The ELRA Newsletter

EUROPEAN LANGUAGE RESOURCES ASSOCIATION

ELRA

January - March
2002

Vol.7 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
Magali Duclaux-Jeanmaire

Layout:
Magali Duclaux-Jeanmaire

Contributors:
Anne de Roeck
Mathieu Guidère
Abdelhak Mouradi
Carol Peters
Mike Rosner

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.fr or
WWW: http://www.elda.fr

Dear Colleagues,

Year 2002 will be marked by a milestone event for the HLT community: the Language Resources and Evaluation Conference, LREC, in Spring 2002 (from 27/05 to 02/06). We look forward to meeting you at this occasion, and hope LREC will once again prove to be a useful forum to raise and discuss issues of interest for the HLT community.

Information about the LREC 2002 conference and other practical details can be found at the following address: www.lrec-conf.org.

Considering internal activities and achievements of this last quarter, there are several points which deserve to be mentioned.

First, ELRA's network of technical centres will be completed with a set of validation centres for written language resources (VC_WLR). It will be set up thanks to an Open Call, similar to the one used to select technical centres for the validation of spoken resources (VC_SLR).

Then, as ELRA & ELDA are getting more and more involved in the evaluation activity, already quoted in some previous information bulletins and CEO letters, a new team dedicated to this activity is currently being set up. Its first member joined us at the beginning of March.

The new ELRA web site, which has been completely redesigned, has been open to the public. As for the ELDA web site, it is currently being re-worked and redesigned, and it should be made publicly available before the summer time.

Finally, new members joined the centre in Morocco, MLTC (Mediterranean Language Technology Centre), which was officially launched in Autumn 2001. MLTC will be in charge of all the activities related to the languages of the OrienTel project and of the tasks ELRA & ELDA outsource in Morocco.

ELRA & ELDA are involved in several projects, at the French, European and international levels. Updates of our current activities are listed below, with details of the projects we are participated in at www.elda.fr.

In the framework of the Speecon project, the French recordings will soon be over, while the transcriptions are going on. The recordings for the Swedish and the Italian languages are also progressing.

The monthly newsletter published by the Euromap Language Technologies project is a successful initiative, as well as its French version, produced by ELDA.

In the framework of the C-Oral-Rom project, ELDA, which is responsible for the legal aspects of the distribution of corpora and for the information dissemination, has made publicly available the official web site: http://www.elda.fr/proj/coralrom.html. Its content has been validated by the partners involved in the project and is being updated regularly.

For the OrienTel project, ELRA & ELDA will be innvolved in the distribution of the 26 speech databases which are to be created, and in the recordings for Morocco and Tunisia, in co-operation with the Polytechnic University of Catalunia.

Arabic audio data are being collected in Paris in the framework of the Network-DC project, and we are considering the existing transcription conventions for the Arabic language, which will also be used in the framework of other projects, such as OrienTel.

As for the ISLE (International Standard for Language Engineering) project, ELDA is involved in the NIMM (Natural Interaction and MultiModality) part, and helps to define the criteria for the description of multimodal resources and to write the guidelines. A draft version of these guidelines has been sent to every participant in the project for comments before their release.

Concerning the evaluation, a call for the third campaign of CLEF (CLEF 2002, available in this issue) has recently been disseminated. CLEF has recently come to an agreement with Amaryllis, the French evaluation programme, to combine their efforts.

A new European event in the field of HLT will be organised on 26th and 27th September 2002 in Berlin, LangTech 2002. Complementary to LREC, LangTech will bring together key players in HLT who will be able to present some newly developed products and systems, and real world applications. LangTech will focus on the industrial, professional and commercial aspects of HLT. ELDA is in charge of the exhibition that will constitute a major event within LangTech 2002. If you are interested in exhibiting at LangTech 2002 and for further information, please contact: exhibition@lang-tech.org.

You can also visit the web site dedicated to this event at www.lang-tech.org.

Now, as far as the content of this newsletter is concerned, we have decided to release a special issue dealing with Mediterranean language processing. This issue thus comprises articles written by specialists in NLP particularly interested in Mediterranean languages, mainly Arabic. This issue thus comprises articles written by specialists in NLP particularly interested in those languages, and who conduct research and design applications in the various fields of their processing, i.e. Maltese language resources (Mike Rosner), corpus and lexicon designing (Mathieu Guidéré, Anne De Roeck), or speech synthesis (Abdelhak Mouradi).

Last but not least, the new resources added to the catalogue are listed below. Their detailed description can be found from page 13 to 16: W0030, Arabic Data Set; W0031, GeFrePac; W0032, Modern French Corpus with Anaphors Tagging; W0033, CRATER 2; S0119, Spanish SpeechDat database for the mobile telephone network; S0120, Translanguage English Database (TED) Transcripts database.

Sincerely,

Antonio Zampolli, President

Khalid Choukri, CEO

Arabic for Absolute Beginner

Anne de Roeck

Anne De Roeck started working on Arabic language processing in 1996, whilst at the University of Essex, and the ELRA Al-Hayat dataset was developed in collaboration with Abduelbaset Goweder (http://privatewww.essex.ac.uk/~agowed/). The article describing the profiling work on the dataset can be found on: http://www.elsnet.org/acl2001-arabic.html

Arabic presents many challenging features for language engineering. Some of these features also apply to other Semitic languages - for instance, in Hebrew. On the other hand, Arabic is quite unique in its diversity: text in every day use is written and read by at least 150 million native speakers with different national identities, spread across a wide cultural and geographic area. This is reflected in a rich variety of orthographic conventions and habits, which successful applications would need to handle.

Like other Semitic languages, Arabic is root-based. 80% of Arabic words are derived from roots that are just three consonants long. In the table below, *ktb* and *qtL* are transliterations of such roots. Words sharing a root also share an aspect of meaning, so correctly identifying the root of a word is reported to benefit recall in information retrieval.

Roots are related to words via a complex morphology, which (a) first turns roots into stems by application of a collection of patterns and (b) forms further words from roots and stems by adding affixes.

Root		Pattern	Stem	
ktb	wrote	fa?L	Katb	writer
		mf?wL	Mktwb	document
qtL	killed	fa?L	QatL	killer
		mf?wL	MqtwL	corpus

Fig. 1: Stem Patterns

Looking at how stems are formed, the table shows example derivations using patterns. The starting point is a basic pattern called *f?L* (pronounced as *f'l* - with a strong glottal stop). Think of each symbol (*f*, *?*, *L*) as a placeholder for a consonant. This basic pattern is then "manipulated" by inserting characters to form further patterns.

Given a root, a stem is formed by mapping each consonant in the root onto one of the three placeholder characters in the basic pattern. So, for *ktb* (a root which gives rise to words connected with writing), *k* matches *f*, *t* matches *?* and *b* matches *L*. A stem is formed by inserting characters from a further pattern in the appropriate places around the root consonants. For clarification, in the table, the root consonants and the placeholder letters in the basic pattern are highlighted, whereas letters added by further patterns are not.

In other words, stem patterns "interdigitate" with, or repeatedly interrupt, the sequence of root letters. This is notoriously difficult to parse using standard techniques. The examples also show that each pattern has a specific effect on meaning. There are several hundred patterns but each root only takes about 18 or so patterns.

Apart from pattern application, roots, stems and words can take further affixes to form further words, either as a result of derivation, or to mark grammatical function. The string *walktab*, for example, breaks down as *w* (*and*) + *al* (*the*) + *ktab* (*writers*). Other affixes function as person, number, gender and tense markers, subject and direct object pronouns, articles, conjunctions and prepositions, though some of these may also occur as separate words (e.g. in the example, *wal* from *w* (*and*) + *al* (*the*), may be written separately).

Arabic has two kinds of vowels: long and short. Short vowels are a significant part of words, and they appear in patterns. However, short vowels are not written. As a result, the effects of some patterns are indistinguishable in written text. For instance, the example *walktab* above may also break down as *w* (*and*) + *al* (*the*) + *ktab* (*book*), because the difference between the word for *writer* and the word for *book* lies in the presence of short vowels, but these are not written. Interpretation depends on the voweling. Readers must infer the intended meaning.

The long vowels - *a* (*alif*), *w* (*waw*) and *y* (*ya*) - are quite distinct from short vowels, and can occur as root consonants. In that case, they are considered weak letters, and the root is a weak root. Under certain circumstances, weak letters (i.e. long vowels functioning as root consonants) may appear, change shape (eg *waw* into *ya*) or disappear during derivation. Long vowels also occur as affixes, so identifying a long vowel as either an affix or a root consonant can be difficult.

