The ELRA Newsletter

October - December 2004

Vol.9 n. 4

Contents

Signed articles represent the views of their authors and do not necessarily reflect the position of the Editors, or the official policy of the ELRA Board/ELDA staff.

ISSN: 1026-8200

ELRA/ELDA
CEO: Khalid Choukri
55-57, rue Brillat Savarin
75013 Paris - France
Tel: (33) 1 43 13 33 33
Fax: (33) 1 43 13 33 30
E-mail: choukri@elda.org
Web sites:
http://www.elra.info or
http://www.elda.org

Dear Colleagues,

This is the last issue of 2004 and we would like to quickly highlight a number of topics on which ELRA focussed during this year.

The major event for ELRA is the organisation of LREC 2004, the fourth edition in the series of Language Resources and Evaluation Conference launched by ELRA with the support of a large number of active players in the field. LREC 2004 was a very successful event, which attracted more than 900 participants with over 500 papers and 8 workshops. Another important milestone in the life of the association is the organisation of a "strategic meeting" attended by most of the board members to discuss and finalize the ELRA vision of our domain for the years to come. Among the issues debated during that meeting, we would like to highlight a few ones:

• The board refined the mission and goals of ELRA; this has been revised to consider the evolution of our area during the first decade of ELRA activities. In particular the mission considers the new sub-domains and the new trends (e.g. multimedia, multimodal, video, images). Such extension will be incorporated in the association statutes that will be submitted to the General Assembly for revision and adoption in 2005.

• The membership policy constituted another major debate topic. The board considers of outmost importance to serve its members with new services. The main issue was how to attract new members but also how to reward the loyal ones. The idea is that members joining ELRA and buying resources will accrue LR miles that can be used interchangeably to acquire free membership, free registration for LRECs or extra discount on language resources. The details of such fidelity plan will be announced before the General Assembly.

• The board focused on issues related to the technical activities with respect to LRs: e.g. identification/collection of LRs, distribution of LRs, production and validation of LRs. The leading role of ELRA as a coordination and focal point for such activities is stressed and actions are planned to boost them (e.g., put more emphasis on the work being done on the Universal Catalogue of LRs, more budget allocated to production of new resources...).

• The board also addressed the issue of Language Technology Evaluation. ELRA is involved in evaluation and acts as a booster of a number of evaluation campaigns by supplying the right infrastructure to support the technical institution willing to carry such evaluation. ELRA is capitalizing on such evaluation initiatives to offer a new catalogue of "Evaluation Packages". The board decided to set up an HLT Evaluation portal that would give a clear picture of the past, present and future evaluation initiatives of interest to our members. In addition to these issues, ELRA board emphasized the role of coordination and "synergy", it would like to play for the mutual benefits of all HLT parties.

During this quarter, ELRA and ELDA continued their works on a number of projects funded internally or funded/supported by funding agencies (e.g., French Technolangue Program, FP6 of European Commission), in particular, a survey on "Landscape of LR Scene" has been completed during this quarter with over 30 responses that will help us fine-tune our future activities. Such report will be formatted for distribution via our website to our members and the participants soon.

New resources have been secured for distribution. These are announced in the last section of this newsletter and consist of:

- S0167: SALA II Spanish Mobile Network Database collected in Venezuela
- S0168: French Speecon database
- S0169: Hebrew Speecon database
- S0170: BABEL Romanian database
- S0171: SALA II Spanish Mobile Network Database collected in Mexico
- S0172: C-ORAL-ROM

As for this newsletter it contains a description of the WISPR project ("Welsh and Irish Speech Processing Resources") which is the first collaboration between Welsh and Irish researchers to develop speech technology tools and resources for Welsh and Irish.

It also contains a description of BIOMET: a Multimodal Person Authentication Database Including Face, Voice, Fingerprint, Hand and Signature Modalities. These so-called "biometric" modalities can be used for projects related to speaker identification/verification. The database will be available soon via ELRA.

Finally, the SCALLA conference "Crossing the Digital Divide" is described as well. The SCALLA project aimed at increasing collaborations between European and South Asian players in human language technologies and software localisation.

Once again if you would like to join ELRA and benefit from its services (that are summarized at www.elra.info), please contact us.

Bente Maegaard, President

Khalid Choukri, CEO

Speech technology in Welsh and Irish: the WISPR project

Briony Williams, Delyth Prys, Dewi Jones

Introduction

The WISPR project ("Welsh and Irish Speech Processing Resources") is the first collaboration between Welsh and Irish researchers to develop speech technology tools and resources for Welsh and Irish. The initial focus will be on providing the infrastructure for TTS in both languages. It is funded by the Interreg programme of the EU, together with the Welsh Language Board, and runs until the end of 2005.
The main partners are the University of Wales, Bangor and Trinity College, Dublin. There is also additional input from Dublin City University, University College, Dublin, and Institiúid Teangeolaíochta Éireann (ITE). Together, the Irish partners are active in the "Irish Speech Group" (see http://isg.eeng.may.ie), an informal network of speech researchers in the Irish language.

The requirement for tools and infrastructure to build speech technology applications is particularly pressing in the case of minority languages. This is because the languages themselves are threatened to varying degrees. Industry is unlikely to provide the necessary resources, due to the lack of sufficient commercial return. Furthermore, the leverage represented by speech technology is proportionately much greater for minority languages: these very tools have the potential to assist in stemming the decline in language use.

