The ELRA Newsletter

January - March
2005

Vol.10 n.1

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.

Editor in Chief:
Khalid Choukri

Editors:
Khalid Choukri
Valérie Mapelli
Hélène Mazo

Layout:
Martine Chollet
Valérie Mapelli

Contributors:
Driss Mammass
Ali Rachidi
Tanja Schultz

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,

The ELRA Annual General Assembly was held in Paris in April 2005. This meeting was chaired by Bente Maegaard, President of ELRA and over 15 members attended. Apart from the review of ELRA activities and the presentation of financial data (2004 report and 2005 budget), new services offered to ELRA members were presented:

- The Universal Catalogue is a new service offered to ELRA members who will be given an early access to almost 730 identified LRs.

- The fidelity programme, inspired from the airlines "frequent flyer" programs, is meant to attract and keep members by rewarding faithful members with miles (or glyphs).

- The Language Resources and Evaluation Journal, a quarterly journal dedicated to scientific papers mainly on these two topics and published by Springer/Kluwer, will be issued from mid-2005.

In addition, a number of articles of the association statutes have been modified during this Annual General Assembly in order to account for the changes seen by our area of activities and in order to capitalize on the experience learnt from this first decade of activities. The major changes relate to the extension of our focus to other resources (e.g. Multimodal) and the addition of evaluation as a key activity within ELRA.

Finally, 2 major events were announced:

- The 5th edition of LREC will take place in Genoa (Italy) on 22-28 May 2006.

- A 2-day HLT Evaluation Workshop will be held in early December 2005 to celebrate ELRA's 10th anniversary. This workshop should gather both distinguished speakers and evaluation specialists.

During this last months, ELRA and ELDA continued to work on a number of projects funded/supported by funding agencies, and organised in Paris the General Assembly and progress meeting for both European-funded projects TC-Star and CHIL.

Concerning the production of LRs, the consortium of Nemlar is currently producing a set of databases in standard Arabic: a broadcast news speech corpus of ca 40 hours, a TTS corpus with a male and a female speaker producing.

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

- S0166 : Fixed1frDesign

- S0173 : SALA Spanish Mexican Database

- S0174-01 : FASiL English unimodal "fasil-uk" corpus

- S0174-02 : FASiL Portuguese unimodal "fasil-pt" corpusw

- S0174-03 : FASiL Swedish unimodal "fasil-pt" corpusw

- S0174-04 : FASiL combined unimodal "fasil-all" corpus

- S0174-05 : FASiL multimodal “fasil-mm” corpus

- L0054 : Label-Lex (MW)

- L0055 : Label-Lex (SW)

As for this newsletter, it presents a description of the "Methods and developments in the creation of a computerised Amazigh script", one of the research focuses, considered as a priority, being the design and production of applications that are able to process linguistic data automatically (data expressed in the Amazigh natural language).

It also contains a description of "GlobalPhone: A Multilingual text and Speech Database". The GlobalPhone corpus provides transcribed speech data for the development and evaluation of large vocabulary continuous speech recognition systems in the most widespread languages of the world.

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

Methods and developments in the creation of a computerised Amazigh script

Ali Rachidi, Driss Mammass

Introduction

This paper falls within the scope of a large international movement that aims at ensuring that all peoples have access to all the resources they need to communicate in their language. Asserting or defending a language used to be done with other resources: defining a spelling, creating monolingual or bilingual dictionaries, collecting oral traditions or even developing printing fonts.

Today, the development of personal computers and networks have transformed computers into tools for writing and communication in the same way as paper has been since *Cai Lun* and printing since *Gutenberg*. However, all languages are not equal when it comes to computerisation and speakers of less-endowed languages have limited access to these new resources. This limitation can range from mere discomfort to a complete inability of use. Amazigh is one of these underprivileged languages of the information society.

Consequently, scientific and linguistic research has been launched to work on improving the current situation. One of the research focuses, considered as a priority, is the designing and production of applications that are able to process linguistic data automatically (data expressed in the Amazigh natural language). In this connection, we are proposing methods and strategies for producing a tool for 1) Amazigh word processing under ASCII coding and another in Unicode format after the integration of Amazigh Unicode format into IT applications by the companies responsible and 2) automatic translation and management of an Amazigh lexical database.

This paper is divided into five sections. In the first part, we present the linguistic background and the writing system of Amazigh. The second part describes the resources and software used in computerising the language. In the third part, we present basic elements of the Amazigh computer writing system. The fourth part presents the various methods used in computerisation and their implementations. The last section presents the implementation of certain methods and the tools involved.

Amazigh: a natural language

. Background

The Berber alphabet or Amazigh has undergone many changes and variations from its origins to the present day. We will briefly describe below the different stages the Tifinagh script has undergone starting from the Libyan form, then the Sahara and Tuareg forms to the modern day neo-Tifinagh variant.

Libyan

These are the earliest varieties of Tifinagh. There are two forms of Libyan: Eastern and Western. The western form was used along the Mediterranean coast of Kabylia up to Morocco and most probably in the Canary Islands. The eastern form was used in the Constantinois and Aurès regions and Tunisia.

Saharan Tifinagh

This variety is also called Libyan-Berber or Old Tuareg. It contains additional signs that are not found in the Libyan variety, more particularly a vertical stroke denoting the final /a/ vowel. Although this was used to transcribe Old Tuareg, these inscriptions are not understood at this stage. [IRCAM 04b].

Tuareg Tifinagh

Differences exist within the Tuareg Tifinagh as to the value of the signs used by each dialectal population. Although the appearance and number of signs vary from one area to another, Tuareg texts are generally understood by the various regions.