Arabic has infixes as well as prefixes and suffixes. Any of these may be consonants or long vowels. Infixes are problematic because they break up further the root letter sequences (which tend to be short), and they are easily mistaken for root consonants. The difficulty with affixes can be put like this: it is hard to tell a weak root consonant from a long vowel affix; it is equally hard to tell a consonant affix from a non-weak letter root consonant. If a root consonant is mistaken for an affix and is removed, the root cannot be recovered.

Arabic plurals are a problem in a class of their own. Arabic forms a dual and some plurals with suffixes, like English. These plurals are called "external" plurals. However, the normal way of forming a plural is by applying a collection of patterns which change the internal structure of the word. The following examples show some of the complexity. Masculine external plurals take either a -*wn* or -*yn* suffix, as in *mhnds* (*engineer*), *mhndswn*. Female external plurals add the -*at* suffix, or change word final -*h* to -*at*, as in *mdrsh* (*teacher*), *mdrsat*. Broken plurals, on the other hand, affect root characters. The plural of *mal* (fund from root *mwl*) is *amwal*. The plural of *wSL* (link from root *wSL*) is '*aySaL*. Note too that these examples are rife with long vowels (*a*, *w*, *y*) which may be part of the root, or the plural pattern. The examples show how long vowels cause interference in the detection of broken plural patterns and other ways of segmenting words.

Processing Arabic is complicated further by the presence of regional spelling conventions. For instance, in the same newspaper, three versions of word initial *alif* may occur. One prominent orthographic problem is the behaviour of hamza ('). Hamza is written over a carrier letter and is produced as a soft glottal stop, but it is not always pronounced. When at the beginning of a word, it is always carried by the long vowel *alif*, but it may be written above or below, or even omitted. When occurring in the middle of words, hamza is often, but not always, carried by one of the long vowels, depending on rules. At the end of words, it may be carried or written independently. Spelling rules involving hamza are so complex that they often give rise to mistakes. Like other consonants, hamza may function as a root consonant and an affix, and is subject to the same problems as non-weak letter consonants, compounded by unpredictable orthography: identical words may have differently positioned hamzas and would be considered as different strings.

ELRA's Al-Hayat Dataset: Text Resources in Arabic Language Engineering

Anne de Roeck

This article co-incides with ELRA's release of an 18 million word Arabic dataset (ELRA-W0030), based on more than 45,000 Al-Hayat newspaper articles, covering 7 different subject areas. This is not the first, or largest collection of electronic Arabic text, but it is the first one on this scale that is widely available, and about which some preliminary findings have been published [1].

Building a text collection and experimenting with it takes a bit of nerve. The experience is filled with expectation, but there is also the fear of finding too many surprises. What if the profile of the data turns out to be too extreme to be of general use, or to be unsuitable for our purposes? For Arabic specifically, this might be an issue, because the literature suggests that Arabic text may have an unusual profile, and may be particularly challenging for language processing applications (eg [2], [3]). In this case, it appears all is well. At least at first sight, it looks as if our fears that Arabic text might show an extreme profile were unfounded. The Al-Hayat dataset clearly is not a balanced corpus in the technical sense, but it shows a comfortingly standard word frequency distribution, so there is no a-priori reason to believe that the sample is particularly skewed either. We also know that the texts contain reassuring levels of misspellings, and examples of the different orthographic conventions used throughout the Arabic speaking regions, as many authors had predicted it would.

Perhaps slightly more worrying is the fact that the dataset confirms Yahya's experiment [4], [1]. Using texts up to 20,000 words long, Yahya showed that Arabic type to token ratios are significantly lower than those for English. The general argument is straightforward to understand. Raw text in a morphologically complex language (like Arabic or Finnish) will feature more distinct words, which will occur less frequently, than the same amount of raw text in a morphologically frugal language (like English or Cantonese). Type to token ratios can be quite important because they may reflect data sparseness. Experiments on the Al-Hayat collection suggest that to get a similar ratio (or sparsity level), an Arabic corpus of raw text would need to be about 8 times as large as an English one [1].

Even so, it still seems that all is well. Sparsity can be alleviated by adding more data. As more electronic resources for Arabic become available, and with a growing interest in Arabic language engineering research, we may expect significant and rapid progress. After all, statistical language modelling techniques transfer well across languages, do they not? The successes of the last decade, for instance in Web retrieval applications in different languages, have shown that the key to many problems lies in the availability of sizeable electronic linguistic resources. So is there any reason to believe that large scale processing of Arabic language will remain a challenge for long?

Easy access to large amounts of electronic text certainly has a key part to play in ensuring success. Arabic text data have been very hard to come by. Without doubt, the shortage has slowed progress of both research and applications development. For lack of an alternative, much past work has been conducted on very small collections of text, sometimes only a couple of hundred words [5]. Under these circumstances it is not clear whether the results of such studies would scale up. A good example is the debate on whether to index on stem or root for Arabic information retrieval. Words derived from the same root, and words derived from the same stem, share an aspect of meaning, a fact which may be exploited by retrieval techniques. Root indexing retrieves all words sharing a root and, in previous studies, is reported to outperform stem and word indexing on recall and precision [3], [6]. Stem indexing retrieves all words derived from the same stem, and is reported to outperform root indexing on precision [7]. These are quite important, intriguing, and apparently contradictory findings. Importantly, they stem from studies conducted on small samples of quite specialised text formats (242 conference abstracts [3] , 355 bibliographic records [6], and 590 heterogeneous articles, abstracts and records [7]). Reliable conclusions will only be possible by testing on much larger samples (and particularly so when bearing in mind Yahya's prediction about the relative sparseness of Arabic text samples).

Notwithstanding, there are good reasons to believe that increased availability of linguistic resources will not solve everything. Arabic language processing is likely to prove a rich ground for research, and an exciting challenge, for some time to come. This is so for the simple reason that Arabic is just not very much like English at all. The argument runs as follows. Most mainstream language processing and retrieval techniques have been developed for, and tested on, Western European languages, and English in particular. These techniques transfer well to other languages, as long as certain key features are present. Arabic, however, shows a profile that looks set to stretch the boundaries of what we know how to do with current mainstream approaches.

The in-set article tries to summarise some of the key problems we should expect to find in Arabic text. The true extent of these problems becomes clear, when we realise that they challenge the full spectrum of techniques, from deep, rule-based solutions to rough-and-ready, brute force alternatives.

With respect to rule-based analysis, Arabic morphology escaped routine treatment [8], [9] for quite a while. The reasons lie partly in a number of characteristics which Arabic shares with other Semitic languages. Arabic stem and word formation involves "interdigitation", a phenomenon whereby the sequence of letters in a root or stem is repeatedly interrupted by infixes. There is also the problem of "weak letter change", where quite standard morphological derivations, like the formation of plurals, will add, delete, or change certain letters in the root. Matters are further complicated by an orthographic convention which omits vowels. Derivational affixes, articles, pronouns and some prepositions and conjunctions are added as prefixes and suffixes, each often a single consonant long. As a result, a single string of characters may stand for quite different words and Arabic readers frequently rely on context to arrive at an interpretation. An approximation, perhaps, might make this clearer to the average English reader. Imagine a way of writing English, where the sequence of consonants *nthmt* might stand for "on the mat", or for "in the moat", and where context would provide clues as to which was intended (as in "Thct st nthmt."). Tuned as they are to Western European languages, most mainstream morphological analysis techniques do not fare well when faced with this combination of features.

A breakthrough came only recently [10], with refinements to finite state network compilation. The motivation for these developments originated directly in the desire to treat non-concatenative morphology, of which Arabic was the prime example. This led to the Xerox Arabic Morphological Analyser and Generator (www.xrce.xerox.com/research/mltt/arabic/), an application which looks set to allow morphological processing on a large scale. Whether it will be useful for retrieval on large bodies of text can of course only be verified with the help of sufficient textual resources.

Similarly, there is no reason to believe that shallow techniques would fare any better. Here, stemming algorithms are a good example. They work by stripping, from words, those character sequences which are deemed relatively irrelevant to content. Removing word final *-s* in English, for instance, makes the plural and singular of most nouns co-incide, so they can be retrieved together. Typical stemmers are simple devices and they have been hugely successful in many information retrieval applications.

