Localising Nations, saving languages: moving from Unicode to Language Engineering

Patrick A.V. Hall,
Computing Department, The Open University, Walton Hall, Milton Keynes, UK
p.a.v.hall@open.ac.uk

and

Global Initiative for Local Computing
Localisation Research Centre, University of Limerick, Limerick, Ireland.

Abstract: It is believed that all peoples would benefit from the use of computers and access to the Internet, a belief that was reflected in the World Summit on the Information Society in Geneva at the end of 2003. But computers, and now the Internet, are dominated by western nations working in English and other major languages of the developed world. If computers and the internet are to be widely used, clearly it should take place using the language of that community, just as all other activities do. Using Unicode is a first step, and it has been proposed the Unicode should be funded to provide encodings for all known writing systems. In this paper I will spell out what is required to move beyond Unicode to the language technologies that underpin the use of Unicode and software localisation, looking at developments in South Asia, as the countries of the region move towards using computers and the Internet in their own local languages.

1. Introduction

It is estimated that worldwide there are 6000 to 7000 languages, with these disappearing at an alarming rate so that only some 600 of these can be viewed a ‘safe’ and likely to survive (Nettle and Romaine 2000). Even major European languages like Swedish and Danish are now fearful of survival (Allwood 2004). The response has been to raise projects to document these languages before they disappear altogether, for example the Endangered Languages Documentation Programme (ELDP) funded by the Lisbet Rausing Charitable Fund at the School of Oriental and African Studies, University of London (SOAS), or the projects funded by the Volkswagen Stiftung in Germany. Diversity of languages is seen as precious to humanity– the languages embed within them particular conceptualisations of the world and knowledge that we should not lose.

This documentation and preservation view of languages must be contrasted with language activism which seeks to revitalise languages as human resources in active use and intimately associated with the identity of people (Eisenlohr, 2004). Technology such as broadcast television has been seen as a major cause of language extinction, and equally well technology is seen as a major instrument in language revitalisation.

In this paper we will explore what technology is required to revitalise a language We will look in depth at the situation in South Asia, in particular focussing on Nepal as a case study, concluding that not only must there be the technology available to support the writing system for the language, as characterised by Unicode, but that the whole language itself must be supported through language engineering resources as would be seen as normal for European languages.

2. Languages and their Power

We will start by looking at the languages of South Asia and how some of those languages dominate others. South Asia has well over 1 billion inhabitants speaking between 500 and 1000 distinct languages, falling into three major language groups with a few languages from other language groups. The largest is the Indo-European group to which most European languages also belong, and covers the major languages of Bangladesh, northern India, Pakistan, Afghanistan, as well as Nepal. To the south are the Dravidian languages, the best known of these being Tamil, an official language not just of India but also of Sri Lanka and Singapore. Sinhala, the language of the rest of Sri Lanka, is usually classified as Indo-European, but with a strong Dravidian influence from its long contact with Tamil. The remaining languages in the north of the region, in the Himalayas, are Tibeto-Burmese, which includes Dzongka the official language of Bhutan and covers most of the languages of South East Asia as well.