However, very little work has been done to date on developing speech technology tools for the Welsh and Irish languages. There are no existing usable software packages either for speech synthesis or speech recognition for either language. Both communities suffer through this lack of resources, and applications for communications and spoken information exchange cannot proceed for these languages without the development of speech technology tools. In addition, Welsh-speaking blind people are exceedingly eager to obtain a Welsh-language screenreader. To date, no such software exists, and developing a basic screenreader is one of the early objectives of the Welsh WISPR team.

Both partners believe that the best way to disseminate speech technology tools in a minority language environment is to provide freely distributable and unlicensed tools that are easy for the end user to adopt. These may also be taken up for further development by businesses to incorporate into other applications. To this end, it was decided to adopt the "Festival" speech synthesis framework developed initially at the University of Edinburgh by Alan Black and others (Black and Lenzo 2000). This software provides a complete development environment for concatenative speech synthesis, covering both diphone and unit selection (variable-unit) voices.

Current and future progress

The Welsh team benefits from earlier work in Welsh text-to-speech synthesis which produced a diphone-based voice from a male South Welsh speaker (Williams 1994, 1995). Although the quality of the voice is below current standards, many of the resources developed during the earlier project can be built on: e.g. the phoneset, diphone recording script of nonsense words, and (importantly) the letter-to-sound rules. In addition, the team can draw on speech resources produced during an earlier project that aimed to develop a phonetically-annotated Welsh speech corpus (Williams 1999). The following progress has been made.

1-Welsh

In October 2004, the Welsh WISPR researchers released an alpha-level version of an MSAPI interface to the existing Welsh diphone voice running in the Festival Speech Synthesis System. This was released to selected user groups for initial usability testing. Some minimal MSAPI support was included for basic "speak" functionality and for varying the speech rate, since early feedback from blind users suggests that they prefer to speed up the rate of a screen reader to a far greater degree than other users would be comfortable with.

The speech quality is not optimal, but the alpha release will facilitate the gathering of early user feedback, which can then be used to guide and prioritise work for both the optimisation of the existing Welsh diphone voice, and also the further development of an MSAPI interface to Festival.

The MSAPI synthesised voice has been successfully integrated into various Windows applications, such as a simple speech-enabled word processor called EdWord, developed for dyslexic users (see www.sense.org.uk/nof/software.html). Integration with commercial screen reader software is planned.

Other future directions for 2005 include the development of a unit selection voice, and the development of a voice for North Welsh

(which has a larger vowel inventory than South Welsh). As the older voice was that of a male speaker, the new voice will probably be a female one.

2-Irish

In contrast to the Welsh WISPR team, the Irish researchers have no previously developed resources, and their work centres on producing basic infrastructure elements. A first task on which they are working is to adapt to a specific dialect the "Foclóir Póca", a standardised pronunciation dictionary for Irish. This would produce the first such resource for Irish in digital form. They are also in the process of developing letter-to-sound rules for the Connemara variety of Irish, using the Classification And Regression Tree (CART) functionality included within Festival and the Edinburgh Speech Tools. Since Irish orthography is very far removed from the pronunciation and hence not transparent, it was not considered feasible to develop letter-to-sound rules by hand, as had been done earlier for Welsh. Instead, the automatic statistically-based CART approach will be used.

Many of the tasks involved will have spin-offs in the area of education and special needs applications for Irish. For example, having a dialect-specific pronunciation facility will be of great value in language teaching. It is hoped that the resources produced in this project might form the basis for future software applications.

Other work in speech technology for the Celtic languages of the UK

1-Welsh

As part of the EU "SpeechDat" program, a small amount of telephone speech data in Welsh was collected from a wide range of subjects, using a highly restricted vocabulary (Jones et al. 1998). However, this work was not further developed, so far as is known. As regards text-to-speech synthesis for Welsh, a hardware synthesiser capable of synthesising both Welsh and Irish via separate language ROMs is produced by Dolphin Computer Systems Ltd as the "Apollo 2" synthesiser (see http://www.dolphinuk.co.uk). This is a separate mains-powered unit used in conjunction with a desktop computer. In addition, researchers at the University of Manchester Institute of Science and Technology have produced a set of letter-to-sound rules for Welsh, sending the output phones to an English synthesiser for sound production. This approach inevitably displays phonetic mismatches, as there are several sounds in Welsh which have no exact equivalents in English (e.g. the voiceless lateral fricative). However, in the absence of a purely Welsh synthesiser this approach has proved helpful to many Welsh blind people. The same approach (based on an English synthesiser) has been used by UMIST to synthesise several other languages (Evans et al. 2002).

2- Scottish Gaelic

A diphone-based text-to-speech synthesiser for Scottish Gaelic was produced using the "Festival" synthesis framework at the University of Edinburgh (Wolters 1997). This work has not been further developed. However, as Gaelic is within the same language sub-family as Irish, it may be possible for the Irish WISPR researchers to adapt some of the strate gies used for this work.

Looking to the future

It is hoped that the WISPR project will be merely the first step in developments for speech technology applications in the Celtic languages. There is a need for speech recognition software for Welsh and Irish, ideally in more than one variety of each language (e.g. versions for both North and South Welsh, which differ in the vowel phonemes). Also, it is to be hoped that specialist software providers might, in the future, be willing to collaborate in the production of Welsh and Irish speech applications.