In contrast, much work has been reported on Arabic stemming algorithms, but success on the scale of English stemmers has been elusive, because affix stripping in Arabic is fraught with danger. Popular stemming algorithms are inspired by Western European languages, and they make at least two assumptions which do not hold universally. First of all, they assume that affixes are quite easily identifiable. Secondly, they assume that the meaningful part of a word is comparatively large and invariant, so that "overstripping" the odd character is not disastrous. In other words: stemming can be fairly hit and miss, as long as not too much information is lost from the meaningful part of the word. So, whether "computing", "computational", "computers" and "compute" are conflated into "comput-" or "compu-" will not matter a great deal.

It now becomes easy to show why attempts at stemming Arabic in a similar fashion have been far less successful. A very large part of the meaning of an Arabic word resides in just three root consonants. Though these occur in a sequence, they are not necessarily adjacent as they may be interspersed with infixes. There is no visible distinction between affixes and root characters: they are all just consonants. Furthermore, weak root consonants in words which are clearly related may have disappeared or undergone transformation. With only three root consonants per average word, the price for one consonant lost to erroneous stemming is an information loss of at least 33%. For these reasons, research into alternatives to stemming, for instance how to cluster related words without affix stripping [5], may deliver exciting results.

The conclusions are clear. Easy access to large scale electronic resources is crucial for a sustained research and development programme in Arabic language engineering. Hopefully, the ELRA Al-Hayat collection is only one of the first of many resources to be released. Success, however, will require more than that. Straightforward application of established techniques to Arabic is not likely to suffice, because the language shows many features which seriously stretch the boundaries of what we know how to do. This is why the field of Arabic language processing is of interest, not just to researchers and developers working on Semitic languages, but to the whole language engineering community. Collectively, we have some methodological lessons to learn: the language engineers' horizon is only beginning to reach out to the majority of the world's languages, many with a great variety of notations and flexible orthographic conventions. We have little reason to believe that the standard tools of our trade will fit them any better than they do Arabic. Anything we can learn here will benefit all.

References

[1] Goweder, A. and A. De Roeck. 2002.

Assessment of a significant Arabic Corpus. Workshop on Arabic Language Processing, Proceedings of the 39th ACL, Toulouse.

[2] Ali, N. 1988.

Computers and the Arabic Language. Al-khat Publishing Press, Ta'reep, Cairo.

[3] Hmeidi, I, Kanaan Ghassan and Martha Evens. 1997.

Design and Implementation of Automatic Indexing for Information Retrieval with Arabic Documents. Journal of the American Society for Information Science. 48(10):867-881.

[4] Yahya, A. H. 1989.

On the Complexity of the Initial Stages of Arabic Text Processing. Birzeit University, Birzeit, West Bank.

[5] De Roeck, A.N. and Waleed Al-Fares. 2000.

A Morphologically Sensitive Clustering Algorithm for Identifying Arabic Roots. Proceedings of the 38th ACL, HongKong.

[6] Al-Kharashi, I. and M. Evens. 1994. *Comparing words, stems, and roots as Index terms in an Arabic Information Retrieval system*. Journal of the American Society for Information Science, 45/8, pp. 548-560

[7] Al-Tayyalh, M.S. 2000 *Arabic Information Retrieval System based Morphological Analysis*, PhD Thesis, De Montfort University, UK

[8] Kiraz, G. 1994. *Multi-tape two-level Morphology: a case study in Semitic non-linear morphology*. Proceedings of COLING-94, pp180-186.

[9] Beesley, K.B. 1996. *Arabic Finite-State Morphological Analysis and Generation*. Proceedings of COLING-96, pp 89-94.

{10] Beesley, K. and L. Karttunen. 2000. *Finite-state Non-concatenative Morphotactics*. Proceedings of the 38th ACL, HongKong.

Professor Anne De Roeck
Computing department
The Open University
Walton Hall
Milton Keynes MK7 6AA (UK)
Tel:. + 44 1908 654991 (direct)
Tel:. + 44 1908 858462 (secretary)
Fax:. + 44 1908 652140
Email: A.DeRoeck@open.ac.uk
Web site: http://mcs.open.ac.uk/anr29

Alignment of a French-Arabic Corpus and Extraction of a Bilingual Lexicon

Mathieu Guidère

The increasing utilisation of corpus-based methods is due to the increasing availability of textual databases. Regarding the Arabic language processing, existing and exploitable monolingual corpora are rare; besides, no real reference corpus can be easily and freely accessed. Thanks to the Internet, we are able to elaborate extended work corpora, but patience and perseverance may be needed.

The research work described below lies in the framework of the corpus linguistics activity. This activity aims to extract lexicons and develop terminological tools that may be integrated into Arabic machine translation systems. The corpus considered as a basis for work comprises French articles extracted from "Le Monde Diplomatique", with their translation in Arabic extracted from the Arabic version of this newspaper, covering a time span of 12 months. The whole bilingual corpus contains over 1 million words.

We will not enter into further details concerning the acquisition of the translated corpus, but rather present the tasks that had to be performed for its processing. First, both corpora, in French and in Arabic, had to be aligned, i.e. matching and synchronisation of the texts in both languages. The second processing task consisted of formalising the extraction of the bilingual lexicon, so that it may be reused in machine(-assisted) translation applications.

You will find below a description of the main steps which have been followed to perform both tasks, as well as the utility of such a bilingual corpus.

Constitution and Alignment of the Bilingual Corpus

As no tools are made publicly available or effective for Arabic language processing, (e.g. word frequency counters, concordance programmes, analysers, etc.), all the phases described below have been completed by a human being, manually (most of the time). The bilingual corpus was constituted by progressive stages, downloading from the Internet every new complete release of the newspaper "Le Monde Diplomatique", from November 2000 to November 2001. All downloaded articles were "cleaned up", and saved into a text format.

The process was the same for the Arabic version, despite the various problems we encountered, such as a slightly delayed publication, the contingencies regarding the edition, and some difficulties for accessing the on-line articles made available by the Arabic partner of the newspaper.

Once both corpora were ready, the next crucial step consisted of aligning the French documents and their corresponding Arabic version: at the text level; at the paragraph level; at the sentence level and at the word level. A few comments need to be mentioned to understand the nature and the challenges of this processing task:

1) As the Arabic version of "Le Monde Diplomatique" does not consist of an exhaustive translation of the original French newspaper, and rather contains a selection of some articles which may be of interest for the Arabic-speaking community, the matching of the Arabic and French texts had to be done manually.

The titles of the articles as well as the authors' names were precious indications to establish the matching between the texts during this macro-structural alignment phase. Nevertheless, we encountered numerous difficulties, because the translations of the titles were approximate, if not totally free, several articles included in the corpus were written by the same author, which did not facilitate an author's name-based alignment, and the transcription of the authors' names in Arabic was often incorrect[1] . The textual metadata attached to the articles (titles, authors' names) were not sufficient to proceed to a correct alignment, and we had to make further enquiries.

2) The matching between both corpora was achieved thanks to the alignment of the texts' and paragraphs' segments. Some basic but efficient principles were defined to complete this tiresome and time-consuming task:

- Two articles are considered to be reciprocally translated (therefore aligned), if at least two paragraphs (the first and/or the last one) are translations of one another;
- Two paragraphs are considered to be reciprocally translated (therefore aligned), if at least two sentences (the first and/or the last one) are translations of one another;

The alignment of the Arabic and French corpora proved to be easier, and more reliable, thanks to these two basic principles. Actually, the texts were compared to one another, at the sentence level, i.e. at a microstructure level.

3) Though the alignment at the sentence level is the most tricky phase of the processing, it is also the most important. The feasibility or not of the matching that was envisaged initially is determined by this operation, completed by a human person who has to compare manually the sentences to decide whether they are translations of one another or not. To optimise the process, a tested and validated procedure was set up to compare the sentences. This procedure might be reused in the future. The potentially alignable sentence segments were examined according to the following:

- Two sentences are reciprocally translated if at least two full words[2] are translations of one another.

According to this principle, two sentences are compared and matched with one another following purely lexical criteria.

At this stage, the structure of the sentences is not taken into account, though it may generate some errors. In fact, in order to ensure a reliable alignment, one has to choose a lexical anchor (“pivot word”) which acts as a marker for the matching operation.

