Computational Linguistics (CL) in Pakistan: Issues and Proposals

Sarmad Hussain
Center for Research in Urdu Language Processing (CRULP)
National University of Computer and Emerging Sciences
852 B Block, Faisal Town, Lahore, Pakistan
sarmad.hussain@nu.edu.pk

Abstract

Internet Communication Technology has opened new venues for CL. Because of this information revolution, research and development is now viable for many languages of Pakistan. This paper briefly presents the current work in CL in Pakistan, issues in its development and some proposals for accelerating the current pace of work in computational modeling of Pakistani Languages.

1 Introduction

There are fifty seven languages[1] spoken in Pakistan[2] (Rahman 2002). English is only understood by about 5% of this population. Therefore, for a Pakistani to benefit from the IT revolution (e.g. to give them access to services including e-governance and e-commerce), solutions must be provided to this population in local languages. This paper introduces the work in progress in computational modeling of local languages spoken in Pakistan and current issues in pursuing such work. Further, the paper also presents proposals to promote CL in Pakistan.

Most of the research and development work related to languages has focused on modeling orthography to develop word processors. These solutions were developed by private sector in 1980's, which could not continue this development because of losses incurred due to insufficient enforcement of copyright laws in Pakistan.

2 Current Work

Though limited work has been done, with growing need, interest in CL is increasing. Work is currently being done in the following areas:

- Lexical development and corpus based lexical data acquisition at CRULP
- Grammar Modeling at CRULP
- Machine Translation at Karachi University and Pakistan Institute of Engineering and Applied Sciences
- Linguistic research at CRULP and National University of Modern Languages
- Optical Character Recognition at Ghulam Ishaq Khan Institute
- Speech Synthesis and Recognition at CRULP

3 Issues in CL

Following challenges are currently faced by the researchers who are working in computational linguistics (and related areas) in Pakistan.

3.1 Linguistic Research

With such a rich breeding ground containing fifty seven not-so-well-studied languages, it is interesting to note that even up till 1999 "Pakistan [did] not have a university department or institute of higher education and research in linguistics" (Rahman 1999). With growing realization, few organizations are now established. However, much ground in basic research in these languages needs to be covered. Some original work is available, but most of it is either old (e.g. Platts 1909, Shackle 1976, etc.) or not to the level of detail required for computational modeling. However, there is some recent work available (e.g. for Urdu: Butt 1995, Hussain 1997, Moizuddin 1989), but more needs to be done.

As an example, there is still controversy on existence of Urdu phonemes including /l^h, m^h, n^h, r^h/ (Saleem et al. 2002). Similarly, only recently have Urdu (phonological) sound change rules been partially documented (e.g. Zia 2002). With such basic issues still unsettled, it is difficult to

[1] Ethnologue estimates sixty-six languages (Grimes 1992)
[2] Population of 0.127 billion (1981 census) (Rahman 2002)

develop speech synthesis or recognition applications. Similarly, work in other areas, including Morphology, Syntax, Semantics is also limited. Work in other Pakistan languages is lagging behind Urdu, to the extent that even major Pakistani dialects of Punjabi (the most spoken language of Pakistan) have not yet been documented.

3.2 Standardization

Another significant problem faced by researchers is lack of standardization of languages. Though literature is available on many of these languages, different views presented have still not been debated and consolidated. This issue is highlighted through the following examples.

Script of many languages, e.g. Balti, Burushaski, Shina, Khowar, etc., does not exist and is currently being proposed by researchers (Baart 1997, pp. 50-56). This limits ways to process these languages using computers. Where scripts exist, there is lack of consensus on the writing styles. For example, currently Punjabi Rnoon (nasal retroflex flap) is written in three different ways. Though this variation may be handled through fonts, it also puts obstacles in developing and usage of language processing applications.

Worse problem is whether a character exists in a language. Character sets of many languages are not final. As an example, characters in Urdu vary from fifty-three (Siddiqui & Amrohi 1977) down to thirty-eight (e.g. Platts 1911). Similarly, new combined character set has been introduced for Kandhari and Yusufzai dialects[3] of Pashto, but has only been partially accepted[4]. This lack of consensus poses serious impediments in development of computational lexica and for other applications as well. There has been some development recently (e.g. Hussain and Afzal 2001), but much more work needs to be done.

Equally significant is the problem of order of characters in a language. All applications which depend on sorting and indexing (including computational lexica) cannot be developed unless collation sequence has been standardized for a language. Though data for languages is being collected and being finalized, standardized collation sequences are still not available for most Pakistani languages.

Many other basic standards required for computing are not available, which pose problems in the development of applications. These include standards for keyboards and fonts for many languages (Afzal 1999).

3.3 R&D Funding

Though work is being done, progress is slow because of limited funding available. Much of the work being done in linguistics and CL is being funded through foreign support. There is growing awareness in public[5] and private sectors of importance of this work, but it will perhaps take some time before adequate amount of funds are diverted to these areas. For the first time, Rs. 100 Million were allocated for language software development by Ministry of Science and Technology in 2001, but most of it lapsed as no projects were actually awarded.