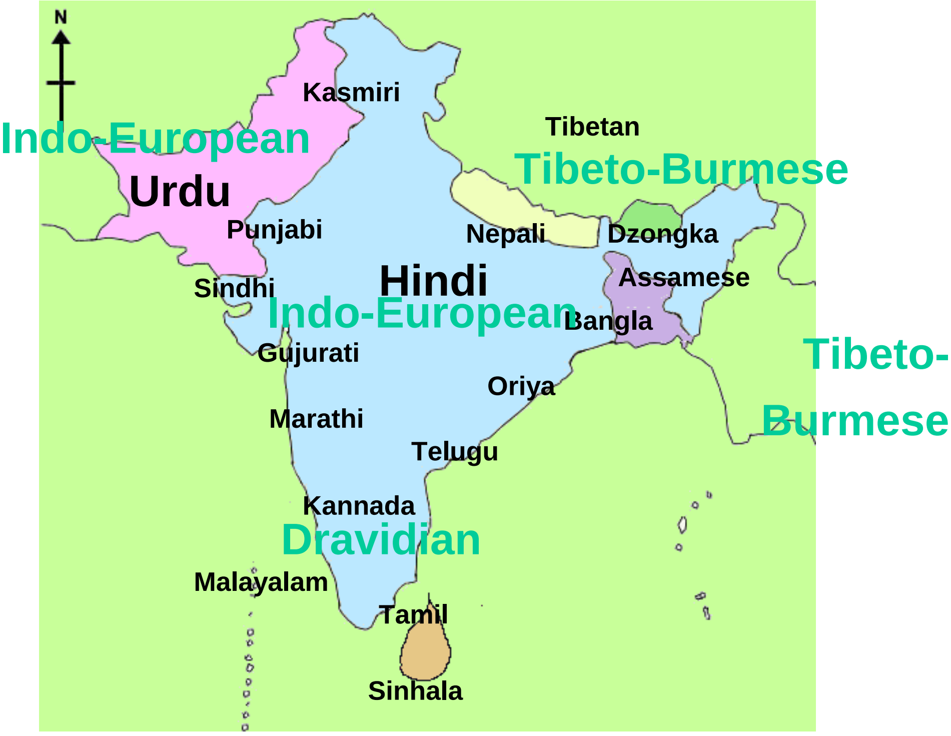

Figure 1. The major languages of South Asia.

Figure 1 shows the distribution of the major languages of South Asia. In India the 1991 census identified 1576 'mother tongues' which were subsequently rationalised into 216 and then 114 languages from five language families. In this rationalisation, linguistic communities of less than

10,000 people were dropped, and thus the number of languages is very much higher than that final 114 (see Mallikarjun 2004 or the Central Institute for Indian Languages (CIIL) web-site). Hindi is the national language of India and spoken as first language by about one third of the inhabitants, there are 21 other 'scheduled' or official languages with large populations of mother-tongue speakers, while there are very many minor languages frequently characterised as 'tribal'. In Pakistan there are around 70 distinct languages (Rahman 2004) with nearly half the population speaking Punjabi while Urdu the official language is the mother tongue of only 7.5% of the population. In Bangladesh there is the single dominant language, Bangla, spoken by maybe 96% of the population, though it is contested that Sylheti, seen by some as a dialect of Bangla, is in fact a distinct language. In Nepal there are about another 100 or so languages (Turin 2004), with the major ones shown in Figure 2. Only in Sri Lanka is the variety limited, with just 2 languages, Sinhala and Tamil.

Total Population = 20,055,632

Tibet-Burmese Group of Languages

Language	**Number**	**%age**
Gurung	252,381	1.26%
Limbu	28,224	0.14%
Magar	476,445	2.38%
Newari	764,067	3.81%
Rai-Kirat	486,464	2.43%
Sherpa & Bhote	134,894	0.67%
Tamang	1,001,533	4.99%
Total	3,144,008	15.68%

Indo-Aryan Group of Languages

Language	**Number**	**%age**
Awadhi	414,849	2.07%
Bhojpuri	1,527,805	7.62%
Dunuwar	26,267	0.13%
Maithili	2,427,161	12.10%
Nepali	10,301,376	51.36%
Rajbangsi	94,741	0.47%
Tharu	1,100,010	5.48%
Total	15,892,209	79.24%

Other languages

Total	1,019,415	5.08%

Figure 2. The number of speakers of the major languages of Nepal (source National Research Associates, projection for 1996 from 1981 and 1991 censuses)

English, the dominant world language (eg Crystal 1997), is often used as the language of business and politics, but is only spoken by about 5% of the population. Many people are bilingual or even trilingual, possibly speaking a minor language as mother tongue, and also Hindi and the state official language. About 60% of the population of the region is literate, people might be illiterate in their mother tongue but literate in Hindi. Note that Urdu and Hindi are essentially the same language, though written in different writing systems (Kachru 1987) – together they are second only to Chinese in number of mother-tongue speakers. .