Basically, the alignment procedure of the corpus is the opposite of its interrogation procedure. Indeed, in practice, the alignment of the texts is completed from the top level to the bottom level, i.e. from the largest to the smallest textual items (paragraph, sentence, word), while the corpus interrogation starts from the smallest items (lemma) to the largest ones (sentence, paragraph, text).

Once the first step, i.e. the alignment at the sentence and word levels, is achieved, the next step consists of extracting from the corpus a bilingual lexicon.

From the bilingual corpus to the bilingual lexicon

Statistical methods have been used for years for the extraction of lexicons from corpora. Among the different studies carried out on Arabic language processing, Mr Kouloughi's lexicon (1991) and more recently, Mr Guidère's (2001)[3] illustrate the use of such methods.

A major default of such lexicons, in spite of their undeniable interest, is the translation of these terms selected through statistical calculation. As the corpora used for these studies are monolingual, the translations which are given for the selected terms, which are taken out of their original context, may be incorrect. If the most recent lexicon (Guidère 2001), which was created following statistical methods, corrects the problem by illustrating the use of the term with an example found in the corpus, the translation which is proposed for a sentence or a term was nevertheless initially provided by a human translator, and its use has not been attested in a bilingual corpus.

That is why we decided to use a validated bilingual corpus to set up the lexicon. To use an homogeneous set of French texts, that had been translated beforehand by professional translators, allowed not only to exploit a valid complete corpus in the process, but also to provide coherent and homogenised translations, because all the data derived from the same corpus.

Below are listed various aspects of the processing which should be performed on the corpus before starting the extraction of the bilingual lexicon. As we are still elaborating on the process, we will focus on the most significant conclusions that could be drawn up to the present days, from the translation point of view:

1) *No translation*: there are some French words that are not translated in Arabic and for which there is no equivalent available in the corpus.

2) *Double-translation*: a French word may have two corresponding words in Arabic, because of the redundancy principle, typical of the journalistic style.

3) *Reduction*: several French words may be translated thanks to a single Arabic word which meaning comprises all the different French words' meanings.

4) *Expansion*: the translation of a single word in language A is made of several words in language B, as this is often the case for compound words in French.

5) *Set Phrase* as a Translation: a word group in French may be translated thanks to a set phrase in Arabic.

6) *Multiple Translations*: a single word may have several translations in a document; this is often the case for prepositions and conjunctions.

Considering these observations, which terms, among the whole translation of the original version, should be selected as entries for the lexicon?

Applications in Arabic language processing

Both translation (Machine Assisted Translation) and lexicography (elaboration of a bilingual dictionary based on a corpus with attested uses of the words) are some potential applications.

To elaborate the bilingual corpus, we focused on the lexicon. The alignment of the corpus at the word level, i.e. to search for lexical equivalents, seems to be more complex because of the various translation possibilities that could be observed during the synchronisation phase. In fact, these “many translations” refer to models of equivalence between the various units comprised in the texts of the corpus. Up to now, we have been able to characterise a few structures, in French and in Arabic, that are relevant for the lexical items, typical of the journalistic style. Later on, we will refine the results in order to get some exploitable patterns for automatic alignment of the units corresponding to these structures.

Linguistic Specifications of the French and Arabic Units:

Some morpho-syntaxic structures which seem to be stable enough between the French and Arabic languages have been highlighted in the parallel corpus. These structures include the noun phrases, and deal with the syntactical correspondence between the two languages. As a priority, we focused on lexical units larger than words, i.e. lexical units comprising at least two full words such as adjectives, nouns, etc., abbreviated as follows:

Adj. for Adjective, *N*. for Noun, *Prep*. for Preposition, *Pr*. for Present Participle, *Pa*. for Past Participle, *Det*. for determiner, *Poss*. for possessive, *Pro*. for pronoun, *Suff*. for suffix.

1) - *Adj. N*. (in French) = *N Adj*. (in Arabic)

Ex: “premier ministre” = “wazîr awwal”

- *Det. Adj. N. = Det. N. Det. Adj.* ‘Le premier ministre” = “Al wazîr Al awwal”

N.B. the nature of the adjective is essential to determine the equivalence structure:

Ex: *Adj. poss. N.* = *N. Pro. suff.*
"notre pays" = "bilâdu-nâ"
2) *N. Pr.* (in French) = *N. Pr.* (in Arabic)
Ex: "phénomène inquiétant" = "zâhira muqliqa"
3) *N. Pa.* (in French) = *N. Pa.* (in Arabic)
Ex: "rapport détaillé" = "taqrîr mufassal"
4) *N1 Prep. N2* (in French) = *N1 Det. N2* (in Arabic)
Ex: "ministère des finances" = "wazârat Al mâliyya“
- *N1 Prep. Det. N2* = *N1 Det. N2*
"ministère de la culture" = "wazârat Al thaqâfa"

These few examples of "morpho-syntactic patterns" (representations of some noun phrases), show that there are two ways of aligning the texts at the bottom level of the sentence: each word can be taken out of its context, which means that it has no relations with other words of the sentence, and submitted then to an interrogation procedure of "document query" type, into a tagged corpus (chapters, texts, paragraphs). All the attested equivalents of the word which is being "queried" are proposed.

Another possible way to align the corpus consists of considering the word as part of a whole structure, and of searching the structures and morpho-syntactical patterns which match with the original structures, instead of searching terms-to-terms matching.

We would recommend the latter solution.

[1] Some of the articles were not even signed!

[2] "Full word": a lemma associated to a non grammatical word (preposition, conjunctio, etc.). Ellipsis and anaphors are this way removed, and do not impact on the alignment.

[3] Kouloughli (D.E.), *Lexique fonctionnel de l'arabe standard moderne*, Paris, L'Harmattan, 1991, 288 p.; Guidère (M.), *Lexique bilingue de l'arabe d'aujourd'hui*, Paris, Editions du Temps, 2001, 285 p.

Mathieu Guidère
Maître de conférences
Lyon 2 University
86, rue Pasteur
F- 69002 Lyon (France)
Email: mathieu.guidere@univ-lyon2.fr
Web site: http://nte.univ-lyon2.fr/~mguidere

The Maltilex Project

Michael Rosner

Abstract

This brief article describes the background and current directions of Maltilex, a project of the University of Malta for which the principal aim is the construction of a computational lexicon for the Maltese language. The paper commences with a short description of some of the more particular characteristics of Maltese. This is followed by an exposé of the project aims and achievements, concluding with some ideas for the future.

Introduction

Maltese is the national language of the Maltese Archipelago and is, together with English, the official language. It is spoken by most of the Maltese people who live in Malta and Gozo (approximately 370,000) and also by a substantial number of people who emigrated from Malta in the 1950's and 60's and established communities in Australia, GB and the United States and Canada. It thus has a total of just under one million native speakers.

Within the context of the EU, a language with so few speakers is often referred to as a 'minority' or 'regional' language and as such, enjoys a special status. This is something of a mixed blessing. On the positive side, the EU is making a great effort to favour regional diversity by providing incentives to encourage the preservation and development of regional languages. But praiseworthy as these efforts are, they tend to miss the point as far as the problems of Maltese is concerned.

Maltese is not a minority language in the usual sense of being spoken by a minority of citizens within a given country. It is the national language and is truly national in scope. It is the language spoken at home, the language of the government, of TV soap operas and of literature. In short, minority status is not the problem of Maltese, and the usual protective remedies for this condition are somewhat inappropriate. A far more pressing problem for Maltese is that of bringing it into the electronic age. This problem manifests itself in a number of different guises:

- Idiosynchratic conventions for the use of Maltese characters in documents. This includes character representations, sorting order, keyboard layout;
- Preference for using English in computer-based communications (e.g. email) and in interfaces. This is partly a question of habit, due to the physical awkwardness of using Maltese with computers. Another factor is the lack of technical vocabulary for most of the terminology associated with computer usage.
- Lack of formal linguistic knowledge as input for language-sensitive support, e.g. spell and style checking. There remains a serious lack of investment in the linguistics profession. Consequently, many areas of the language are not that well understood in comparison with better-studied languages.
- Lack of language resources. There is as yet no organised system for acquiring and cataloguing language resources. What has been acquired so far has been on an ad-hoc basis.

