# The ELRA Newsletter

EUROPEAN LANGUAGE RESOURCES ASSOCIATION

ELRA

April - June 2005

Vol.10 n.2

## *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:***
*Joseph Dichy*
*Mohamed Hassoun*
*Gianni Lazzari*

*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,*

We are happy to announce two major events being organised by ELRA :

- **HLT Evaluation workshop**, taking place in Sliema (Malta) on December 1st and 2nd 2005 to celebrate ELRA‘s 10th anniversary. In organizing this workshop, ELRA intends to bring together the HLT Evaluation key players to discuss HLT evaluation from various perspectives and to allow a fruitful brainstorming on HLT evaluation, starting from what is being done today, what should be done better, differently, which approaches should be followed, etc. All sectors of HLT will be addressed: speech technologies, machine translation and speech to speech translation, information retrieval/filtering, multimodal interfaces, etc. If you would like to participate, please contact Hélène Mazo (mazo@elda.org). Please bear in mind that the workshop will be held with a limited number of participants.

A detailed programme will be available soon.

- **LREC 2006**, taking place in Genoa (Italy) from May 22nd to May 28th 2006. The second call for papers detailing the workshop and abstract submission procedure has been issued and is publicized in the newsletter. The deadline for abstract submission has been extended to October 20th.

We are also proud to announce the setting up of a fidelity program to reward ELRA's loyal members. The principle behind the fidelity program is to earn miles by joining and remaining member of our association. A detailed description of this programme is provided on page 12.

Over the past months, ELRA and ELDA have continued to be involved in a number of co-funded projects, both at a European and national level. The main focus of those projects lies on the evaluation of HLT and the production of language resources. One of these projects is described in this newsletter: TC-Star. The TC-Star consortium brings together prominent European players in the Speech technology field. The consortium offers a unique opportunity to place Europe in a position of leadership in Speech to Speech Translation (TTS) technologies.

This newsletter also contains a description of “The DIINAR.1 Arabic Lexical Resource”, with an outline of contents and methodology. This paper aims at presenting the Arabic monolingual lexical resource DIINAR.1 available soon to researchers and developers as a general lexicon through ELRA/ELDA catalogue. An outline of the format, contents, number of entries and purpose of the resource, as well as a brief historical and methodological survey, are given. Basic concepts and underlying representations are then summarized and followed by a few words on future prospects.

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

- S0175 Mandarin Chinese Speecon database
- S0176 Finnish Speecon database
- S0177 Korean Speecon database
- S0178 Turkish Speecon database
- S0179 Polish Speecon database
- S0180 Portuguese Speecon database
- M0041 Bulgarian WordNet
- L0056: STO SprogTeknologisk Ordbase (Danish Lexicon for NLP/HLT Applications)

Once again, if you would like to join ELRA and benefit from its services (that are summarized at www.elra.info), please do not hesitate to contact us.

Bente Maegaard, President　　　　Khalid Choukri, CEO

# TC-Star Project : Technology and Corpora for Speech to Speech Translation

*Gianni Lazzari*

In Europe, a large number of national and international initiatives have taken place in recent years, covering core language technologies such as speech recognition, machine translation, speech synthesis, information retrieval, and question-answering systems. A number of successful technologies have been developed and some even successfully deployed. However, there remain several areas where technologies are difficult to deploy, a big factor being unknown performance.

The importance of language will become evermore critical as the EU moves forward as an information society. Language is the main instrument of communication for work, travel, and home. The social costs of language translation cannot be ignored.

Clearly public institutions and the private sector need a solution to the issue of language translation. Automated language processing technology can provide solutions for translation, information query, and other cross-lingual applications. In fact, language technologies have been a strategic research topic since the V Framework program. The result is that some European public research institutions and private companies have developed leading language technologies. They are competing well against US institutions at the technology level.

Performance evaluation is one of the most important objectives of a project called TC-STAR (Technology and Corpora for Speech to Speech Translation), funded by the European Commission, VI Framework Programme. The TC-STAR Consortium-www.tc-star.org-brings together prominent European players in the speech technology field, including both basic research institutes and universities, as well as large companies with years of experience in the field. The Consortium itself is a first step towards reinforcing the European Research Area, and offers a unique opportunity to place Europe in a position of leadership in speech-to-speech translation (SST) technologies.

SST technology is a combination of Automatic Speech Recognition (ASR), Spoken Language Translation (SLT) and Text to Speech (TTS) (speech synthesis). The objective of the project is ambitious: achieve a breakthrough in SST that significantly reduces the gap between human and Machine Translation (MT) performance.

The project targets a selection of unconstrained conversational speech domains-speeches and broadcast news-and three languages: European English, European Spanish, and Mandarin Chinese. Accurate translation of unrestricted speech is well beyond the capabilities of today's state-of-the-art research systems. Therefore, incremental as well as breakthrough advances are needed to improve state-of the-art technologies for speech recognition and speech translation.