With the advent of the Web has come the opportunity for interactive speech-enabled websites, using synthesised speech for educational or leisure purposes. Minority languages in particular have a need for widely-disseminated educational tools, both teaching the language itself and also teaching other subjects through the medium of the language. Speech-enabled websites would form invaluable educational tools for enhancing language skills. They would also help to ensure the continuance of the language as a fully-fledged modern language, capable of functioning even in the most technological of contexts. For this reason, speech technology tools and resources are a vitally important step in the evolution of minority languages today.

References

Black, Alan and Lenzo, Kevin (2000): *Building voices in the Festival speech synthesis system*. Carnegie-Mellon University.

Evans, D.G.; Polyzoaki, K. and Blenkhorn, P (2002): *An Approach to Producing New Languages for Talking Applications for use by Blind People*. In: K. Miesenberger, J. Klaus and W. Ziegler (eds.): *Computers helping People with Special Needs"*, Proc. 8th ICCHP, Lecture Notes in Computer Science, No. 2398.

Jones, Rhys James; Mason, John S.; Jones, Robert Owen; Helliker, Louise; Pawlewski, Mark (1998) *SpeechDat Cymru: A large-scale Welsh telephony database*. Workshop on speech and language technology for minority languages, Language Resources and

Evaluation Conference, Granada, Spain.

Williams, Briony (1994): *Diphone synthesis for the Welsh language*. Proceedings of the 1994 International Conference on Spoken Language Processing, Yokohama, Japan.

Williams, Briony (1995): *Text-to-speech synthesis For Welsh and Welsh English*. Proceedings of Eurospeech 1995, Madrid, Spain.

Williams, Briony (1999): A *Welsh speech database: preliminary results*. Proceedings of Eurospeech 1999, Budapest, Hungary.

Wolters, Maria. (1997): *A Diphone-Based Text-to-Speech System for Scottish Gaelic*. MSc thesis, University of Bonn. http://citeseer.ist.psu.edu/309369.html

Briony Williams, Canolfan Bedwyr, University of Wales, Bangor,UK,
Email: b.williams@bangor.ac.uk

Delyth Prys, Canolfan Bedwyr, University of Wales, Bangor,UK,
Email: d.prys@bangor.ac.uk

Dewi Jones, Canolfan Bedwyr, University of Wales, Bangor,UK,
Email: d.b.jones@bangor.ac.uk

BIOMET: a Multimodal Person Authentication Database Including Face, Voice, Fingerprint, Hand and Signature Modalities

GET (Groupe des Ecoles de Télécommunication)

Introduction

Every individual has characteristics unique to him/her: voice, fingerprints, facial features, hand shape, signature ... right through to the DNA. These so called "biometric" data can thus be used to identify the individual. Whilst such methods have been used in the past by the police, it is now necessary for an individual to be identified in many contexts: to access an apartment block or open the door to his/her apartment, withdraw money from a cash machine, access buildings and move around freely, access his/her workstation, email, the Internet and, more generally, anywhere security is essential.

Currently, to verify identity, codes, passwords and other personal identification numbers are used, which have two drawbacks:

- they have to be memorized;
- there is the risk that other unauthorized people may use them.

So why not replace them with biometric data?

No single biometric modality is 100% reliable in itself. There are situations, relating to data capture mechanisms, the user him/herself or the environment at the time of capture, in which a given modality may turn out to be flawed.

- the fingerprint of a manual worker make it difficult to identify him/her;
- the loss of the user's voice or excessive ambient sound at the time of voice capture interfere with the verification of the user.

In addition, for a given person, a modality may be too variable, which also prevents identification. Such is the case with signatures, which some people rarely reproduce exactly. This fact has led researchers to combine various biometric modalities to achieve more reliable results and make verification systems more appropriate, faced with a whole range of usage contexts. Beyond the technical reasons outlined, many human factors relating to the way these identity verification systems are perceived by the public come into play in the real utilization phase: having ones fingerprints taken can have police connotations; a given modality may be perceived as an intrusion into ones private life; a specific sensor may be rejected by the public for reasons of hygiene, etc. Faced with these pitfalls, both human and technical, a multimodal approach seems to represent a solution with tremendous potential. It is one of the major objectives of research currently being conducted by GET.

In order to take advantage of the particularities of each modality, and to improve the performance of a person authentication system, multimodality can be applied. The choice of a modality depends on several aspects: the performance of the authentication algorithms, the intrusiveness (acceptance by the user), the ease of implementation, the cost of the capturing devices... In order to study how different modalities can be combined, GET have recorded the BIOMET database, a biometric database with five different modalities.

Database Description

Five different modalities are present in the BIOMET database: audio, face images, hand images, fingerprints and signatures. For the face images, besides a conventional digital camera, we also decided to take 3D pictures in order to study the contribution of face surface characteristics for person authentication.

In order to study the influence of aging, three different sessions, with three and five months spacing between them, were realized. The number of persons participating to the collection of the database was 130 for the first campaign, 106 for the second, and 91 for the last one. The proportion of females and males was balanced for all the campaigns. 10% of the persons were students (with a mean age of

20). The age of the others varies from 35 up to 60 years. There are 3 recording sessions for the hand images and fingerprints. The signature and audio-video data were acquired during the second and third campaign. The 3D images were only recorded during the last session. In the following sections, more details about each modality are given.