The Language

Maltese is a so-called 'mixed' language, with a substrate of Arabic, a considerable superstrate of Romance origin (especially Sicilian) and, to a much more limited extent, English. The Semitic (Western/Maghrebi Arabic) element is evident enough to justify considering the language a peripheral dialect of Arabic. Its script, codified in the 1920's, utilises a modified Latin alphabet. This is just one of the peculiarities of Maltese as compared to other dialectal varieties of Arabic. More important ones are its status as a 'high' variety and its use in literary, formal and official discourse, its lack of reference to any Qur'anic Arabic ideal, as well as its handling of extensive borrowings from non-Semitic sources. These features make

Maltese a very interesting area for those working in the fields of language contact and Arabic dialectology.

The Alphabet

Maltese orthography was standardised in the 1920's, utilising an alphabet largely identical to the Latin one, with the following additions/modifications:

Letter	Nearest English Equivalent
ċ/Ċ	mu**ch**
ġ/Ġ	**J**anuary
għ/Għ/GĦ	silent
ħ/Ħ	**h**eadache
ż/Ż	**Z**anzibar
ie/Ie/IE	**ea**r (not exact)

Note that this is not an exhaustive guide to pronunciation. There are other English letters that carry unfamiliar sounds. The interested reader is referred to Falzon (1997), which is more accessible than the more academic Suttcliffe (1936).

Morphological Aspects of Maltese

The morphology is in part based on a root-and-pattern system typical of Semitic languages. For example, from the triliteral root consonants *h-d-m* one can obtain forms like: *hadem* (*to work*); *haddiem* (*worker*); *hidma* (*work*/noun); *hadem* (*be worked*/verb passive); *haddem* (*he caused to work*).

Most of these forms are based on productive templates called *forom* of which Maltese has a subset of those in Classical Arabic. One other typical feature shared with Semitic languages is 'broken' plural. Plural formation in such instances involves an actual change in the pattern of vowels and consonants.

Singular	Plural
qamar (moon)	qmura (moons)
tifel/tifla (boy/girl)	tfal (children)

In contrast to this, sound plural formation involves affixation of suffixes such as *-i* (very common with words of Romance origin), *-iet* or *-a* as in:

Singular	Plural
karozza (car)	karozz-i (cars)
ikla (meal)	ikl-iet (meals)

Maltese has taken on a very large number of Romance lexical items and incorporated them within the Semitic pattern. For example, *pizza*, a word of Romance origin, has the broken plural form *pizez* (cf. Italian *pizza*/*pizze*), and *cippa*, a very recent borrowing from English (*computer chip*) has a broken plural form *cipep*. In certain cases, one gets free variation between the broken plural form and a sound plural based on (Romance) affixation, e.g.:

Singular	Plural
kaxxa (box)	kaxex/kaxxi (boxes)
tapit (carpet)	twapet/tapiti (carpets)

The stem, as opposed to the consonantal root, also plays an important role in word formation, in particular in nominal inflection. Typical stem-based plural forms in which the stem remains intact are:

Singular	Plural
ahbar (news item)	ahbar-ijiet (news)
omm (mother)	omm-ijiet (mothers)

Verbs are also often borrowed and fully integrated into the Semitic verbal system and can take all of the inflective forms for person, number, gender, tense etc. that any other Maltese verbs of Semitic origin can take. For example, '*spjiega*' (*to explain*) which clearly derives from the Italian '*spiegare*':

Person	Singular	Plural
I	nispjiega	nispjegaw
II	tispiega	tispiegaw
III	jispiega	jispiegaw
IIIF	spjegajt	

The vigour and productivity of these processes is attested by the fact that one keeps coming across new loan verbs all the time (increasingly more from English), both in spoken and in written Maltese, without the language having any difficulty in integrating them seamlessly into its own morphological paradigms.

Within the verbal system, complex inflectional forms can also be built through multiple affixation. For example, the word '*bghatthielux*' (*I didn't send her to him*), contains the suffixes *-t* or 3rd person singular masculine subject (perfective), *-hie* for 3rd person singular feminine direct object, *-lu* for 3rd person singular masculine indirect object, and *-x* for verb negation. This ready potential for inflectional complexity is another Semitic feature of Maltese which applies across the board, whatever the origin of the verb. It also raises a host of interesting questions concerning the nature of lexical entries, the relationship between lexical entries and surface strings, and the kind of morphological processing that is necessary to connect the two together.

Many of the linguistic issues that could help to resolve these questions are themselves unresolved for lack of data - which could take the form of suitably organised language resources.

For this reason, we see the design/implementation of the lexicon, the development of language resources, and the evolution of linguistic theory for Maltese as three goals which must be pursued in parallel.

The Maltilex Project

Notwithstanding the wider context mentioned above, the stated aim of Maltilex is to develop a computational lexicon for Maltese.

At the outset of the project (see Rosner et. al. 1999) it seemed clear that any such undertaking would involve two rather separate subtasks: the identification of a set of lexical entries, and the population of these entries with different kinds of linguistic information.

Identification of Lexical Entries

We initially approached the problem of lexeme identification by using entries from printed dictionaries such as Aquilina (1997). It soon became clear, however, that this approach was not going to succeed, partly because of copyright problems, but more importantly, because of the limited scope of any fixed word list, and the questionable relevance of many of the entries to real-world applications. For these reasons, we decided to opt for a strictly empirical approach based on the extraction of entries from naturally occurring raw text.

Raw text is first pre-processed, leaving a large, completely unstructured set of tokens, some of which will be related to each other through the rich set of morphological transformations that Maltese allows. The task of sifting through a few hundred thousand such tokens in order to discover genuine lexical entries is not only extremely laborious, but requires considerable linguistic expertise: a very good argument for postulating computer-supported methods. A very general formal outline of the algorithm is sketched in Micallef and Rosner (2000).

This work has now come to fruition in the form of Dalli's (2002) Lexicon Structuring Technique, which identifies lemmas in an unstructured list of words without appeal to any predefined rules. It works using a set of statistical techniques adapted from bioinformatics algorithms that are usually used to structure genome data. To give a concrete example, the algorithm is capable of 'discovering' not just that the following words, occurring within a much larger set, can be clustered together, but also what sequence of letters can be regarded as the optimal intersection that best characterises the set.

___kel__ma__
___k_l_iem___
___kel__miet_
__tkellem___
__tkell_mu__
__tkell_met_
__tkellim_tu

This turns out to be the sequence "*kelm*", which is very close to the Semitic root (*k-l-m*) of the words. Of course, not all cases are as clear-cut as this one.

Semantic Tagging

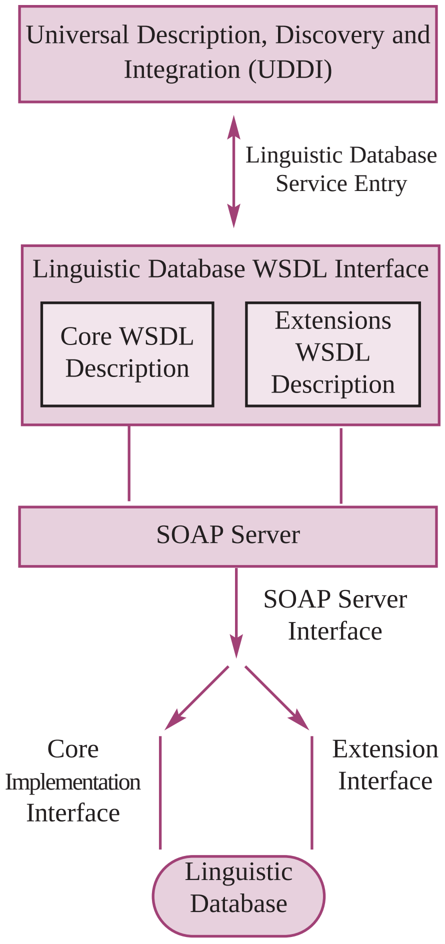
The aim of Dalli's algorithm is to relieve the linguist of some of the more tedious aspects of lexicon building. Amongst the tasks that are left over are population of the entries with appropriate information. For this to be possible, a language for expressing the content of lexical entries must be used and efforts are underway to define what Gatt (2001) refers to as a "knowledge backbone through which re-classification of lexical items according to their grammatical category and morphosyntactic characteristics can take place". This goes hand in hand with the development of a suitable tagset for Maltese based on EAGLES guidelines. Some of the special features of the tagset are described in Gatt and Dalli (2002, forthcoming).