Neo-tifinagh

Neo-tifinagh refers to the writing systems that were developed to represent the Maghreb Berber (Amazigh) dialects. The first variant was the one proposed at the end of the 1960's by the Académie Berbère (AB) on the basis of Tuareg Tifinagh letters. It has become widespread in Morocco and in Algeria especially in Kabylia. This term "neo-tifinagh" also covers other variants that appeared to complement or correct the imperfections of the system proposed by Académie Berbère.

The wider family of Berber people who have in common the traditional Tifinagh script represents around 20 million people. In Morocco, the term Berber (or Amazigh) encompasses the three main Moroccan variants: *Tarifite*, *Tamazighe* and *Tachelhite*. More than 40% of the country's population speak Berber. However, Tifinagh now concerns all Moroccans since the teaching of Amazigh, written in Tifinagh, will be brought into general use and made compulsory in Morocco. The Royal Institute of Amazigh culture (IRCAM) is working to reach this objective.

. Tifinagh: the Amazigh alphabet

Tifinagh characters

IRCAM has proposed the Tifinagh alphabet to the International Organisation for Standardisation (21/06/2004) [IRCAM 04a] who confirmed the proposal. It is made up of four sub-sets of Tifinagh characters:

1. The basic Ircam set;
2. The extended Ircam set;
3. Other neo-Tifinagh letters in use;
4. Modern Tuareg letters that have been attested as used.

Below is the list of the Tifinagh alphabet and the associated Unicode plan allocated by the ISO illustrated by figure 1:

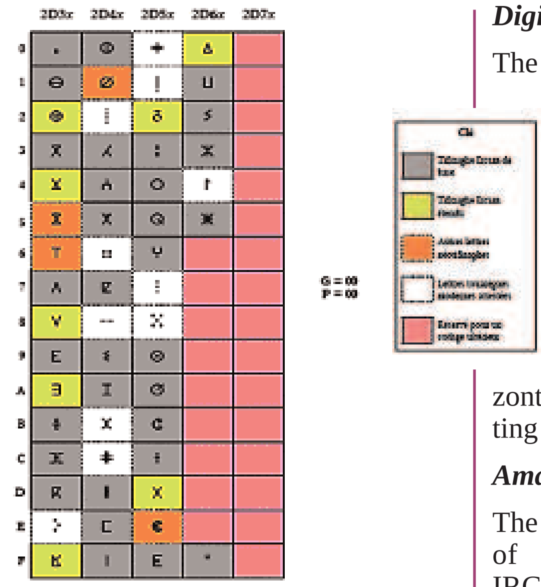

2D3x
2D4x
2D5x
2D6x
2D7x
0
1
2
3
4
5
6
7
8
9
A
B
C
D
E
F
G = 00
P = 00
Clé
Tifinaghe Ircam de base
Tifinaghe Ircam étendu
Autres lettres néotifinaghes
Lettres touarègues modernes attestées
Réservé pour un usage ultérieur

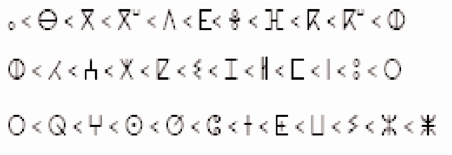
Figure 1: 2D Unicode range: Tifinagh

Punctuation

We do not know of any specific punctuation marks for Tifinagh. Ircam has recommended that the conventional signs found in Latin script be used: " " (space), " . ", " , ", " ; ", " : ", " ? ", " ! ", " … ", etc. Consequently, this proposal does not present any Tifinagh punctuation sign.

Sorting order

Only Ircam has defined a precise sorting order, described by the following expression (a < b means that a is sorted before b). The institute has developed a draft Moroccan standard for classifying character strings according to the Tifinagh, Latin and Arab alphabets that it will propose for inclusion in the next amendment to the international standard ISO/CEI 14651. It also provides information on what is considered to be the best way of dealing with the following difficulties:



- disambiguation of ligatures that have the same appearance as the letters of the basic alphabet;

- sorting of Latin and Arabic transcriptions of Amazigh, whether or not there are lists in the languages using the Latin or Arabic alphabet.

Digits

The Ircam proposal uses Western "Arabic" digits (0, 1, 2, 3, 4, 5, 6, 7, 8, 9) in Tifinagh script.

Consequently, this proposal does not present any new digits or numbers.

Directionality

IRCAM has chosen the horizontal left to right direction for writing Tifinagh.

Amazigh keyboard

The Tifinagh alphabet is made up of thirty-three characters. The IRCAM's computer centre (CEISIC) has proposed a keyboard in ASCII format (font, driver) illustrated in figure 2 [IRCAM 03a] [IRCAM 03b].

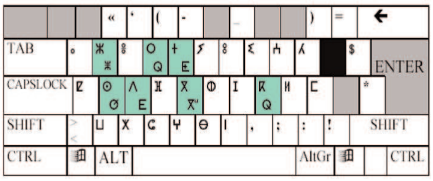


Figure 2: Amazigh keyboard in ASCII format

The first 26 characters can be accessed directly. Emphasis characters are obtained by using the black box (the " ^" on the Latin keyboard) in the same way as the "^" is used in French (to type "â" for example). The institute has proposed a project to standardise two keyboard groups by specifying two compliance levels: one for the strict keying of the thirty-two letters of the basic Tifinagh alphabet taught in Moroccan schools, and the other for keying in the basic alphabet, plus the twenty-two letters of the extended Tifinagh alphabet and ligatures (when the underlying technology makes it possible to process two command characters used to create or prevent the formation of ligatures). This project will be confirmed in the next amendment to the international standard ISO/CEI 14651.