Only about 100 of the languages have an established written tradition. Most of these use writing systems derived originally from the Brahmi alphabetical system which arose a few centuries BCE. Some authorities claim that the Brahmi system was invented independently, others that it derived from the earlier Semitic systems. The languages of Pakistan and Afghanistan, including Urdu and Pushto, are written in Arabic-derived writing systems. While printing using type setting came to India in the early 1800s, Urdu was only printed using type for the first time in the 1980s.

However not all languages are equal. While in India there are no prohibitions about the use of languages in the media, 22 languages are given the special status as official languages named in the constitution, favoured with financial support for technology investment. In Nepal all languages have been viewed equally since 1991, but the previous two centuries of favouring Ghorkali/Nepali under the slogan of ‘one nation, one culture, one language’ has left a legacy of Nepali dominance, with other languages still lagging far behind in the resources to support them. In Pakistan language policy has been to favour Urdu over all other languages even though Urdu is a minority language spoken by only 7.5% of the population. In Bangladesh it is Bangla that is favoured over all other languages. Minor languages are often completely overlooked in accounts of languages of the region.

It is not simply a matter of treatment under the law that determines a language’s fate, there are significant social processes that are critical. The languages chosen by the powerful elite can bias people towards preferring those languages, and while this has usually meant favouring the official and national languages, is has often also meant a preference for English. One particularly surprising statistic from Pakistan is that Punjabi is by far the most used language, and yet it has low status and little support. While the people are drawn towards these languages of power, the elite themselves might use political and financial means to suppress other languages. This includes the level of technology provision to support these languages - the powerful languages get the support, the others may not.

This situation can be compared with that in Europe, where there are also many minority languages, both indigenous and migrant, which are also inadequately supported with technology. Many languages native to Europe have already disappeared. English is a ‘killer language’ and Swedish and other national languages may themselves be endangered as people move to English. With thousands of years of conceptual development stored in a language that would be lost if the language was lost, we need to distinguish between museum preservation and sustaining a language in use. Allwood (2004) has suggested that we need ‘language survival kits’, that language technology may be an important part of these kits.

Eisenlohr (2004) has emphasised the importance of a language being able to participate fully in technological developments to ensure the continued vitality of that language. We thus now look at information technology and how languages and their speakers stand with respect to access to IT.

3. Information Technology and its Power

In the West we now see ourselves as knowledge economies and information societies where computers and communications are centrally important – for example Castells (1996) has given a very comprehensive account of this shift. We believe that it is important that developing economies also become based on the use of computers and communications, on software and the Internet, on knowledge and information. This led to the World Summit on the Information Society (WSIS) last December in Geneva, with discussions still going on round the world based on the Plan of Action that came from that meeting. The EU responded to the WSIS meeting by declaring “information and communication technologies (ICT) are among the most important contributors to growth and sustainable development" and the "challenge is to make ICT available and affordable". But how can this be achieved?

At the moment there are the digital haves, and the digital have-nots. This is the digital divide. If we passionately believe that ICTs are good for us and good for everybody we need to understand this divide in all its facets, and through that understanding remove that divide.

The first component of the digital divide is economic, the cost of both communications and computers. The division between the rich and the poor in the world is enormous, and growing: this happens between communities but also within communities. When we think of the gaps between countries we usually contrast the developed countries (DCs) of the industrialised "West" or "North" of Europe and North America and Japan, and the least developed countries (LDCs) of the "South" of Africa, Central and South America and much of Asia. The UN's Human Development and World Development Reports (UN 2000 and 2001) show that the per capita GDP in 1997 was an average of $19,285 in DCs; but $3,610 across the world and only $245 in LDCs. A poverty line of $1 a day puts 84.6% of the population of Zambia below the poverty line – the per capita GDP was $300. Similar division between the richest and poorest exist within countries, with the strongest contrasts in South America.