Lexicon Server

There is a tendency within the linguistic community to regard the lexicon as a static and passive repository of linguistic data that is used in much the same way as a store cupboard for linguistic data. For Maltilex, the lexicon is somewhat different, being based on the view that the main point of a lexicon is to provide different kinds of services to different categories of human and non-human user.

For example, a common service provided by a lexicon is "lookup". A string is supplied and a definition of some kind is returned, if the word is present. Clearly, the way in which this operates will depend on the kind of user. An ordinary human user will probably require form-based input, with a carefully designed visual presentation of results. The interaction protocol would be quite different for non-human application programs. The lexical interface of a sentence parser, for example, should be completely functional, reliable and efficient. In many cases, the only result required is a list of possible categories. All this can be determined at the level of a suitably defined API. Besides lookup, a whole other range of other services is associated with the lexicon - maintenance and extension for example.

In short, the Maltilex view of the lexicon is as a collection of services delivered using different protocols. To accommodate this, we are currently experimenting with architecture shown below. At the lowest is the core lexical information, stored in an efficient relational database. Basic lexicon services are delivered using a SOAP (*Simple Object Access Protocol*) server which provides XML-based interactions between different linguistic databases and systems over the HTTP protocol. Data records can be imported and exported in XML format and converted into efficient relational records transparently. WSDL (*Web Services Description Language*) is used to describe the services provided by the linguistic database system in a standard manner, significantly reducing the development time for the implementation of new clients. Finally, recent developments like UDDI may be used to facilitate the development of flexible and secure but easily accessible linguistic databases and processing resources.


Universal Description, Discovery and Integration (UDDI)
Linguistic Database Service Entry
Linguistic Database WSDL Interface
Core WSDL Description
Extensions WSDL Description
SOAP Server
SOAP Server Interface
Core Implementation Interface
Extension Interface
Linguistic Database

The Future: Integrated Maltese-Enabled Applications

It is still early days to be talking about applications although a number of prototypes have been produced as undergraduate dissertations in the following areas:

- OCR for Maltese
- Maltese Legal Documentation Classifier and Server
- Maltese Spell Checker
- Email Classification and Response Processing

One of our aims is to develop these prototypes into well engineered artifacts that could in principle be made into products. There is also considerable scope for integration. For example, a spell checker can be used to improve OCR, and OCR can be used to provide further content for a text archive which we hope to develop along the lines of British National Corpus.

These and other developments are currently under discussion. Further information about the Maltilex project is available at http://mlex.cs.um.edu.mt.

Acknowledgements

The work reported in this document is partly supported by the University of Malta and the Mid-Med Computer Commerce Foundation. I extend my thanks to other members of the Maltilex team (Angelo Dalli, Albert Gatt, Ray Fabri, Joe Caruana) for their helpful comments and advice.

References

J. Aquilina, *Maltese-English Dictionary*, Valletta: Midsea Books, 1977.

Box, Don et al. *Simple Object Access Protocol (SOAP)* 1.1. W3C Note, http://www.w3.org/TR/SOAP, May 2000.

Christensen, Erik et al. *Web Services Description Language (WSDL)* 1.1. W3C Note, http://www.w3.org/TR/wsdl, March 2001.

Dalli, *Interoperable Extensible Linguistics Databases*, presented at IRCS Workshop on Linguistic Databases, University of Pennsylvania, December 2001

Dalli, *Biologically Inspired Lexicon Structuring Technique*, to be presented at Human Lan Language Technology Conference, University of San Diego, March 2002.

G. Falzon Basic *Maltese Grammar*, www.aboutmalta.com/grazio/maltese-grammar.html, 1977.

A. Gatt, *Linguistic and Computational Aspects of the Design of an XCES-EAGLES Compatible Tagset for Maltese*, Maltilex Technical Report 2001.

A. Gatt and A. Dalli, *A Formal Framework for the Evaluation, Optimisation and Mapping of Annotation Models and Lexical Representation*. To appear, 2002

P. Micallef & M. Rosner, *The Development of Language Resources for Maltese*, Proc. Workshop on Developing Language Resources for Minority Languages, LREC 2000, Athens.

M. Rosner et. al. 1999. *Linguistic and Computational Aspects of Maltilex*. Proc. of the ATLAS Symposium, Tunis.

E.F. Suttcliffe, *A Grammar of the Maltese Language*, Valletta: Progress Press (1936)

Michael Rosner
Department of Computer Science and AI
University of Malta
MSD06 Msida, MALTA
Tel. +356 3290 2505
Fax. +356 32 05 39
Email: mike.rosner@um.edu.mt
Web site: www.cs.um.edu.mt/~mros
Maltilex project Web site:
http://mlex.cs.um.edu.mt.

IHM Group, ENSIAS, Rabat, Morocco
Overview of the research activities

Abdelhak Mouradi

Introduction

Human Language Technologies in general and Speech Technologies in particular play a major role within the field of Information and Communication Technologies. As speech is the most natural means of communication between human beings, it seems quite natural to try to develop it as a means of communication with computers.

Language Technologies have a multiple economical and social impact. The language of an application is a decisive factor with respect to the level of acceptance of the technology by the user and it is recommended (if not compulsory) that man-machine communication takes place in the user's native language.

For these reasons, and given the human potential existing in Morocco for contributing to the development of language technologies in the Arabic language, ENSIAS has initiated a number of activities in this field, both from the educational and research viewpoints. This work takes place mainly within the IHM Group, itself embedded in a network of other research groups with similar focus.

Presentation of the IHM Group

The IHM Group, which stands for Interfaces Homme-Machine, i.e. Man-Machine Interfaces, is one of the three research groups within the Information Technology Engineering Group of ENSIAS (Ecole Nationale Supérieure d'Informatique et d'Analyse des Systèmes). The focus of the IHM Group is to promote research in the following fields :

1. Speech Processing
   a. Speech Synthesis
   b. Speech Recognition

2. Mobile Communication Systems
   a. Data Protection Methods in Noisy Channels
   b. Error Detecting and Correcting Codes
   c. Algorithms and Implementation Issues

3. Real-Time Applications
   a. Parallel System Control
   b. Synchronous Languages-Application to Real-Time Systems

4. Multimedia and Databases
   a. Image Databases in a Multimedia Environment
   b. Image Processing and Object-Oriented Databases

Research in the field of Speech Synthesis and Recognition

In the field of Speech Processing, most of our activities are focused on speech synthesis in Arabic, especially the investigation of various approaches and techniques, using different types of units in concatenative synthesis.

1. Speech recognition

Most research activities worldwide in the domain of Large Vocabulary Continuous Speech Recognition (LVCSR) have been focused mainly on English, French and other European or Asian languages. Languages such as Arabic have been poorly studied. The Arabic language has a number of features that must be taken into account in Automatic Speech Recognition systems. For instance, in most cases, texts in Arabic do not contain short vowels, which makes them hard to exploit as written data for language model learning.

We are aware of how important Automatic Speech Recognition is becoming and of the limited amount of work on Arabic in this field. We intend to put some effort in this direction.

2. Speech synthesis

Our main activity in speech processing has been focused on Speech Synthesis in Arabic.

We have designed one of the first Speech Synthesis systems for the Arabic language, based on the diphone concatenation approach, using the Linear Prediction Coding (LPC) technique. There are 28 consonants and 6 vowels in Arabic (3 short vowels and 3 long vowels). Therefore, if we include the silence that can occur at the beginning and at the end of a word, this yields 35 x 35 = 1225 diphones.

The input of the system is a text in Arabic characters including vowels, as well as numbers and punctuation signs. The synthesis of texts without vowels raises additional issues, requiring syntactic, semantic and sometimes pragmatic analysis, which goes beyond the framework of speech synthesis. Our grapheme-phoneme transcription module is based on pronunciation rules and a dictionary of exceptions. The transcription algorithm uses a variable-width sliding window for text analysis. The phonetic transcription is produced as soon as the width of the window is large enough so that a pronunciation rule can be applied.

The validity and the limits of the diphone as a speech synthesis unit for standard Arabic have been identified and discussed in a paper published in the Journal d'Acoustique in 1989. One of the conclusions of this article is that the diphone is poorly adapted to units composed of emphatic consonants and vowels, because emphasis can range beyond the neighbouring vowel. Therefore, we have proposed to enlarge the vocalic system in Arabic by including six other vowels, which we call emphatic vowels, and which complement the six conventional vowels.

In the context of PhD works, two synthesis systems have been implemented.