Resources and software chosen for the computerisation of Amazigh

The Webster's Dictionary defines *computerisation as "To put in a form that the computer can use"*. Ideally therefore, to computerise a language is to place at the disposal of human users all the resources that they will need in their language, whether or not it is a written one: dialogue with the machine, tools for writing or reading a text (locally), sending e-mail (via a network), computerised translation into another language, etc.

More precisely, we have listed the resources and software that we have chosen for computerising Amazigh below:

Resources:

Dictionaries: bilingual and usage,

Software:

1.Word-processing software,
- Entry and viewing,
- Search and replacement of text,
- Text selection,
- Lexicographic sorting,
- Spell checker,
- Grammar checker,
- Style correction,

2.Voice processing software:
- Voice synthesis,
- Speech recognition,

3.Automatic translation and computer-assisted translation (written and oral) software,

4.Amazigh optical character recognition (OCR) software,

5.Software that provides advanced services (automatic summary, etc.),

6.Adaptations to existing software: These are software applications that were initially developed for well-endowed languages and adapted to Amazigh, with modifications that do not require any language processing skills, for example: translation of menus and messages in Amazigh, cultural adaptations, choice of fonts compatible with encoding and display technology.

Table 1 illustrates the developments required to create the resources that were listed above.

Amazigh computerised writing system

Current operating systems of microcomputers integrate the Unicode capacity in that they have a programming interface that is compatible with Unicode. They are therefore natively multilingual, once the writing system under consideration is in Unicode and has fonts that function for this writing system. The main operating systems are:

- Windows as from version NT 3.1 in 1993;

- MacOS as from version 8.5 in 1998;

- Linux as from XFree86 4.0 in 2000.

The basic element used in creating text is the edit window (the window in which text can be entered). Advanced edit windows that allow editing in several, or even all the writing systems contained in Unicode are included in the development environments in the form of objects or programming interfaces (API). The use of these edit windows saves a lot of time since these objects have become very complex with the Unicode. They carry out the following functions:

- Management of the actions of the keyboard and mouse,

- Display of text,

- End of line breaks,

- Text justification,

- Management of the cursor's movement,

- Text selection (inverted video),

- Copy and paste.

In addition to these basic functions, the current edit windows such as HTML, RTF and Word manage the association of attributes - bold, italics, underlined, font, etc. - to certain parts of the text, usually by placing tags in the text.

These functions, which are already quite difficult to develop for text in Latin characters, become extremely complex when you have to take into account all the constraints linked with all the writing systems:

- Shape of characters depending on the characters next to them (for example Arabic, Hebrew, Thai and Hindi), which is not the case for Amazigh because there is no cursive writing at the moment.

- Bidirectionality (for example a text that contains a part in Amazigh and a part in Arabic or Latin),

For instance, seemingly basic functions such as basic as text selection and even the management of cursor position become real headaches, in particular with texts that include both Amazon and Latin scripts.

Many applications compatible with Unicode have been developed, in particular automated office suites and Internet browsers. Some propose linguistic services such as automatic detection of the language, automatic date format, word breaks at the end of the line, segmentation (for scripts without separators between words), spell checkers, grammar and style checkers, lexicographic sorting, thesaurus, automatic summary, etc.

For example, Office XP, one of the most common automated office suites, includes language tools for forty-eight languages [LREC 98] [TALN 03]. Some of these applications are themselves objects that can be used as platforms to computerise Amazigh.

One of the objects that we will use in our developments is the EDIT class, a simple edit window that displays only one font at a time, the *CRichEditCtrl* class, RTF edit window and the *Word* a word-processing application that includes the powerful editing class, *WwG*

Methods for computerising Amazigh

. Computerisation with limitations

The computerisation of Amazigh is made difficult for several reasons. These include:

1. linguistic difficulties:
 - low level of description of the language,
 - a language that is not very written,

2. the small number, low income and weak skills of speakers, which results in:
 - fewer potential users,
 - fewer potential linguists,
 - fewer potential developers.

These difficulties have to be assessed at the beginning of the computerisation project, together with the resources available for the language: dictionaries, grammars, competent and motivated people to complete the project.

. Finding the right solutions: key concepts

Take advantage of developments made for related languages

Language speakers usually communicate with people outside their group through a common language or pivot language that may be central or super central. Although the two languages may not necessarily belong to the same language family, they often share common elements that can facilitate the computerisation of the language (script, vocabulary, bilingual dictiona-

Service	Development
Simple entry, vizewing and printing	Character font and virtual keyboard (format unicode)
Lexicographic sorting	Appropriate sorting tool
Spellchecker	Spelling checking tool
Grammar and style checker	Grammar and style checking tool
Speech synthesis	- Morphological analyser and generator for Amazigh - Syntactical analyser for the Language - Speech synthesis tool
Speech recognition	Speech recognition software
Automatic translation	Bilingual or multilingual lexical databases
Optical character recognition	Character recognition software

Table 1: developments needed for Amazigh

ries). The synergies with these pivot languages will therefore have to be assessed and must result in a development strategy that includes them.[LREC 98].