What can we do about it? While basic needs go unmet, how can they contemplate ICTs? You cannot eat computers. One way is simply to redistribute wealth, through graduated taxation and benefits programmes within countries and through aid programs between countries. However such redistribution must be accompanied by programs which lead to self-sufficiency, to national development. Most international aid is aimed at making other nations become just like us, to develop countries to that they become western capitalist democracies. Such aid programs view development as a linear process like that undergone in the west, but with some potential for leapfrogging particular stages in this process.

While most countries have telephone cable networks, most people do not have telephones – those networks were set up to serve the elite in the major cities – the rural poor are not provided for. But there is hope here for leapfrogging using wireless technologies, and mobile phones are increasingly popular. In Bangladesh there are thriving businesses based on using mobile phones like phone boxes, using Grameen micro-credit to get started. The prospect for communications looks good as wireless technologies increase their coverage world-wide.

The prospect for computers look less good. While under Moores law the cost of computer processors and RAM memory has been rocketing down, software providers have escalated their software's "capabilities" to require the extra power and plan the software and hardware's obsolescence. Further, the reductions in the cost of CPUs and RAM has not been matched by reductions in the cost of displays and hard drives. Several projects have been established to reduce the cost of computers, for example the Simputer with Linux in India aimed at a palmtop for $200, the recently announced PCtvt from Raj Reddy with MS Windows for £250, the Jhai Foundation. Many organisations will ship discarded computers from DCs to LDCs, western cast-offs. And, to make things worse, often the environment in developing countries is less friendly – heat, dust, high humidity, intermittent power supply. The SOLO computer was developed to overcome this – this is a great system, but it does cost more initially, though would claim a lower total cost of ownership.

The cost of software has been a barrier, particularly the operating system without which nothing else can be done. To escape from the commercial demands of Microsoft, open-source 'free' software is proving popular, with the Linux movement supported globally. While Microsoft has responded with new reduced and cheaper Windows product, this has been strongly criticised by the Gartner Group. Other basic software is also available in open source and proving very popular. Free from commercial constraints, this does seem to be the way forward.

UN records show that while 59% of the US population has access to computers, in Zambia the figure is 0.67%, with 35% of the US having access to the internet contrasted with 0.19% in Zambia. These figures assume personal ownership, as in the individualistic West, but many communities prefer shared communal ownership, accessing ICTs through telecentres, or sharing PCs and internet addresses. The real access might be ten times what the raw figures seem to portray.

Historically the internet has been dominated by Websites in English, these are increasingly being developed in other languages, but only in a few other languages such as Japanese. Noronha (2004) reports that in India there are 1.5 million websites in English, but only 20,000 in Indian languages. He notes "The new technology has also made us slaves of a new type of 'information colonialism'. One which encourages us to think that the centre of the world is somewhere in Europe or the US."

What we see in information technology is not good for minor languages and economically disadvantaged communities. Just how easy is it to enable technologies for these minor languages?

4. Localisation as Unicode plus a bit

Localisation has focused on the user interface, on text and its translation, and on the presentation of numerical data. It has now become a very sophisticated industrialised process aimed at shortening lead-times and reducing costs. There are many books now available covering this, for example the general book by Esselink (2000) and the platform specific text by Schmitt (2000). There also are many very sophisticated tools to assist in localisation, and workflow and content management tools to streamline the process.

The computer platform needs to be prepared for the writing system (also known as the script), if it is not already there. For current platforms this requires that the script is represented in Unicode, that there is a rendering engine available for the script, and that there are a number of fonts readily available. While the Unicode script may be used by several languages, the fonts may need to be language specific. For example the Devanagari script of Hindi is also used for Nepali and several other languages, however the style of writing sometimes uses different letter forms and has a different visual appearance that needs to be captured in language specific fonts. For an example of just how difficult it can be to produce a font for a complex writing system, see Hussain's account of Urdu (Hussain 2004).