Audio and Video Data

A digital camera is used for recording conventional audio-video sequences of a talking person. The capture of the combined audio-video data is done with the person facing the camera, and pronouncing in French: his/her identification number, digits from 0 to 9, digits from 9 to 0, "oui", "non", and 12 phonetically balanced sentences.
After the acquisition of the audio and frontal face video data, the acquisition continues, and the person is asked to turn the head roughly 15° to the left, to the right, up, down, and finally 45° to the left and right, to capture his/her left and right profiles. The average duration of the acquired audio-video sequences is 1.5 min.

Images Acquired with the 3D Acquisition System

Information about a person's identity can also be extracted from its facial surface images. We use a 3D acquisition system to capture most of the facial surface with a natural cooperation of the subject. A prototype has been developed with a high resolution color camera and a low-cost slide projector. It is based on structured light projection, and consists of a camera and a projector. A specific pattern of color lines is projected on the face and serves as reference points in the obtained image. The color sequence of lines allows the determination of each line, and enables the localization by means of triangulation (camera, projector) of surface points in the space defined by the camera-projector axis system. Such 3D images have been acquired only during the last session.

Hand Images

People's identities can also be verified through the information conveyed in the geometry of their hands. The characteristics gathered from the contour of the hand, could be eventually combined with the lines extracted from the hand palmprint, in a fused system. For the BIOMET database, we have used a scanner to capture 2 dimensional images of the hand (Fig. 1 shows an example). A conventional scanner is used to acquire the left hand images of the people participating to the database collection. The person is asked to put his/her left hand on the scanner with the fingers spread naturally, and as flat as possible (in order to capture the palmprints). The resolution of the hand images is 150 dpi, and 5 images of the left hand were captured during the 3 acquisition sessions.

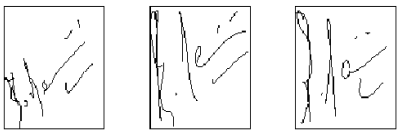

Figure 1: Captured hand image (left), and middle finger fingerprints with the capacitive GEMPLUS sensor (middle), and with the optical SAGEM sensor (right)

Signature Data

Everybody is familiar with signing documents. The paper can be replaced by a digitizing tablet, that can capture the signatures on-line, resulting in the acquisition of the the dynamic informations of the signing process.
The signatures for the BIOMET database are acquired with the A6 (1024x768 pixels) graphical tablet. The following five parameters characterizing the signature are available: the x and y coordinates, the pressure of the writing device, the azimuth and the altitude of the pen. These parameters are captured by the digitizer at a rate of 100 Hz.

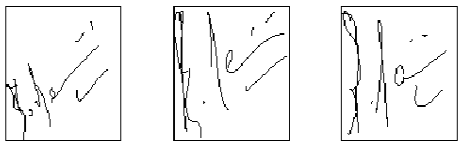

Figure 2: Example of one genuine signature, and two different impostors signatures, acquired with a graphical tablet.

During the second and the third campaigns 15 genuine and 17 impostor signatures per person, with five different impostors realizing the imitation signatures were acquired. Fig. 2 shows examples of one genuine signature, and two impostor signatures, realized by two different impostors.

Fingerprint Images

Fingerprints are known from a while as delivering reliable person specific characteristics. Currently, fingerprints can be captured with sensors of different types, detecting differences between the ridges and the valleys of the fingerprints. We have chosen two commercially available devices based on different types of sensors. The SAGEM Morphotouch 3.0 device is an optical sensor, detecting differences in reflection. The GEMPLUS PC Touch 430 device detects differences in capacitance.
During the three campaigns, we have acquired 6 fingerprints of the middle finger and 6 images of the index finger for each of the sensors. The resolution of these images is 500 dpi. Examples of fingerprint images of the middle finger with both sensors are shown on Fig. 1.

Acknowledgments

This work has been supported by the French "Groupe des Écoles des Télécommunications" (GET), within the research project BIO-

MET. We wish to thank all the participants of this project, with special thanks to Marc Sigelle, who coordinated the first year of the BIOMET project. We also express our thanks to Martine Chollet for her help during the recording campaigns, as well as Geoffroy Fouquier for the realisation of the BIBLOS database capture software.

GET (Groupe de Ecoles de Télécommunication) :
INT, Dept. EPH, Evry, France and
ENST Paris, Dept. TSI Paris, France

The SCALLA working conference " Crossing the Digital Divide "

Pat Hal

This conference was the final conference of the SCALLA project, funded by the European Union under the Asia IT&C programme. The SCALLA project aimed to increase collaborations between European and South Asian workers in human language technologies and software localisation. As well as revisiting computational linguistics for South Asian languages, we considered in greater detail the localisation of software to those languages and the reasons why this is important. As well as technical support for languages we also considered social and policy issues concerned with linguistic diversity. We aimed to identify gaps in current support, both technical and sociopolitical, contrasting the situation in South Asia with that in Europe. This led us to consider what advice and actions might be appropriate to help advance technical support for languages across South Asia, to help cross the digital divide. In this process we also expected to identify advice and actions which in turn should be taken back to help the European Union in its own efforts to remove its own internal digital divide.