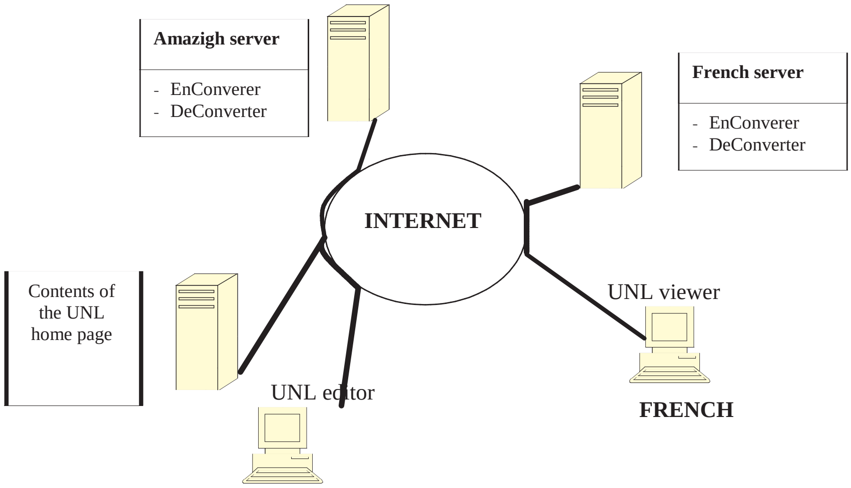
Amazigh is a language that is close to Arabic and French. Consequently, it is very possible to use the ties that link the three languages (Arabic and French are already very computerised) and to use the resources available for French to facilitate development.

Join generic, open source and projects and environments

UNL project (www.undl.org): UNL is the abbreviation of "Universal Networking Language". This is an artificial language that replicates the functions of natural languages on the Internet into human communication. It also enables computers all over the world to communicate with each other and to have a linguistic infrastructure to distribute, receive and understand multilingual information. It thus enables the different populations of the world to express all the knowledge communicated by natural languages. In this context, it is increasingly necessary to create multilingual documents that include Amazigh [Rachidi 04]. The current idea is to carry out collaborative manual translation on the Internet, using a multilingual polyphrase memory (MPM). These are tools that are being developed by the research team of Christian Boitet (GETA, CLIPS, IMAG in Grenoble, France) [Berment 04]. The result of the translation (sentences in Amazigh) will then be added to the document in UNL-XML. Lastly, we will therefore have to create a UNL-Amazigh DeConverter and EnConverter as shown in figure 3.

Figure 3: UNL system that will integrate Amazigh

Papillon Project (www.papillon-dictionary.org): the purpose of this project, which is backed by the international community, is to create a multilingual lexical database. Amazigh will be added to the project once the Unicode format is well established in the software platforms.



Using general linguistic contributions: pooling

We believe that it is possible to obtain language resources of quality by working in cooperation on the Internet [Berment 04], replacing a local team with a distributed working group, that comes at no cost and is potentially of a larger size. This idea of a "generalised linguistic contribution" on the Internet is the focus of the Montaigne project proposed in 1995 by the GETA and the company SITE-Eurolang, but it was unable to take off at the time because of inadequate funds [Berment 04].

Recycle existing dictionaries

The creation of a good quality lexical database is a complex task in the Amazigh computerisation process. An alternative to the creation of a complete and costly lexical database is the reprocessing of files containing dictionaries that are to be used in print version. Given that print dictionaries do not meet the same requirements as lexical databases, the data cannot be retrieved immediately. To facilitate retrieval, we can use tools like *RecupDic* in which the textual structure of dictionaries is formally described in order to automate their transformation into a lexical database. We must point out that such a process can often be implemented only after an ad hoc semi-automatic "standardisation" phase, which cuts down on "noise".

. Apply an adapted management

Determine which product to create

Given the limited resources at our disposal for computerising Amazigh, it is important to clearly define the product we wish to create.

- Is it a product intended for the general public or a laboratory prototype?

- Is it a word processor, an online translation service or a speech-processing tool?

- Is the software to be used with a Palm Pilot or for a microcomputer under Linux, Windows or Mac OS?

- Do the target users have access to the above resources?

- Is it a free or fee-paying product?

- When will it become available?

- How will it be distributed?

These marketing issues are essential if the project is to produce a useful result.

Determine who is producing the software and resources

The computerisation of Amazigh by native speakers may be made very difficult by the fact that they are not well-trained enough to carry out the work. More generally, the computerisation of Amazigh requires IT experts and/or linguists who may or may not be native speakers. The various scenarios are presented below.

The computerisation may be carried out by:

- a local group created specifically for the project

- a group working on an Open Source software or freeware project,

- members of the Diaspora working in a network,

- a specialised company,

- a scientific university laboratory,

- a language and linguistics institute,

- an association between several of these possibilities.

We might be obliged to work with a local group depending on the financing obtained for the project. This is the case of development aid funding (ICT4D) [Berment 04] that is aimed at the acquisition of IT expertise in developing countries. Whatever the case, funding requirements must be assessed at the beginning of the project.

Draw up a development plan

The computerisation of Amazigh is a complete project that can be mastered by breaking up the elementary tasks to be organised into a development plan.

.Development plan proposed for Amazigh

Development will be in two distinct phases, each of which requires a specific development plan:

- the creation of word processing that is adapted to Amazigh script,

- the development of computer-assisted translation and a lexical database by integrating the UNL and Papillon projects.

Implementation and tools involved

. Current phase (before Unicode format): Create word processing for Amazigh

We can build a software platform that will be used to process a text in Amazigh in ASCII with C / C++ by using the Software Development Kit (SDK) by Windows, that was aimed at the first levels of service of the processing of the text. It includes the following functionalities:

- Keying in of Amazigh regardless of the font used and using an intuitive keyboard,

- Change of font (and therefore conversion of code),

- Canonical formatting of the selected text (standardisation, unequivocal entry)

- Selection facilitated with the Amazigh syllable and word keyboard,

- Formatting of Amazigh texts, to make them useable by off-the-shelf word processors,

- Export in TeX and RTF formats,

- Creation of a glossary from texts (add, change and delete an entry in a local glossary),

- Translation of Amazigh words into French,

- Phonetic transcription of the selected text.