However the language may not be written. Where languages are not yet written, well meaning people from outside may create a writing system for the language, but often have only one use in mind – a researcher may wish to describe the language as part of a PhD project, a missionary may wish to translate the bible. Often several people develop competing writing systems, so for example in South Africa where the languages are written in a Roman derived system, there are several orthographies for Tsutu which differ as a result of the linguistic background of the originators – English or Dutch missionaries. We need some systematic method for creating new writing systems, but this can be a complex process which is largely undocumented; van Dyken and Lojenga (1993) illustrated the deep knowledge of language structure that is needed. In addition in doing this today we would also create an encoding in Unicode for the writing system, since any account of the new writing system would need to be produced digitally. This writing system and its Unicode encoding need to be subject to standardisation.

If the language is already written, but has not yet been encoded in Unicode, the situation is rather similar to the above. It is not a simple process of looking at the character set, collation sequences, etc, and encoding the characters, a much more subtle investigation may be necessary, paring the process right back in order to validate the candidate writing system(s). We are seeing much of this debate currently in South Asia, particularly in Dravidian language communities who find the encodings developed in the north of India as part of the ISCII standardisation do not adequately represent their writing system.

Even when languages are written, not everybody may be able to use the written form, they may be illiterate. An illiterate person is 'text-blind' but can see its layout and see the graphics. They could access the written information through Text-to-Speech generation, and here global initiatives are under way (Tucker and Shalonova 2004). To be able to generate speech we need a large database or corpora of typical speech including transcriptions of that speech. We could also in principle enable contribution of written information through speech recognition, though this is a much harder problem. Why don't we base much more use of the technology on speech? We would not need a written form, speech recording and playback works for all languages.

Once the writing system has been enabled, localisation of particular software packages can begin. The software will initially be internationalised by separating all language text that will need translation from the rest of the software, and by identifying input and output routines for numbers, dates, currency etc and using standard procedure calls for these from some standard API such as ICU (IBM 2002).

Different communities may have distinct conventions concerning how they write numbers and numerical data like dates and money. For example, in South Asia it is usual to segment large numbers not by thousands, millions, thousand-millions, and so on using powers of a thousand, but in terms of lakhs and crores, 100 thousand and 10 million. No major commercial platform handles this.

While most of the world has moved to the Christian Gregorian calendar, some countries like Saudi Arabia stay with the Islamic Hejira calendar, Nepal stays with the Hindu Vikram Sambhat, and Japan stays with dates within eras determined by the current Emperor. Calendar systems are important for they celebrate important events in the life of a nation or religion. Converting between them is not simply a matter of recalculating years from a different starting point: the Hejira calendar is lunar and not solar, and the start of key months is determined empirically and not by astronomical calculation; Hejira days start at sunset, Vikram days start at sunrise.

Colours may also be important, indicating totally different things in different cultures. Red for danger in the west is a colour of celebration in China, mourning and death is black in the west but white in the east.

More subtle are the designs of web-pages, which are different in different cultures, not just in colour and in choice of icons, but also by what is seen as important in that culture – so for example a bank might also give a customer the opportunity to make donations to charities.

Our cultural values and practices also get embedded in our software – so if we expect software produced in one country to fit another we will be disappointed. For example, ERP systems are usually claimed to embed best industrial practice but Jose Abdelnour-Nocera (2004) has found that even between northern Europe and Spain there can be differences in business practices that make systems unacceptable.

Some applications need further knowledge about languages, and the use of language resources. Spell checkers in word processors and search engines for the web need dictionaries, and possibly also the morphological rules of the language. Language teaching would benefit from language corpora. As we have already seen, speech generation needs speech corpora.

We see that when we localise software we bring to bear a wide ranging knowledge about the language and also its culture. This moves us from Unicode into language engineering, and beyond.

