front unrounded vowel</td>
</tr>
<tr>
<td>η</td>
<td>velar nasal</td>
</tr>
</tbody>
</table>