The proposed interface of this platform is illustrated in figure 2 below:

It will however not be user-friendly as compared with word processing market standards. For example, the edit window will allow only one font at a time, and several functions, such as printing, will not be developed. We will have to go one step further and offer a word processor of quality for the general public with Amazigh-specific functionalities. This is what makes us consider developing applications that take the Unicode format into account.

After the Unicode format phase

Using Word Pad

Word Pad can serve as a base for an Amazigh version (Amazigh Pad). These sources are available in Microsoft's C++ language. Word Pad is developed in C++ and is based on classes of the Microsoft Foundation Classes (MFC) Library [Berment 04]:

- Class of the edit window CRichEditCtrl;

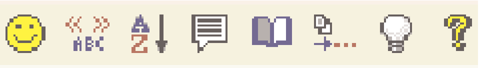
- Rich Text Format

Using WinWord

We can also use the Word developer's kit. This is a set of C modules that are used to interface with Word. The principle is to develop a dynamic library that is loaded with Word and which calls the Word functions through an API called CAPI [Berment 04].

The proposed format of this bar is as follows:

Below are the functions of the buttons, from left to right:
- Amazigh Word configuration;
- Change of Amazigh font;
- Sorting of tables in Amazigh;
- Amazigh transcriptions;
- Electronic dictionary;
- Text formatting;
- Choice between Amazigh and Latin to enter texts;
- Online help.



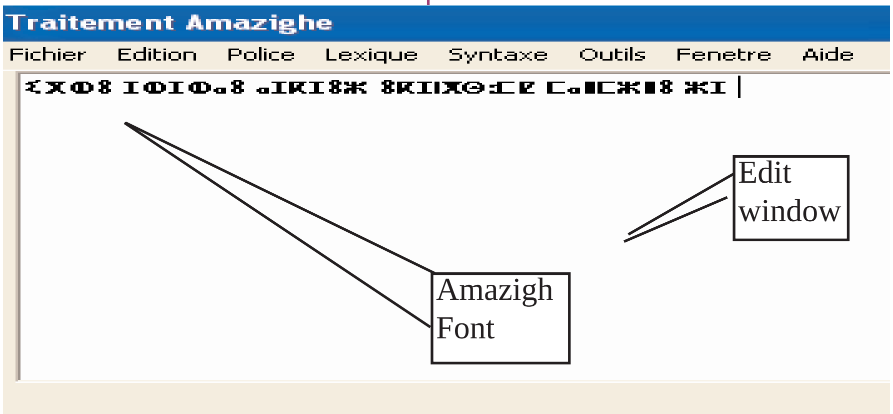


Figure 2: Amazigh editor interface

Conclusion

The computerisation of languages, which has developed with the explosion of information technology, has evolved, offering an increasing number of services an ever-increasing number of languages. This is however an expensive process that currently benefits only a small proportion of the languages of the world (less than 1%) [LREC 98].

Computerising Amazigh demands the efforts of several researchers and the use of levers that will enable them to obtain quality software very quickly. The Unicode standard, which has recently been extended to Amazigh, has led to the creation of operating systems and software that cover many different writing systems while avoiding the multiplication of incompatibilities between platforms. Amazigh can thus benefit from powerful editing tools because encoding and basic principles are shared by the different writing systems. The momentum provided by Unicode has thus led to the development of powerful, generic software that covers a large part of the word processing service levels. This includes the creation of highly multilingual edit windows (entry, viewing, printing, search/replace and text selection).

Bibliography

[IRCAM 04a] Proposal to add Tifinagh script to the ISO/CEI 10646 directory (Unicode format), 21/06/2004, CEISIC centre, Ircam, Rabat, Morocco.

[Berment 04] Vincent Berment, Méthodes pour informatiser des langues et des groupes de langues "peu dotées", PhD thesis, Université Joseph Fourier, Grenoble 1, UFR d'informatique et mathématiques appliquées, 18 May 2004.

[Rachidi 04] Ali Rachidi, Les Graphes UNL : un concept unificateur pour l'intégration de l'Amazighe dans des Documents Multilingues, international seminar: La typographie entre les domaines de l'art et de l'informatique, Ircam, Rabat, Sept. 2004

[IRCAM 04b] Institut Royal de la Culture Amazighe, Centre de l'Aménagement Linguistique (forthcoming), Writing of the Amazigh language, Coordinator El Mehdi Iazzi, Publications de l'IRCAM, Rabat, 2004.

[IRCAM 03] Institut royal de la culture amazighe, Centre des études informatiques et des systèmes d'information et de communication, Creating the first version of the Amazigh keyboard, L. Zenkouar, Y. Belkasmi & Y. Aït Ouguengay, Rabat, 2003 action plan, see also

http://www.ircam.ma/Telechager/pilote.htm

[IRCAM 03b] Institut royal de la culture Amazighe, centre des études informatiques et des systèmes d'information et de communication, Design and development of the following fonts:

- Tifinagh-Iracm standard font, L. Zenkouar, Y. Aït Ouguengay & H. Jaa ;

- Tifinagh-Iracm izzuren font, Y. Aït Ouguengay, H. Aarab & L. Zenkouar ;

- Tifinagh-Iracm taromit font, Y. Aït Ouguengay, H. Aarab & L. Zenkouar ;

- Tifinagh-Iracm tissnat n'irrumin font, Y. Aït Ouguengay, H. Aarab & L. Zenkouar,

Rabat, 2003 action plan, cf. http://www.ircam.ma/Telecharger/polices.htm

[TALN 03] workshop associated with the natural language processing conferences (TALN) 2003 "Automatic processing of minority languages and small languages:

http://www.sciences.univ-nantes.fr/irin/taln2003/page/acte_sommaire.html#atelier.