5. Sustainability and Language Engineering.

In any development programme, we must be concerned about what happens when that programme is over and the flow of aid and help from us to them is no longer available, They need to be able to help themselves, the changes we have introduced must then be sustained by the beneficiaries of that aid.

This is also the case with making ICTs available. Even in localising software for a new language we need competent translators who have access to well established dictionaries and authorities on language structure and cultural conventions for numbers and similar. We need computing platforms enabled for the writing system in use, and that in turn requires local authorities on how the language is written, and how it ‘should be’ written, its orthography. Scholars in distant western universities do not suffice, we need local scholars in local institutions.

We also need the capability to produce original software entirely for use locally, we need a local software industry.

To illustrate what is intended by sustainable development, consider the project that a group of us have recently had accepted for funding by the European Union. The project is Nepali Language Resources and Localisation for Education and Communication (NeLRaLEC) and aims to move Nepal forward to approach the position with respect to their national language Nepali that we would take for granted for any major western language. We aim for Nepal to being self-sufficient with respect to software localisation and language engineering, even to act as a leader in the region. The principle partner is Madan Puraskar Pustakalaya (MPP) who manage the archival library for Nepali, working closely with linguists in Tribhuvan University. They will draw upon expertise in Europe in corpus linguistics, dictionary compilation, speech technologies, software localisation.

At the moment access to ICTs in Nepal is predominantly through shared computers and shared internet addresses. There is some public access through Internet Cafes which are widely spread. Yet currently English is the only language in which access is readily available, which mitigates against any widespread use of ICTs in business and administration and education which takes place in Nepali and other local languages. This effectively excludes from the information society all those who are not literate in English.

People in Nepal have an active interest in moving to using computers through the national language, Nepali. Soon after PCs became available in Nepal, desk top publishing with ad hoc capabilities for Nepali became freely available, but these non-standard formats were severely limited in use. By the mid 1990s it became appreciated that to exchange information in Nepali, particularly via the Internet, required conformance to standards. Yet by this point there was already a proliferation of bespoke, non-standard encodings of Nepali, which made the task very difficult indeed. In response to this problem, in 1997 a number of activists - publishers, software developers, academics, civil servants and ex-patriates - drafted a proposal for a Unicode standard for Nepali and acquired funding from IDRC for initial development of fonts and code converters. This work was picked up by MPP, one of the partners in this project, and successfully completed. This is being further developed at MPP (Chalmers and Gurung 2004) based on a second grant from IDRC, this time as part of a regional Pan Asian Networking (PAN) project based in Pakistan (see. www.PanAsia.org.sg or www.crulp.org).

PAN will take the provision of ICT support for Nepali through to a version of Linux and Open Office localised to Nepali, together with a number of language facilities to support computing in Nepali – simple versions of a lexicon, thesaurus, spell checker and grammar checker. This is still a long way short of what we would take for granted for any major European language, and indeed for most of the minor languages like Catalan and Welsh. As has been shown by previous surveys of language processing technology, South Asia has lagged behind as a region in the development of the resources, standards and software necessary develop such utilities (McEnery, Baker and Burnard 2000). While this situation has been partly addressed of late for Bangladesh, India, Pakistan and Sri Lanka by joint European/South Asian research projects such as EMILLE (McEnery, Baker, Gaizauskas and Cunningham 2000), the situation for Nepali has remained unchanged. In a region which had fallen behind in language processing research, Nepal has continued to fall farther behind while the rest of the region has moved forward towards catching up with East Asia, and even Europe and the US.