The first system relies on formant-based synthesis for a rule-based system in Arabic. This system is inspired of the well-known Klatt's system, simulated with a C++ software.

The rule-based system takes as input a file of phonetic symbols and punctuation signs. These symbols are converted by a set of rules and an acoustic-phonetic dictionary into a series of time-varying parameters, which control the synthesizer.

The second system relies on a concatenative text-to-speech approach using di-syllabic units and the TD-PSOLA (*Time-Domain Pitch-Synchronous OverLap and Add*) synthesis technique. A di-syllable can be defined as a speech segment that ranges between the stable vocalic centre of a given syllable and the stable vocalic centre of the next one.

Speech obtained with the two systems described above is intelligible and is judged of acceptable quality by a large majority of the subjects who participated in the listening tests.

In order to improve the quality, it is necessary to investigate on prosody and include the results of these studies in the synthesis system. In the field of computational linguistics, similarly to the field of speech recognition, the Arabic language requires specific research, in order to define adequate prosodic models.

We have therefore decided to improve the system based on di-syllables and PSOLA, which we have called PARADIS (*Psola ARAbic DI-syllable concatenation based System*), by incorporating prosodic aspects. Research has started in the context of PhDs preparation, to study intonation, duration and rhythm.

The first work deals with the generation of pauses and syllable duration. The goal is to derive two models: one which accurately renders syllable lengthening corresponding to the various modalities of Arabic sentences, and another which describes the distribution and the duration of pauses.

The second work is dedicated to intonation. The goal is to analyse local events in relation to stress, as well as the melody declination, which corresponds to the overall intonation pattern in an utterance.

Past and ongoing projects

These research works have benefited from the support of the FRANCIL thematic network set up by AUPELF-UREF, in the framework of the Joint Research Action between our group and LIMSI, entitled “study of the dialectal variations in Moroccan French”. Several researchers from our group have been able to spend some time in LIMSI and benefited from the scientific context of this laboratory.

In the framework of the PROTARS Programme (*Programme Thématique d'Appui à la Recherche Scientifique*) initiated by the Ministry of Higher Education, Executive Training and Scientific Research of the Kingdom of Morocco, and following a call for proposals, we have submitted a project entitled “Study, design and implementation of a man-machine dialogue system in Arabic based on speech synthesis and recognition”. This project has been accepted by the National Commission for Evaluation and a budget has been allocated to it. The contract is currently being finalised.

Abdelhak Mouradi
ENSIAS B.P. 713 Agdal
Rabat, Maroc
Tel.: 037 77 73 17
Fax: 037 77 72 30
Email: mouradi@ensias.ma

CLEF 2002 (Cross-Language Evaluation Forum)- CALL FOR PARTICPATION

The CLEF series of system evaluation campaigns aims at promoting research and development in mono- and cross-language information retrieval for European languages.

Registration is now open for CLEF 2002.

The main track in CLEF2002 tests multilingual IR systems. Additional tracks will offer evaluation for bilingual and monolingual (non-English) systems on general-purpose and scientific test collections. There will also be a track testing interactive cross-language systems.

The CLEF test collection for 2002 consists of a multilingual corpus of newspaper and newswire documents for English, French, German, Italian, Spanish, Dutch and Finnish plus collections of scientific documents in French and German.

Important Dates:

Data Release: 1 February 2002
Topic Release: 1 April 2002
Submission of runs by participants: 15 June 2002
Release of relevance assessments and individual results: 1 August 2002
Submission of paper for Working Notes: 1 September 2002
Workshop - 19-20 September 2002, Rome, Italy (in conjunction with ECDL 2002)

For full details on the CLEF Agenda and Task Description for 2002 and instructions on How to Participate, see:

http://www.clef-campaign.org

For further information, contact:
Carol Peters - IEI-CNR
Tel: +39 050 315 2987/ Fax: +39 050 315 2810 / E-mail: carol@iei.pi.cnr.it

LREC 2002 News

Third International Language Resources and Evaluation Conference

Main Conference: 29th, 30th & 31st May 2002

Over 360 submissions have been accepted by the LREC 2002 Programme Committee. These proposals are listed on the LREC 2002 web site, in the 'Program' section at the following address: http://www.lrec-conf.org/lrec2002/index.html.

Besides, four panels have been selected: "ethical and legal issues in corpus construction", organised by Tony McEnery, "language resources strategy panel" by Mark T. Maybury, "the Open Language Archives Community" by Steven Bird, and "Standards & best practices for multilingual computational lexicons", by Nicoletta Calzolari.

As for the workshops, 18 pre-conference (9) and post-conference (9) workshops are to be organised. These are shown below.

Pre-conference workshops

Pre-Conference Workshops: 27th & 28th May 2002

26th & 27th May 2002

- International workshop about resources and tools in Field Linguistics (Peter Wittenburg)

27th May 2002

- OntoLex'2 - Ontologies and lexical knowledge bases (Kiril Simov)
- Machine translation evaluation - Human evaluators meet automated metrics (Maghi King)
- Workshop on annotation standards for temporal information in natural language (Andrea Setzer)

28th May 2002

- Question answering - Strategy and Resources (Mark T. Maybury)
- Customizing knowledge in NLP applications (Federica Busa)
- International standards of terminology and language resources managament (Key-Sun Choi)
- Language resources in translation work and research (Elia Yuste Rodrigo)
- Workshop on Wordnet structures and standardization, and how these affect Wordnet applications and evaluation (Dimitris N. Christodoulakis)

Post-conference workshops

Post-Conference Workshops: 1st & 2nd June 2002

1st June 2002

- Linguistic knowledge acquisition and representation: bootstrapping annotated language data (Alessandro Lenci)
- Arabic language resources and processing: status and prospects (Khalid Choukri)
- Multimodal resources and multimodal systems evaluation (Mark T. Maybury)
- First International workshop on UNL, other interlinguas and their applications (Jesus Cardenosa)
- Portability Issues in HLT (Bojan Petek)

2nd June 2002

- Towards a roadmap for multimodal language resources and evaluation (Steven Krauwer)
- Using semantics for information retrieval and filtering (Claude de Loupy)
- Event modelling for multilingual document linking (Roberta Catizione)
- Beyond PARSEVAL - Towards improved evaluation measures for parsing systems (John Carroll)

New Resources

ELRA-W0030 Arabic Data Set

The corpus was developed in the course of a research project at the University of Essex, in collaboration with the Open University. The corpus contains Al-Hayat newspaper articles with value added for Language Engineering and Information Retrieval applications development purposes. Data has been organised in 7 subject specific databases according to the Al-Hayat subject tags. Mark-up, numbers, special characters and punctuation have been removed. The size of the total file is 268 MB. The dataset contains 18,639,264 distinct tokens in 42,591 articles, organised in 7 domains

	ELRA Members	Non Members
Price for research use	480 Euro	720 Euro
Price for commercial use	960 Euro	1,440 Euro

ELRA-W0031 GeFRePaC - German French Reciprocal Parallel Corpus

The German-French Reciprocal Parallel Corpus (GeFRePaC) was produced by the Multilinguale Forschung/Multilingual Research Abteilung Lexik, Institut für Deutsche Sprache (Germany) through a funding from ELRA in the framework of the European Commission project LRsP&P (Language Resources Production & Packaging - LE4-8335). The German-French Reciprocal Parallel Corpus (GeFRePaC) is a 30 million word corpus (15 million for each language) for the purpose of developing, enhancing and improving translation aids (dictionaries, lexicons, platforms) for French-German and German-French translation.

The database consists of the following parallel corpora:

- European Union CELEX Database: Treaties, Foreign relations, Law, Complementary Law and all the published documents of the "European Parliament".
- Celex-Database: 22,000,000 words (German+French) (http://www.outlaw-web.com)
- Europarl: 8,320,000 words (German+French) (http://www.europarl.eu.int)

It covers natural general language as used in public socio-political discourse and it has a focus on multilingual administration and commercial and legal documentation. GeFRePaC comprises a large variety of text types for which there is a rapidly growing need for translation but which currently defy successful machine translation. The corpus is encoded according to the PAROLE guidelines, it was aligned on the sentence level and also for single word translation units on the lexical level, POS-tagged in conformity with EAGLES recommendations and validated according to the most current version of the ELRA guidelines. The parallel German-French texts were aligned using a program developed at the Equipe Langue et Dialogue, Laboratoire Loria, Nancy. The text files containing markup for paragraphs and sentences were processed by the Tree Tagger developed at the IMS Stuttgart. The text files are automatically converted into TEI-conformant SGML format.