[LREC 98] workshop on minority languages of the International Conferences on Language Resources and evaluation (LREC) (once every two years since 1998):

http://www.lrec-conf.org/fr/index.html,

http://www.lrec-conf.org/lrec98/ceres.ugr.es/_rubio/elra/minority.html,

http://www.lrec-conf.org/lrec2000/www.cstr.ed.ac.uk/SALTMIL/lrec00.html,

http://www.lrec-conf.org/lrec2002/lrec/wksh/WP15agendaF.html,

Driss Mammass, Laboratoire de Traitement d'Images et Systèmes d'information (LTISI), Faculty of Science, Agadir, Morroco

driss_mammass@yahoo.fr

Ali Rachidi, Ecole Nationale de Commerce et de Gestion, BP 37/S Hay Slam, Agadir, Morroco

rachidi.ali@caramail.com

GlobalPhone: A Multilingual text and Speech Database

Tanja Schultz

The GlobalPhone corpus provides transcribed speech data for the development and evaluation of large vocabulary continuous speech recognition systems in the most widespread languages of the world. GlobalPhone is designed to be uniform across languages with respect to the amount of text and audio per language, the audio data quality (microphone, noise, channel), the collection scenario (task, setup, speaking style etc.), speaker population, and the transcription conventions. As a consequence, GlobalPhone supplies an excellent basis for research in the areas of (1) multilingual speech recognition including multilingual acoustic model combination and multilingual language modeling, (2) rapid deployment of speech processing systems to new languages, (3) language and speaker identification tasks, (4) monolingual speech recognition in a large variety of languages, as well as (5) comparisons across major languages based on text and speech data.

To date, the GlobalPhone corpus covers 15 languages Arabic (Modern Standard Arabic), Chinese-Mandarin, Chinese-Shanghai, Croatian, Czech, French, German, Japanese, Korean, Portuguese (Brazilian), Russian, Spanish (Latin American), Swedish, Tamil, and Turkish. This selection covers a broad variety of language peculiarities relevant for Speech and Language Research and Development. It comprises wide-spread languages (Arabic, Chinese, Spanish), contains economically and politically important languages (Korean, Japanese, Arabic), and spans over wide geographical areas (Europe, America, Asia). The spoken speech covers a wide selection of phonetic characteristics, e.g. tonal sounds (Mandarin, Shanghai), pharyngal sounds (Arabic), consonantal clusters (German), nasals (French, Portuguese), palatized sounds (Russian), syllable-based languages (Japanese), and more. The written language contains large orthographic variations, such as phonologic scripts (alphabetic scripts such as Roman, Cyrillic, Arabic; syllable-based scripts like Japanese Kana, Korean Hangul), and ideographic scripts (Chinese Hanzi and Japanese Kanji). Among the phonologic scripts are those with close grapheme-to-phoneme relations (Spanish, Croatian), reasonable relations (German, Russian), and those with weaker relationships between letters and sounds (Swedish). The GlobalPhone languages cover many morphological variations, e.g. agglutinative languages (Turkish, Korean), compounding languages (German), and also include scripts that completely lack word segmentation (Chinese).

The Table below shows the ranking of the most widespread languages of the world, the primary locations where the languages are spoken, and their pertaining speakers population (numbers according to Webster's New Encyclopedic Dictionary, published by Black, Dog & Leventhal, 1992). The languages covered by the GlobalPhone corpus are marked by "*". English is not collected in the framework of GlobalPhone, however the Wall Street Journal database is very similar with respect to domain, recording conditions such as microphone equipment, number of speakers per language, and amount of collected data. The data acquisition was performed in countries where the language is officially spoken. In each language about 100 adult native speakers (of both genders, in the range of 20 to 60 years) were asked to read 100 sentences. The read texts were selected from national newspaper articles available from the web to cover a wide domain with large vocabulary. The articles report national and international political news, as well as economic news mostly from the years 1995-1998. Each speaker was recorded in one session taking place in a quiet setting. The speech data was recorded with a Sennheiser 440-6 close-speaking headset microphone and is available in identical audio characteristics for all languages: PCM encoding, mono quality, 16bit quantization, and 16kHz sampling rate. The transcriptions are available in the original script of the corresponding language. In addition, all transcriptions have been romanized, i.e. transformed into Roman script applying customi-

1. *	Mandarin	China	907 Mio	Sino-Tibetan (Sinitic)
2. (*)	English	USA, UK, Can, Australia	456 Mio	Indo-European (Germanic)
3.	Hindi	India	383 Mio	Indo-European (Indo-Irania)
4. *	Spanish	Latin-America, Spain	362 Mio	Indo-European (Romance)
5. *	Russian	Russia, Indep. States	293 Mio	Indo-European (Slavic)
6. *	Arabic	N. Africa, Mid East	208 Mio	Afro-Asiatic (Semitic)
7.	Bengali	Bangladesh, India	189 Mio	Indo-European (Indo-Irania)
8. *	Portuguese	Brazil, Portugal, Angola	177 Mio	Indo-European (Romance)
9.	Malay-Indo.	Indonesia, Malay, Brunei	148 Mio	Austronesian (Polynesian)
10. *	Japanese	Japan	126 Mio	Isolate
11. *	French	F, Can, Africa, Switzerland	123 Mio	Indo-European (Romance)
12. *	German	G, Austria, Switzerland	119 Mio	Indo-European (Germanic)
:				
15. *	Korean	Korea, China	73 Mio	Isolate
17. *	Tamil	India, SriLanka, Malaysia	67 Mio	Dravidian
20. *	Wu/Shanghai	China (Shanghai)	64 Mio	Sino-Tibetan (Sinitic)
25. *	Turkish	Turkey	57 Mio	Altaic (Turkik)
43. *	Serbo-Croatian	Balkan Europe	20 Mio	Indo-European (Slavic)
85. *	Swedish	Sweden, Finland	9 Mio	Indo-European (Germanic)

Table 1: Most widespread languages of the world, GP languages marked by '*'

zed mapping algorithms. The transcripts are validated and supplemented by special markers for spontaneous effects like stuttering, false starts, and non-verbal effects such as breathing, laughing, and hesitations. Speaker information, such as age, gender, occupation, etc., as well as information about the recording setup (room characteristics and environmental noise conditions) complement the database.