While Nepali on the computer is in need of help, Nepali produced independently of the computer is also in need of support. Setting aside the issue of machine readable dictionaries and focusing simply on paper dictionaries, we find that the situation is poor - the only official dictionary of Nepali was produced in 1983 (2040 Bikram Sambhat) and contains a mere 30,000 words, with many serious omissions and errors. Plans during the 1990s to update and extend the dictionary to 100,000 words ran into funding difficulties after only those entries beginning with the first letter of the alphabet had been completed. In the wake of that project little has been done though isolated efforts have continued, such as student project work at CIIL in Mysore, India, which produced an experimental Nepali dictionary of a few thousand words. Additionally there are rumours that a commercial project to produce a dictionary may be launched; however it is unlikely that such a project would produce a state-of-the-art dictionary as such dictionaries – such as the Oxford English Dictionary or La Petit Larousse – are invariably based on corpus data. Given that there is no corpus of Nepali, it is difficult to see how any serious, modern, dictionary building could be undertaken for Nepali.

While there is considerable expertise in linguistics in Nepal, there has been a lack of resources to develop this further. As a first step NeLRaLEC will gather text and speech corpora to form the Nepali National Corpus, modelled on the British national Corpus. We do not anticipate any difficulties in obtaining access to material in order to build our corpora, though we recognise that not all of the material will be machine readable and some data entry will be required. There is a shortage of suitable trained people and we will need to undertake training in order to carry out the work. However we see all of this as a necessary part of the task of technology transfer.

The corpora will facilitate the construction of a state-of-the-art dictionary of Nepali: dictionary compilation is expensive and laborious, and simply using the funds granted to the project we do not expect to complete the dictionary, but to have made a good start that can then continue afterward. We will have produced a word list sufficient for a high quality Nepali spell checker.

We will also use the corpus to build speech generation capabilities, using the technologies developed by the Local Language and Speech Technology Initiative (LLSTI) funded by the UK's Department for International Development (see Tucker and Shalonova 2004).

As we develop the language resources for Nepali, we will also be developing local language software, further fonts, facilitating the development of a Nepali localisation industry. We will draw upon expertise from the Localisation Research Centre in Ireland and knowledge there of how the localisation industry in Ireland has grown and thrives.

Our software developments will enable us to be able to evaluate our technology in trial applications in schools in Nepal. Schools in Europe would be expected to use computers both directly in classroom teaching and indirectly in the preparation of teaching materials. We want to be sure that while we cannot remove the economic barriers, our technology will at the least have removed the language barrier which stops this happening in Nepali.

As a final step we are aiming to facilitate the introduction of a Nepali Language Engineering masters degree programme at the Tribhuvan University in Kathmandu, based on mature Masters programmes in the UK and Sweden. The corpus and other resources will be an important asset here, and graduates from this degree will help deepen knowledge about Nepali and aid the long term future of local software and language engineering in Nepali. The technology for using the computer in Nepali will also be of great value within the university at large as it moves from teaching in the medium of English to teaching in the medium of Nepali.

Note that while we have focused on Nepali, the dominant language of Nepal and neighbouring regions, there is a case for us devoting equal effort on supporting smaller and more endangered languages. Nepali is almost universally spoken in Nepal, but all together there are over 100 languages. 5 of these have mature written traditions, but more than 60 are completely unwritten. Education in languages other than Nepali has been permitted since 1990 and since the introduction of 'education for all' in 2003 teaching in the mother-tongue is now required. There is much more to be done in Nepal.

6. Conclusions.

The majority of the world's population still has no access to computers and the Internet. Much of this is due to economic constraints, the communications infrastructure has not yet reached them, and where it has they cannot afford the connection charges or the equipment needed. Where physical access has been achieved, there are still major barriers. The interfaces and data must be in the language of the user, preferably their mother tongue, but otherwise in their local official language. The user may not be literate, and yet may need to acquire knowledge from the Internet, and even supply knowledge to the Internet. The conventions used for presenting information to the user, and the methods for doing things, must fit into their own culture.

We argued that a failure to support a language with technology could lead to the disappearance of that language. Supplying technology for the language could revitalise it. However it is not enough to simply enable the technology for the writing system of the language, nor even to also provide basic localised software, we must enable the community to make these developments itself. This means training and educating local people in language engineering methods and techniques, and establishing academic programmes to build up expertise in language engineering for the language.