4 Proposals for Development of CL

Following are few recommendations which can accelerate the research and development activity in computational linguistics in Pakistan.

- Research work in basic linguistics in Pakistani language must be started. This can be achieved by starting university level research departments and other research organizations.
- Linguistic research can be further enhanced if research funding is allocated for Europeans and Pakistanis to do doctoral and (eventually) post-doctoral work in linguistic aspects of not-so-well-studied Pakistani languages. Collaborations between Pakistan and European organizations for survey related work should be encouraged through such programs
- Of the fifty seven languages, Punjabi, Pashto, Sindhi, Siraiki, Urdu[6], Balochi and Hindko are most spoken languages (in order of speaking population), and cover almost 96 percent of the population of Pakistan (Rah-

[3] Spoken in Afghanistan and Pakistan respectively.
[4] Personal communication with Dr. Raj Wali Shah, Chairman, Pashto Academy, Peshawar Univ., Dec. 2002.

[5] Urdu and Regional Languages' Software Development Forum (URLSDF) was recently devised by Ministry of Science and Technology (see www.tremu.org.pk).
[6] Urdu is the lingua franca and used by people speaking different languages to communicate with each other.

man 2002). Therefore, work should first be done in these languages.

- Standardization of various aspects of languages, which have been highlighted, must be achieved. URLSDF is currently comprised of volunteers, which slows progress. Dedicated resources and funds should be allocated to achieve this task
- Government should prioritize projects and should develop a roadmap for their completion (both in linguistics and CL). Accordingly, government should allocate funding for these projects to relevant R&D organizations for development
- Government should provide better copyright support for private sector investors
- Relevant European organizations (e.g. EACL, ISCA, EAA, ELRA, ELSNET) should come forward to help local organizations do R&D in these areas through collaborative programs, training and funding. For example, EuroMasters program (by EACL and ISCA, which is currently limited to Europe) should be extended to help universities institute similar programs in Pakistan. This could be further achieved if relevant European organizations develop local chapters or regional chapters in South Asia
- Exchange programs between Europe and South Asia should also be initiated for accelerated transfer of technology and expertise (both in Linguistics and CL)

5 Conclusions

Much ground work needs to be done before reasonable activity in CL can be triggered in Pakistan. This may only be achieved if serious efforts and funding are diverted towards it. Government of Pakistan is a key player which can make this happen. However, support by European universities, research centers and organizations can help accelerate this process.

References

M. Afzal. 1999. Urdu Software Industry: Prospects, Problems and Need for Standards. *Science Vision* 5(2). Islamabad, Pakistan.

J.L.G. Baart. 1997. *Sounds and Tones of Kalam Kohistani: With Wordlist and Texts*. National Institute of Pakistan Studies, Quaid-e Azam University, Islamabad, Pakistan.

M. Butt. 1995. *The Structure of Complex Predicates in Urdu.* CSLI Publications Stanford, CA USA.

B. Grimes. (ed.) 1992. *Ethnologue: Languages of Pakistan.* 13th Edition. Summer Institute of Linguistics.

S. Hussain and M.Afzal. 2001. Urdu Computing Standards: UZT 1.01. *Proceedings of the IEEE International Multi-Topic Conference*. Lahore University of Management Science, Lahore, Pakistan.

S. Hussain. 1997. Phonetic Correlates of Lexical Stress in Urdu. *Unpublished PhD thesis*. Northwestern University, Evanstion, IL, USA.

M. Moizuddin. 1989. *Word Forms in Urdu.* National Language Authority, Islamabad, Pakistan.

J. Platts. 1911. *A Dictionary of Urdu, Classical Hindi and English.* Crosby, Lockwood and Son, London, UK.

J. Platts. 1909. *A Grammar of the Hindustani or Urdu Language.* Crosby, Lockwood and Son, London, UK.

T. Rahman. 1999. *Language, Education and Culture.* Oxford University Press, Karachi, Pakistan.

T. Rahman. 2002. *Language Ideology and Power: Language Learning Among the Muslims of Pakistan and North India*. Oxford University Press, Karachi, Pakistan.

M. Saleem, H. Kabir, K. Riaz, M. Rafique, N. Khalid and R. Shahid. 2002. Urdu Consonantal and Vocalic Sounds. *CRULP Annual Student Report 2002,* CRULP, NUCES, Pakistan.

C. Shackle, 1976. *The Siraiki Language of Central Pakistan: A Reference Grammar.* SOAS, University of London, London, UK.

A. Siddiqui, and N. Amrohi (eds.). 1977. *Urdu Lughat: Volume 1.* Urdu Dictionary Board, Karachi, Pakistan.

A. Zia, 2002. Assimilation and Dissimilation Rules in Urdu. *CRULP Annual Student Report 2002,* CRULP, NUCES, Pakistan.