Data collections in additional languages are currently under way and are planned to be released when finished. To date, the GlobalPhone corpus contains over 300 hours of speech spoken by more than 1500 native adult speakers. On average the audio recording length is about 8.8 seconds with roughly 18 word units per utterance. The data of each language is divided in speaker disjoint sets for training, development, and evaluation in the ratio of 80:10:10.

GlobalPhone has been extensively used for studies on the comparison of monolingual speech recognition systems between multiple languages, multilingual acoustic model combination, rapid language adaptation, experiments on multilingual and crosslingual articulatory features, non-verbal cues identification based on multilingual phone recognizers (speaker identification, language identification, accent identification), non-native speech recognition, grapheme-based speech recognition, and multilingual speech synthesis. Publications of these studies are available in widely accessible journals and conference proceedings. More details about the design and collection of GlobalPhone and the application to language independent and language adaptive acoustic modeling are reported in the following publications:

Tanja Schultz, *GlobalPhone: A Multilingual Text and Speech Database developed at Karlsruhe University.*

Proceedings of the International Conference 5on Spoken Language Processing (ICSLP), Denver, CO, 2002.

Tanja Schultz and Alex Waibel, *Language Independent and Language Adaptive Acoustic Modeling*. Speech Communication, Volume 35, Issue 1-2, pp 31-51, August 2001.

Tanja Schultz
tanja@cs.cmu.edu, tanja@xlingual.com

New Resources

ELRA-S0166 : Fixed1frDesign

The Fixed1frDesign includes all database specifications including the full list of a designed corpus: a set of phonetically rich sentences and a set of application oriented utterances used within the French SpeechDat(II) database (ref. ELRA-S0076, produced in the framework of SpeechDat(II)). The SpeechDat common specification totals 40 utterances per call, comprising a mixture of spontaneous and read speech. The purpose of each telephone call was to record the basic structure of the utterances mentioned below. All utterances are read speech unless marked as spontaneous.

Statistics are supplied for each corpus, which are computed on the repetition of digits, letters or phonemes (diphones and triphones) depending on the corpus type. These statistics are reported in a separate file for each corpus.
More information on the website: *www.elda.org*

	ELRA members	Non-members
For research use	6,500 Euro	8,500 Euro
For commercial use	6,500 Euro	8,500 Euro

ELRA-S0173 : SALA Spanish Mexican Database

The SALA Spanish Venezuelan Database comprises 1260 Mexican speakers (554 males, 706 females) recorded over the Mexican fixed telephone network. The corpus design was performed by Universidad Politècnica de Catalunya (UPC). Collection and part of the transcription were performed by a partner who withdrew from the project, final transcription and formatting were performed by ATLAS. This database is partitioned into 7 CD-ROMs The speech databases made within the SALA project were validated by SPEX, the Netherlands, to assess their compliance with the SALA format and content specifications.
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. Each prompt utterance is stored within a separate file and has an accompanying ASCII SAM label file.
More information on the website: *www.elda.org*

	ELRA members	Non-members
For research use	13,000 Euro	16,000 Euro
For commercial use	16,000 Euro	20,000 Euro

ELRA-S0174-01 : FASiL English unimodal "fasil-uk" corpus

The corpus was collected in the context of the FASiL project, EU FP5 IST-2001-38685 (www.fasil.co.uk), as a wizard-of-oz experiment. Therefore, there are sound recordings of subject and wizard. A total of 70 subjects were recorded.

The corpus is formatted as .wav files (u-law) for audio, plain ASCII text (.txt) for transcriptions, and a masterfile which binds .txt and .wav together. The masterfile is a "lattice" of the interaction in time, and contains the exact order of the interaction plus timings. The masterfile is loosely related to the HTK-SLF lattice format.

The woz experiment is about the voice interaction with a Virtual Personal Assistent (VPA) for an email, calender and contacts task. Hesitations are marked as "UH", noise as "NOISE" and other irrelevant stuff as "IRRELEVANT". All annotations are in lower case, except for the former mentioned cases.

Exact documentation of experiment in FASiL deliverable D.2.2

The interactions contain mostly sentences but also spelled names, email addresses, telephone numbers, yes/no questions.

	ELRA members	Non-members
For research use	4,000 Euro	8,000 Euro
For commercial use	8,000 Euro	10,000 Euro

ELRA-S0174-02 : FASiL Portuguese unimodal "fasil-pt" corpusw

The corpus was collected in the context of the FASiL project, EU FP5 IST-2001-38685 (www.fasil.co.uk), as a wizard-of-oz experiment. Therefore, there are sound recordings of subject and wizard. A total of 70 subjects were recorded.

The corpus is formatted as .wav files (u-law) for audio, plain ASCII text (.txt) for transcriptions, and a masterfile which binds .txt and .wav together. The masterfile is a "lattice" of the interaction in time, and contains the exact order of the interaction plus timings. The masterfile is loosely related to the HTK-SLF lattice format.

The woz experiment is about the voice interaction with a Virtual Personal Assistent (VPA) for an email, calender and contacts task. Hesitations are marked as "UH", noise as "NOISE" and other irrelevant stuff as "IRRELEVANT". All annotations are in lower case, except for the former mentioned cases.