	ELRA Members	Non Members
Price for research use	1,500 Euro	2,500 Euro

ELRA-W0033 CRATER 2

The CRATER corpus was built upon the foundations of an earlier project, ET10/63, which was funded in the final phase of the Eurotra programme. The Corpus Resources and Terminology Extraction project (MLAP-93 20) extended the bilingual annotated English-French International Telecommunications Union corpus produced within ET10/63 to include Spanish.

The CRATER 2 corpus was produced by the Department of Linguistics & Modern English Language, Lancaster University (United Kingdom) with funding from ELRA. The ELRA funding in turn was provided by the European Commission project LRsP&P (Language Resources Production & Packaging - LE4-8335). This project has enhanced the CRATER corpus, available under the reference ELRA-W0003 in the ELRA catalogue. CRATER 2 has significantly expanded the French/English component of the parallel corpus by increasing the size of the English/French corpus from 1,000,000 words per language to approximately 1,500,000 tokens per language.

CRATER 2 is sold in with CRATER in a single package.

	ELRA Members	Non Members
Price for research use	25 Euro	125 Euro
Price for commercial use	25 Euro	125Euro

ELRA-W0032 Modern French Corpus including Anaphors Tagging

The corpus that includes the tagging of the anaphors was created by the CRISTAL-GRESEC (Stendhal-Grenoble 3 University, France) team and XRCE (Xerox Research Centre Europe, France) in the framework of the call launched by the DGLF-LF (national institution for the French language and the languages spoken in France), for the creation of modern French corpora).

Over 1 million words have been annotated. The corpora have been selected so that they represent a wide sampling of the French language (scientific and human science articles, extracts from newspapers and magazines, legal texts, etc.) and according to the points of interest of the teams working on the project. The processed corpora supplied by ELRA are listed below:

1) Two books edited by the CNRS:

- *La protection des oeuvres scientifiques en droit d'auteur français*, Xavier Strubel. Paris, CNRS Editions, 1997 (77 591 words).
- *Cinquante ans de traction à la SNCF. Enjeux politiques, économiques et réponses techniques*, Clive Lamming. Paris, CNRS Editions, 1997 (124 990 words).

2) Newspapers' articles:

- 204 articles extracted from *CNRS Info*, a magazine which contains short popular scientific articles from the CNRS laboratories (201 280 words).
- 14 articles dealing with *Hermès* Human Sciences (111 886 words).
- 136 articles extracted from *Le Monde*, dealing with economics (roughly 180 760 words).
- 13 booklets of the *Official Journal of the European Communities* (roughly 337 000 words).

Below the tagged anaphoric elements:

- Person pronouns: 3rd person pronoun, anaphoric.
- Possessive determiners: 3rd person possessive determiner.
- Demonstrative pronouns: anaphoric pronouns (celui, celle, ceux, celles-ci, celles-là).
- Indefinite pronouns: Aucun(e), chacun(e), certain(e)s, l'un(e), les un(e)s, tout(es), etc, when they are anaphoric.
- "Proverbs": "le" + "faire".
- Anaphoric and cataphoric adverbs: Dessus, dedans, dessous , when they have an anaphoric function.
- Ellipsis of head nouns: Nominal adjectives or quantifiers determiners ellipsis.
- Textual headers like "ce dernier": Ce dernier, le premier , etc.

The annotation scheme was defined in XML format. The texts were divided into sections, paragraphs (<p>) and sentences (<s>). The sentence segmentation was carried out with NLP tools developed by XRCE, the annotation part was done manually by two qualified linguists. A large subset of anaphoric phrases was automatically pre-annotated. The antecedents and the tagging of the anaphoric relations were manually processed, but editing tools (emacs, macros from Author/Editor software) were used to make it easier. 5% of the corpora were evaluated to check the annotation reliability.

	ELRA Members	Non Members
Price for research use	250 Euro	1,000 Euro

ELRA-S0119 Spanish SpeechDat database for the mobile telephone network

The Spanish SpeechDat database for the mobile telephone network comprises 1066 Spanish speakers (526 males, 540 females) calling from GSM telephones and recorded over the fixed PSTN using and ISDN-BRI interface. The database was produced by Applied Technologies in Language and Speech S.L. (Spain). The MDB-1000 database is partitioned into 6 CDs in ISO 9660 format. This database follows the specifications given in the framework of the SpeechDat(II) project.

Speech samples are stored as sequences of 8-bit 8 kHz A-law. Each prompted utterance is stored in a separate file. Each signal file is accompanied by an ASCII SAM label file which contains the relevant descriptive information.

Each speaker uttered the following items:

· 2 isolated digits

· 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 prompted date (word style), 1 relative and general date expression.

· 1 word spotting phrase using an application word (embedded).

· 6 application words

· 3 spelled words: 1 spontaneous name (own forename), 1 city name, 1 real / artificial word for coverage.

· 1 currency money amount.

· 1 natural number.

· 6 directory assistance names: 1 surname (set of 500), 1 city of birth / growing up, 1 most frequent cities (set of 500), 1 most frequent company / agency (set of 500), 1 'forename surname' (set of 150), 1 spontaneous forename.

· 2 questions including 'fuzzy' yes / no: 1 predominantly 'Yes' question, 1 predominantly 'No' question.
· 9 phonetically rich sentences.
· 2 time phrases: 1 time of day (spontaneous), 1 time phrase (word style).
· 4 phonetically rich words.
· Call environment.

The following age distribution has been obtained: 5 speaker are below 16 years old, 543 speakers are between 16 and 30, 307 speakers are between 31 and 45, 202 speakers are between 46 and 60, 9 speakers are over 60.

A pronunciation lexicon with a phonemic transcription in SAMPA is also included.

	ELRA Members	Non Members
Price for research use	22,000 Euro	25,000 Euro
Price for commercial use	28,000 Euro	35,000Euro

Joint cooperation between ELDA and LDC - Distribution of language resources

Networking Data Centers, "*Net-DC*",(MLIS-5017), aims to improve the infrastructure for language resources, by designing and implementing new modes of cooperation between the Linguistic Data Consortium (LDC) and the European Language Resources Distribution Agency (ELDA). In the framework of this cooperation, LDC and ELDA are happy to announce the following joint distribution of language resources.

- **TED** (*Translanguage English Database*):

ELRA reference: ELRA-S0031 http://www.elda.fr/cata/speech/S0031.html

LDCreference: http://www.ldc.upenn.edu/Catalog/LDC2002S04.html)

- **TED** (*Translanguage English Database*) **Transcripts Database**:

ELRA reference: ELRA-S0120 http://www.elda.fr/cata/speech/S0120.html

LDC reference: http://www.ldc.upenn.edu/Catalog/LDC2002T03.html)

ELRA-S0031 Translanguage English Database (TED)

The Translanguage English Database (TED) is a corpus of recordings made of oral presentations at Eurospeech'93 in Berlin. The corpus name derives from the high percentage of oral presentations given in English by non-native speakers of English. Two hundred twenty-four (224) oral presentations at the conference were successfully recorded, providing a total of about 75 hours of speech material. These recordings provide a large number of presenters, speaking multiple variants of English, over a relatively large amount of time (15 minutes for each presentation + 5 minutes of discussion), on a specific topic. This release of TED (6 CDROMs) includes 188 speeches without the ensuing discussion period. This database was produced with the support of ELSNET. Associated text materials consist of ASCII versions of over 400 proceedings papers and oral preparations that were supplied by the authors, as well as, 250 speaker questionnaires.

	ELRA Members	Non Members
Price for R & C use	Free	300 Euro

ELRA-S0120 Translanguage English Database (TED) Transcripts database

The Translanguage English Database (TED) Transcripts corpus contains transcriptions of thirty-nine of the 188 presentations of the TED Corpus (ELRA ref.: ELRA-S0031; LDC ref.: LDC2002S04) made at Eurospeech'93 in Berlin. The thirty-nine transcripts in this publication are in Universal Transcription Format (UTF) and were prepared by the LDC. All utf files in the transcript publication were validated against an included utf.dtd. Tables containing speaker demographic information and a cross-reference of file names from the TED audio corpus are included.

	ELRA Members	Non Members
Price for R & C use	Free	115 Euro