Setting the Scene

The keynote address for the conference was given by **Mr. Kanak Mani Dixit,** the member secretary of the Madan Puraskar Pustakalaya, the library of record of the Nepali speaking world. Mr. Dixit spoke of the need for full technological support for languages like Nepali, to enable citizens at large to benefit from the use of Information Technology. He illustrated his argument using Nepal, with its 23 million people most of whom are fluent in Nepali but have little English, who are barred access to computers not by cost since computers have become relatively cheap but by language. But he was optimistic, looking to the example set by the growing use of Radio, which shows such significant evidence of the creative application of technology within the Kingdom.

Language, culture, politics, and the digital divide

The paper by **Dr. B. Mallikarjun** well illustrated this difficulty in counting languages.
What is important to note is that there is a great variety of languages, written in a variety of writing systems. This variety and richness is repeated in other countries, and we heard from **Dr. Tariq Rahman** about the richness of languages in Pakistan with around 70 distinct languages, and from **Dr. Mark Turin** about the 100 or so languages of Nepal. Only in Sri Lanka, presented by **Vincent Halahokonege**, is the variety limited, with just 2 languages.
This situation was compared with that in Europe, where there are also many minority languages, both indigenous and migrant, that are also inadequately supported with technology. Describing this situation **Prof. Jens Allwood** spoke of English as a 'killer language' (quoting Skutnabb-Kangas) and the fear that Swedish may itself be endangered as people move to English.

Linguistics and Language Technology

Prof. B.N. Patnaik suggested that we should start by developing language technology to support tribal languages in order to overcome the 'English divide', the global dominance of English.
This moves away from universal views of language, a theme picked by **Prof. Harold Somers** in the context of Machine Translation which necessarily must use some common view of the two languages being translated - the conclusion from MT is that a universal interlingua is not possible, not even for closely related languages.
Prof. Rajeev Sangal spoke about the development of language resources, corpora and lexicons, for South Asian languages and in particular for Indian languages.
Prof. Yogendra Yadava explained the situation in Nepal, with around 100 languages but where even the national language, Nepali, has very limited corpora and a dictionary that is in urgent need of updating and correcting and extending.
Prof. Pushpak Battacharya then described a machine translation system based on the interlingual system UNL (Universal Network Language). A source sentence in one language is converted into UNL undergoing various normalisations in the process, and then the target sentence is generated by deconversion of the UNL. Studies are being made into English, Hindi, Marathi, and Bengali for application of this UNL-based approach to machine translation.
Dr. Ruvan Weerasinghe reported an investigation into the use of statistical machine translation SMT) applied to the languages of Sri Lanka - English, Sinhala, and Tamil.
Dr. Peter Juel-Henrichsen then described his application of statistical methods to the description of unwritten unfamiliar languages: two different clustering methods were presented, which Battacharya conjectured resembled Wordnet similarity measures.
Dr. Roger Tucker and **Dr. Ksenia**

Shalanova presented their work using text-to-speech (TTS) software to give speech access to information sources, motivated by the need to overcome literacy barriers. They provide a toolset based on diphones and a speech database, give advice on the use of this toolset, and support knowledge sharing among a network of groups using these tools.
Dr. Hema Murthy described work at IIT Madras on multi-modal interfaces, and in particular speech recognition for Indian languages.

Software localisation and local software

Underlying all software that works in local languages are the computer encodings of the writing systems of those languages. In order to clarify the issues here, a short introduction into the history of the encoding of languages and writing system in the computer was given by **Prof Pat Hall**.
The writing of Urdu using the Nastaleeq font was described by **Dr. Sarmad Hussain**. Urdu writing is cursive and multi-directional, with the 36 alphabetical characters changing into 15 to 20 different shapes as a function of context.
Rendering for Indian writing systems came earlier, starting in the 1970s and 1980s, as described by **Dr. S.P. Mudur.** The renderers developed for Microsoft are also available for Linux. Keyboards were also raised in general discussion, and picked up by **Prof B.B. Chaudhuri** in the context of input devices in general. Keyboards should be reconfigurable so that a range of layouts can be used. OCR has been developed in Calcutta and is now being developed into a product at CDAC. Handwriting recognition and speech recognition are not well developed for Indian languages.
Dr. Shailey Minocha talked about human computer interaction and the way assumptions about users become embedded in the software and its interface.
The social reasons for localisation were then picked up by **Venky Hariharan**, who pointed out that only some 10% of India's billion people were competent in English - which means that some 900 million people have no access to computers.
The contrasting situation in Nepal was described by **Amar Gurung** and **Dr. Rhoddy Chalmers**, with around 150,000 PCs mostly assembled locally and mostly without licensed software.
Prof. Reinhard Schaeler explained that 95% of Irelands output was localised US software and digital content, with exports of software from Ireland exceeding the exports of software from the US.

Linguistics and language technology futures.

Language corpora are important for current linguistics, and **Prof Tony McEnery** has considerable experience in this, from gathering material in English in the UK for the British National Corpus and on the EMILLE project gathering material in South Asian languages in both the UK and in South Asia.
Vincent Halakonege recounted similar experience in Sri Lanka where all contact had to be personal and obtaining the support of the top person in an organisation was critical. While the volume of corpus data for South Asian languages is very significantly less that that available for other world languages, the gathering of data has begun.
Prof. Harold Somers described his project to use language technology help health care patients with limited or no English (PLONEs), where all too often untrained people act as interpreters during medical consultations. The current language of study is Somali, with plans to move to Punjabi or Bangla, using speech and icons to interact with the patient.
Prof. Anthony Pym gave us this translator's perspective. Localisation is a major employer of translators, but with the tools available to take care of equivalence, they become more concerned with controlling the source, with internationalisation.