Exact documentation of experiment in FASiL deliverable D.2.2

The interactions contain mostly sentences but also spelled names, email addresses, telephone numbers, yes/no questions.

	ELRA members	Non-members
For research use	4,000 Euro	8,000 Euro
For commercial use	8,000 Euro	10,000 Euro

ELRA-S0174-03 : FASiL Swedish unimodal "fasil-pt" corpusw

The corpus was collected in the context of the FASiL project, EU FP5 IST-2001-38685 (www.fasil.co.uk), as a wizard-of-oz experiment. Therefore, there are sound recordings of subject and wizard. A total of 70 subjects were recorded.

The corpus is formatted as .wav files (u-law) for audio, plain ASCII text (.txt) for transcriptions, and a masterfile which binds .txt and .wav together. The masterfile is a "lattice" of the interaction in time, and contains the exact order of the interaction plus timings. The masterfile is loosely related to the HTK-SLF lattice format.

The orginal recordings were 16bit PCM which are converted to 8bit u-law.

The woz experiment is about the voice interaction with a Virtual Personal Assistent (VPA) for an email, calender and contacts task. Hesitations are marked as "UH", noise as "NOISE" and other irrelevant stuff as "IRRELEVANT". All annotations are in lower case, except for the former mentioned cases.

Exact documentation of experiment in FASiL deliverable D.2.2

The interactions contain mostly sentences but also spelled names, email addresses, telephone numbers, yes/no questions.

	ELRA members	Non-members
For research use	4,000 Euro	8,000 Euro
For commercial use	8,000 Euro	10,000 Euro

ELRA-S0174-04 : FASiL combined unimodal "fasil-all" corpus

The corpus was collected in the context of the FASiL project, EU FP5 IST-2001-38685 (www.fasil.co.uk), as a wizard-of-oz experiment. Therefore, there are sound recordings of subject and wizard. A total of 210 subjects were recorded in the three project languages Swedish, Portuguese and English, all data for the same application. The corpus is formatted as .wav files (u-law) for audio, plain ASCII text (.txt) for transcriptions, and a masterfile which binds .txt and .wav together. The masterfile is a "lattice" of the interaction in time, and contains the exact order of the interaction plus timings. The masterfile is loosely related to the HTK-SLF lattice format. The woz experiment is about the voice interaction with a Virtual Personal Assistent (VPA) for an email, calender and contacts task. Hesitations are marked as "UH", noise as "NOISE" and other irrelevant stuff as "IRRELEVANT". All annotations are in lower case, except for the former mentioned cases. Exact documentation of experiment in FASiL deliverable D.2.2 The interactions contain mostly sentences but also spelled names, email addresses, telephone numbers, yes/no questions.

	ELRA members	Non-members
For research use	10,000 Euro	20,000 Euro
For commercial use	25,000 Euro	30,000 Euro

ELRA-S0174-05 : FASiL multimodal "fasil-mm" corpus

The corpus was collected in the context of the FASiL project, EU FP5 IST-2001-38685 (http://www.fasil.co.uk), as a wizard-of-oz experiment. Therefore, there are sound and interaction recordings of subject and wizard. A total of 90 subjects were recorded (30 per language: English, Portuguese and Swedish).

The corpus is formatted as .wav files (u-law) for audio, plain ASCII text (.txt) for transcriptions, and a TASX .XML for annotations which binds everything together. The multimodal woz experiment is about the voice interaction with a Virtual Personal Assistent (VPA) for an email, calender and contacts task. Hesitations are marked as "UH", noise as "NOISE" and other irrelevant stuff as "IRRELEVANT". All annotations are in lower case, except for the former mentioned cases. Exact documentation of experiment in FASiL deliverable D.2.2_b.

	ELRA members	Non-members
For research use	9,000 Euro	20,000 Euro
For commercial use	30,000 Euro	30,000 Euro

ELRA-L0054: LABEL-LEX (MW)

LABEL-LEX (MW) is a Portuguese formalized lexicon, containing 88 619 inflected multiword lexical units (formally, sequences of simple words).

From a linguistic point of view, multiword lexical units exhibit distributional and selectional constraints; they lack compositionality, and have, most of the time, idiomatic interpretations.

MWUs occur frequently in both everyday language and technical and scientific texts to express ideas and concepts that in general cannot be stated by "free" linguistic structures.

So it is impossible to envisage automatic text analysis without adequate identification and treatment of multiword lexical units. The meaning of a text is mostly supplied by frequent occurrence of multiword units, especially by compound nouns.

Other formats and other services may be supplied by the data owner upon request (e.g. conversion into buyer's formalism, selection of subsets of the words missing from your own dictionary).

	ELRA members	Non-members
For research use	3,000 Euro	5,000 Euro
For commercial use	10,000 Euro	15,000 Euro

ELRA-L0055: LABEL-LEX (SW)

LABEL-LEX (SW) is a Portuguese formalized lexicon, containing 1.545.156 simple inflected words. Each dictionary entry is associated to a lemma; information about POS and morphological attributes - such as gender, number, person, case (for personal pronouns), tense, mood, diminutives, augmentatives, and superlatives - is systematically formalized for each lexical entry.

Syntactic and semantic information is being encoded incrementally. For instance, verbs are sub-classified (transitive, intransitive auxiliary), adjectives are being refined with information about their syntactic sub-classification.

Other formats and other services may be supplied by the data owner upon request (e.g. conversion into buyer's formalism, selection of subsets of the words missing from your own dictionary).

	ELRA members	Non-members
For research use	2,500 Euro	5,000 Euro
For commercial use	10,000 Euro	15,000 Euro