This process is expensive, even for just one language. Most nations have a number of languages, and really we ought then to be developing language resources and software for all those languages. However to do this could be far too expensive and we will need to make compromises. This means, regrettably, that we must anticipate losing a substantial proportion of those 6000 languages. But lets revitalise as many as we can, and improve upon the 600 surviving languages anticipated by Nettle and Romaine.

7. References

Abdelnour-Nocera, Jose; and Pat Hall (2004) 'GLOBAL SOFTWARE, LOCAL VOICES. The Social Construction of Usefulness of ERP systems', *CATaC 2004 conference*, Sweden, July 2004.

Allwood, Jens; 'Linguistic Diversity and the Digital Divide' in SCALLA 2004

Castells, Manuel (1996) *The Rise of the Network Society*. Blackwell

Chalmers, Rhoderick and Amar Gurung; (2004) Localising software: some experiences from Nepal' in SCALLA 2004

Crystal, David (1997) *English as a Global Language*, Cambridge University Press

Comrie, Bernard (Editor) (1987) *The Major Languages of South Asia, the Middle East and Africa*. London: Routledge.

van Dyken, Julia R. and Constance Kutsch Lojenga (1993) 'Word Boundaries: Key Factors in Orthography Development' in Hartell 1993.

Eisenlohr, Patrick (2004) 'Language Revitalisation and New Technologies: Culture of Electronic Mediation and the Refiguring of Communities'. *Annual Review of Antropology 2004*. pp21-45

Esselink, Bert (2000) *A Practical Guide to Localisation*. Amsterdam and Philadelphia: John Benjamins

Hartell, Rhonda L (Editor) (1993) *Alphabets of Africa* Unesco-Dakar and Summer Institute of Linguistics.

Hussain, Sarmad; Complexity of Asian Writing Systems: A Case Study of Nafees Nasta'leeq for Urdu in SCALLA 2004

IBM (2002) International Components for Unicode (ICU) <http://oss.software.ibm.com/icu/userguide>

Kachru, Yamuna, (1987) 'Hindi-Urdu', Chapter 3 in Comrie.(1987)

Mallikarjun, B; (2004) Indian Multilingualism, Language Policy and the Digital Divide, in SCALLA 2004

McEnery, Tony and Andrew Wilson (2001) Corpus Linguistics. 2nd Edition. Edinburgh University Press.

McEnery, A, Baker, JP and Burnard, L (2000). 'Corpus Resources and Minority Language Engineering.' In: M. Gavrilidou, G. Carayannis, S. Markantontou, S. Piperidis and G. Stainhauoer (eds) *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC)*: Athens.

McEnery, AM, Baker, JP, Gaizauskas, R and Cunningham, H (2000) 'EMILLE: Building a Corpus of South Asian Languages.' In: *Vivek, A Quarterly in Artificial Intelligence*, 13(3):23–32

Nettle, Daniel and Suzanne Romaine (2000) *Vanishing Voices, the extinction of the world's languages* Oxford University Press.

Noronha, Frederick (2004) in www.i4donline.net June 2004

Rahman, Tariq; (2004) Language Policy and Localization in Pakistan: Proposal for a Paradigmatic Shift . in SCALLA 2004

SCALLA (2004), conference in Kathmandu Nepal supported by the EU through the SCALLA project, papers through ELDA in Paris. http://www.elda.fr/proj/scalla.html

Schäler, Reinhard ; A Framework for Localisation in SCALLA 2004

Schmitt, David (2000) *International Programming for Microsoft Windows*. Microsoft.

Tucker, Roger and Ksenia Shalonova; (2004) The Local Language Speech Technology Initiative – localisation of TTS for voice access to information in SCALLA 2004

Turin, Mark; (2004) Minority Language Politics in Nepal and the Himalayas in SCALLA 2004

Unicode Consortium (2003) *The Unicode Standard, Version 4.0* Reading, MA, Addison-Wesley, 2003. ISBN 0-321-18578-1)