Localisation and local software futures

The Language Observatory project of **Prof. Yoshiki Mikami** was inspired by his travels around Asia and the many forms of typewriter he discovered.
Dr. S.P. Mudur described the work that NCST did for Microsoft in enabling the input and output of Indic Languages at the code level, and the sad experience that this the only worked in Notepad because other Office packages did their own input and output.
Prof. Pat Hall then described the Glossasoft project, in which software had a high-level knowledge model about the software, and used language engineering methods to generate error and help messages, with an API that enabled language specific components to be replaced as required.
The various localisation issues were then brought together by **Dr. M. Sasikumar**, distinguishing between localisation which enabled content creation by encoding the script and providing fonts, and changing the interface and look and feel.

Policy implications.

We then contrasted the support for language technologies in South Asia and in Europe. **Dr. Om Vikas** has been leading TDIL in India, coordinating technology and language engineering developments.
India through TDIL has focused on adapting technology to meet core priorities, then developing its own technologies in collaborative programmes, and now focuses on creative technology within new sustainable structures. Areas of focus have been machine translation, OCR, text-to-speech generation and speech recognition, and many appliation areas like e-government and education. International collaborations are important. TDIL run an information and resource distribution service with a website and newsletter. **Dr. Khalid Choukri** is the CEO of the European Language Resources Association (ELRA), which started in 1995. Europe is a multi-lingual multicultural region and faces many of the same problems as faced in South Asia.

The EU is now funding ambitious projects on speech-to-speech translation and multi-modal interfaces that will produce resources that will in due course be deposited with ELRA. Like TDIL, ELRA runs an information service with website and newsletter.
The final presentation to round off the conference was by **Prof Ken Keniston**, has observed localisation developments in India for more than a decade. Even though there have been significant developments in South Asia of software that handles South Asian languages, there is still little software that is standard compliant or robust enough for sustained use, and projects to remedy this make promises but have yet to deliver.
The solution throughout is standards and collaborations, with resources devoted to support these.

Patrick A.V. Hall
Professor of Computer Science,
Computing Department,
Open University, Milton Keynes
MK7 6AA, United Kingdom
tel: 01908 652694 (work at OU)
email: p.a.v.hall@open.ac.uk
http://computing.open.ac.uk/

NEW RESOURCES

ELRA-S0167: SALA II Spanish Mobile Network Database collected in Venezuela

The SALA II Spanish Mobile Network Database collected in Venezuela was recorded within the scope of the SALA II project.

The database has been collected jointly by the Universidad de Los Andes (ULA) and Applied Technologies on Language and Speech, S.L. (ATLAS) from Spain. The owner of the database is Applied Technologies on Language and Speech, S.L. (ATLAS).

The SALA II Spanish Venezuelan database contains the recordings of 1,179 Venezuelan speakers (576 males and 603 females) recorded over the Venezuelan mobile telephone network.

The following acoustic conditions were selected as representative of a mobile user's environment:

- Passenger in moving car (160 speakers),
- Public place (461 speakers),
- Stationary pedestrian by road side (236 speakers),
- Home/Office environment (272 speakers),
- Passenger in moving car using a hands-free kit (160 speakers).

This database is distributed as 1 DVD-ROMs.The speech files are stored as sequences of 8-bit, 8kHz a-law speech files and are not compressed, according to the specifications of SALA II. Each prompt utterance is stored within a separate file and has an accompanying ASCII SAM label file. This speech database was validated by SPEX (the Netherlands) to assess its compliance with the SALA II format and content specifications.

Each speaker uttered the following items: 6 application words, 1 sequence of 10 isolated digits 4 connected digits (1 sheet number -6 digits, 1 telephone number –9/11 digits, 1 credit card number –14/16 digits, 1 PIN code -6 digits), 3 dates (1 spontaneous date e.g. birthday, 1 word style prompted date, 1 relative and general date expression), spotting phrase using an embedded application word, 1 isolated digit, spelled words (1surname, 1 directory assistance city name, 1 real/artificial name for coverage), 1 currency money amount, 1 natural number, 5 directory assistance names (1 surname out of a set of 500, 1 city of birth/growing up, 1 most frequent city out of a set of 500, 1 most frequent company/agency out of a set of 500, 1 “forename surname” out of a set of 150), yes/no questions (1 predominantly “yes” question, 1 predominantly “no” question), 9 phonetically rich sentences, 2 time phrases (1 spontaneous time of day, 1word style time phrase), 4 phonetically rich words

The following age distribution has been obtained:
- 7 speakers are under 16,
- 624 speakers are between 16 and 30,
- 368 speakers are between 31 and 45,
- 160 speakers are between 46 and 60,
- 20 speakers are over 60.

	ELRA members	Non-members
For research use	20,000 Euro	22,500 Euro
For commercial use	25,000 Euro	30,000 Euro

ELRA-S0168: French Speecon database

The French Speecon database is divided into 2 sets:
1. The first set comprises the recordings of 550 adult French speakers (275 males, 275 females), recorded over 4 microphone channels in 4 recording environments (office, entertainment, car, public place).
2. The second set comprises the recordings of 50 child French speakers (20 boys, 30 girls), recorded over 4 microphone channels in 1 recording environment (children room).
The database has been collected by ELDA. The owner of the database is NSC Natural Speech Communication Ltd. This database is partitioned into 23 DVDs (first set) and 3 DVDs (second set).
The speech databases made within the Speecon project were validated by SPEX, in the Netherlands, to assess their compliance with the Speecon format and content specifications. Each of the four speech channels is recorded at 16 kHz, 16 bit, uncompressed unsigned integers in Intel format (lo-hi byte order). To each signal file corresponds an ASCII SAM label file which contains the relevant descriptive information.
Each speaker uttered the following items:
- Calibration data: 6 noise recordings, the “silence word” recording;
- Free spontaneous items (adults only): 5 minutes (session time) of free spontaneous, rich context items (story telling) (an open number of spontaneous topics out of a set of 30 topics);
- 17 Elicited spontaneous items (adults only): 3 dates, 2 times, 3 proper names, 2 city name,s 1 letter sequence, 2 answers to questions, 3 telephone numbers, 1 language
- Read speech: 30 phonetically rich sentences uttered by adults and 60 uttered by children, 5 phonetically rich words (adults only), 4 isolated digits, 1 isolated digit sequence, 4 connected digit sequences, 1 telephone number, 3 natural numbers, 1 money amount, 2 time phrases (T1 : analogue, T2 : digital), 3 dates (D1 : analogue, D2 : relative and general date, D3 : digital), 3 letter sequences, 1 proper name, 2 city or street names, 2 questions, 2 special keyboard characters, 1 Web address, 1 email address, 208 application specific words and phrases per session (adults), 73 toy commands and 46 general commands (children).
The following age distribution has been obtained:
- Adults: 245 speakers are between 15 and 30, 210 speakers are between 31 and 45, 95 speakers are between 46 and 60.
- Children: 17 speakers are between 8 and 10, 33 speakers are between 11 and 14.
A pronunciation lexicon with a phonemic transcription in SAMPA is also included.in

	ELRA members	Non-members
For research use	50,000 Euro	60,000 Euro
For commercial use	67,000 Euro	75,000 Euro

ELRA-S0169: Hebrew Speecon database

The Hebrew Speecon database is divided into 2 sets:
1. The first set comprises the recordings of 550 adult Hebrew speakers (273 males, 277 females), recorded over 4 microphone channels in 4 recording environments (office, entertainment, car, public place).
2. The second set comprises the recordings of 50 child Hebrew speakers (24 boys, 26 girls), recorded over 4 microphone channels in 1 recording environment (children room).
The database has been collected and is owned by NSC Natural Speech Communication Ltd. This database is partitioned into 20 DVDs (first set) and 3 DVDs (second set).
The speech databases made within the Speecon project were validated by SPEX, in the Netherlands, to assess their compliance with the Speecon format and content specifications. Each of the four speech channels is recorded at 16 kHz, 16 bit, uncompressed unsigned integers in Intel format (lo-hi byte order). To each signal file corresponds an ASCII SAM label file which contains the relevant descriptive information.
Each speaker uttered the following items:
- Calibration: 6 noise recorded, the “silence word” recording;
- Free spontaneous items (adults only): 5 minutes (session time) of free spontaneous, rich context items (story telling) (an open number of spontaneous topics out of a set of 30 topics);
- 17 Elicited spontaneous items (adults only): 3 dates, 2 times, 3 proper names, 2 city names, 1 letter sequence, 2 answers to questions, 3 telephone numbers, 1 language ;
- Read speech: 30 phonetically rich sentences uttered by adults and 60 uttered by children, **4** isolated digits,1 isolated digit sequence, 4 connected digit sequences, 1 telephone number, 3 natural numbers, 3 natural numbers, 1 money amount, 2 time phrases (T1 : analogue, T2 : digital), 3 dates (D1 : analogue, D2 : relative and general date, D3 : digital), 3 letter sequences, 1 proper name, 2 city or street names, 2 questions, 2 special keyboard characters, 1 Web address, 1 email address, 208 application specific words and phrases per session (adults), 74 toy commands and 48 general commands (children)
The following age distribution has been obtained:
- Adults: 313 speakers are between 15 and 30, 174 speakers are between 31 and 45, 63 speakers are over 46.
- Children: 16 speakers are between 8 and 10, 34 speakers are between 11 and 14.
A pronunciation lexicon with a phonemic transcription in SAMPA is also included.

	ELRA members	Non-members
For research use	50,000 Euro	60,000 Euro
For commercial use	67,000 Euro	75,000 Euro

ELRA-S0170: BABEL Romanian database

The BABEL Romanian Database is a speech database that was produced by a research consortium funded by the European Union under the COPERNICUS programme (COPERNICUS Project 1304). The project began in March 1995 and was completed in December 1998. The objective was to create a database of languages of Central and Eastern Europe in parallel to the EUROM1 databases produced by the SAM Project (funded by the ESPRIT programme).

The BABEL consortium included six partners from Central and Eastern Europe (who had the major responsibility of planning and carrying out the recording and labelling) and six from Western Europe (which role was mainly to advise and in some cases to act as host to BABEL researchers). The five databases collected within the project concern the Bulgarian, Estonian, Hungarian, Polish, and Romanian languages.

The Romanian database consists of the basic "common" set which is:
- The Many Talker Set: 50 males, 50 females; each to read 4 connected passages, 1 block of 2-3 "filler" sentences, 4 phonemically compact sentences, 3-7 individual sentences, and 26 numbers.
- The Few Talker Set: 5 males, 5 females from the Many Talker Set; each to read additionally 3 blocks of syllables and, in 4 supplemental sessions, 16 connected passages, 4 blocks of 2-3 "filler" sentences, 4 repetitions of the 26 numbers.
- The Very Few Talker Set: 1 male, 1 female from the Few Talker Set; each to read additionally 5 pairs of context words and the syllables in these 5 contexts.

	ELRA members	Non-members
For research use	300 Euro	600 Euro
For commercial use	4,000 Euro	6,000 Euro

ELRA-S0171 : SALA II Spanish from Mexico database

The SALA II Spanish from Mexico database collected in Mexico was recorded within the scope of the SALA II project.
The database has been collected jointly by Cycmatec from Mexico and Applied Technologies on Language and Speech, S.L. (ATLAS) from Spain.
The owner of the database is Natural Speech Communications Ltd. (NSC) from Israel.
The SALA II Spanish from Mexico database contains the recordings of 1,075 Mexican speakers (539 males and 536 females) recorded over the Mexican mobile telephone network.
The following acoustic conditions were selected as representative of a mobile user's environment:
- Passenger in moving car, railway, bus, etc. (155 speakers),
- Public place (279 speakers),
- Stationary pedestrian by road side (223 speakers),
- Home/office environment (364 speakers),
- Passenger in moving car using a hands-free kit (54 speakers) .

This database is distributed as 1 DVD-ROM The speech files are stored as sequences of 8-bit, 8kHz a-law speech files and are not compressed, according to the specifications of SALA II. Each prompt utterance is stored within a separate file and has an accompanying ASCII SAM label file.
This speech database was validated by SPEX (the Netherlands) to assess its compliance with the SALA II format and content specifications.
Each speaker uttered the following items: 6 application words, 1 sequence of 10 isolated digits, 4 connected digits (1 sheet number -6 digits, 1 telephone number -9/11 digits, 1 credit card number -14/16 digits, 1 PIN code -6 digits), 3 dates (1 spontaneous date e.g. birthday, 1 word style prompted date, 1 relative and general date expression), 2 spotting phrase using an embedded application word, 2 spotting phrase using an embedded application word, 2 isolated digits, 3 spelled words (1surname, 1 directory assistance city name, 1 real/artificial name for coverage), 1 currency money amount, 1 natural number, 5 directory assistance names (1 surname out of a set of 500, 1 city of birth/growing up, 1 most frequent city out of a set of 500, 1 most frequent company/agency out of a set of 500, 1 "forename surname" out of a set of 150), 2 yes/no questions (1 predominantly "yes" question, 1 predominantly "no" question), 9 phonetically rich sentences, 2 time phrases (1 spontaneous time of day, 1word style time phrase), 4 phonetically rich words
The following age distribution has been obtained: 7 speakers are under 16,
643 speakers are between 16 and 30,
248 speakers are between 31 and 45, .
169 speakers are between 46 and 60,
and 8 speakers are over 60.

	ELRA members	Non-members
For research use	34,000 Euro	40,000 Euro
For commercial use	45,000 Euro	51,000 Euro

ELRA-S0172 : C-ORAL-ROM

Integrated reference corpora for spoken romance languages. Multi-media edition; tools of analysis; standard standard linguistic measurements for validation in HLT

The C-ORAL-ROM resource is a multilingual corpus of spontaneous speech for the main romance languages of around 1,200,000 words (IST 2000-26228). The resource comprises three components:

a) Multimedia corpus;
b) Speech software;
c) Appendix.

The corpus consists of four comparable recording collections of Italian, French, Portuguese and Spanish spontaneous speech sessions (around 300,000 words for each Language). The collections are delivered respectively by the following providers:

- Università di Firenze (Dipartimento di Italianistica, LABLITA);
- Université de Provence (Description Linguistique Informatisée sur Corpus);
- Fundação da Universidade de Lisboa/Centro de Linguística da Universidade de Lisboa;
- Universidad Autónoma de Madrid (Departamento de Lingüística, Lenguas Modernas, Lógica y F. de la Ciencia, Laboratorio de Lingüística Informática).

The C-ORAL-ROM corpus provides the acoustic source of each session together with the following main annotations:

- The orthographic transcription, in CHAT format, enriched with the tagging of terminal and non terminal prosodic breaks
- Session metadata
- The text to speech synchronization, in WIN PITCH CORPUS format, based on the alignment of each transcribed utterance,

The multimedia corpus comes with the speech software Win Pitch Corpus (© Pitch France. Minimal configuration: Pentium III, 1 GHz, 252 Mo Ram, S-blaster or compatible sound card, running under Windows 2000 or XP only. GDPLUS.dll installed on the same directory of the program required).

A series of appendix are also provided containing: a) the purely textual corpus in .TXT and .XML format; b) the PoS tagging of all and the corresponding frequency list of lemmas forms in .TXT files; c) a set of linguistic measurements extracted from the main corpus annotations, in .EXCEL files; d) the specifications and evaluations of the resource, e) corpus metadata.

The C-ORAL-ROM resource is delivered in 8 DVDs and 1 CD.

For more information: http://www.elda.org/catalogue/en/speech/S0172.html

	ELRA members	Non-members
For research use	1,500 Euro	3,000 Euro
For commercial use	10,000 Euro	20,000 Euro