The long-term research goals of the project are:

- Effective SLT of unrestricted conversational speech on large domains of discourse.

- Speech recognition able to perform reliably under varying speaking styles and recording conditions, and for different user communities.

| TC-STAR Statistics | |
|---|---|
| EU grant | €11 million |
| Total Cost | €18 million |
| Total Effort | 150 person years |
| Start | April 2004 |
| End | March 2007 |

- Effective integration of speech recognition and translation into a unique statistically sound framework.

- General expressive speech synthesis imitating the human voice.

Project success will be measured by the progress achieved in each component of SST technology as well as in the end-to-end systems. Key actions to meet these grand challenges can be summarized as follows:

Evaluation infrastructure: an evaluation infrastructure is implemented through the organization of periodic competitive evaluations of single components for ASR, SLT, TTS and end-to-end systems. Three evaluation campaigns are planned to measure progress by all partners on common language resources and under equal conditions. Improvements in methods and technology are systematically demonstrated on common test sets using common evaluation metrics. Improvements are measured against state-of-the-art reference baselines established by the project.

Sharing of Knowledge: to complement the competitive evaluations the participants are required to share the knowledge gained in the process. For this purpose, an evaluation workshop, open to external participants, is organized after each evaluation campaign. This ensures that the techniques and ideas that prove most successful in an evaluation campaign become available to participants. Moreover, techniques and components with proven success are shared and combined to obtain the best performing TC-STAR system.

Technological Infrastructure: a technological infrastructure is developed to foster effective delivery and assessment of scientific results. The single components developed by each partner are plugged into the common platform architecture and made accessible to the whole consortium. This ensures that all components can be evaluated and integrated by all partners. The architecture is also the back-bone for the implementation of showcases.

The project brings together key SST actors to form a critical mass of researchers. The project participants are:

*Istituto Trentino di Cultura - Centro per la Ricerca Scientifica e Tecnologica (ITC-irst)
*Rheinisch-Westfälische Technische

Hochschule Aachen (RWTH-AACHEN)
*Centre National de la Recherche Scientifique (CNRS-LIMSI)
*Universitat Politècnica de Catalunya (UPC)
*Universität Karlsruhe (TH) (UKA)
*IBM Deutschland Entwicklung GmbH (IBM)
*Siemens Aktiengesellschaft (SIEMENS)
*Nokia Corporation (NOKIA)
*Sony Deutschland GmbH (SONY)
*Evaluations and Language Resources Distribution Agency (ELDA)
*Stichting Katholieke Universiteit/ Speech Processing Expertise Centre (KUN-SPEX).

ITC-irst is the project coordinator. Project partners have strong expertise in multiple areas of the project: automatic speech recognition, spoken language translation, text-to-speech, and implementation of advanced technology infrastructures. The consortium is well balanced between research and industrial technology partners in the field of SST, and also includes centres for language resource distribution and validation.

Beyond exchanging new knowledge through the evaluation workshops, scientific achievements are disseminated to the worldwide scientific community through participation at major international scientific conferences and publications in journals covering all research areas related to SST.

The most significant results of the project, in terms of overall advances in SST technology, are expected in the mid- to long-term. Due to inadequate current performance, SLT technology is not yet ready for widespread introduction to the SECOND CALL FOR PAPERS market. The purpose of TC-STAR is to push SLT performance together with the required functionality of ASR and TTS in order to prepare for market adoption. This approach should allow improvements of the functionality of existing products based on ASR and TTS and establishes also a base for introducing new translation products for face-to-face and over-the-phone conversations, speeches, documents (or web sites), cross-lingual retrieval in audio streams, etc.

Currently, the main market segments of voice-driven interfaces are network-based services, mobile terminals, and automotive applications. Network-based services represent the largest market segment, currently dominated by IVR systems. Voice-driven mobile phones are the largest segment within the mobile terminal market. In the automotive market, speech recognition is moving from the high end to the upper mid-range cars, offering services such as mobile-phone and navigation system control.

Even though an overestimation of the capabilities of these technologies in past years caused negative perception in the end-user market, it seems that a more mature and positive phase is now ahead of us. This phase is characterized by a more realistic view of the usability of spoken language technologies.

### *Results of the first project year*

At the early stage of the project, a suitable and challenging reference task was identified. It targets the translation of speeches delivered during the European Parliament Plenary Sessions (EPPS). This makes TC-STAR the first European project on spoken language translation working on a non restricted real-life task. Rapid focus on tackling the EPPS tasks pushed TC-STAR far ahead with respect to planned progress. Two translation directions were explored: from English to Spanish and from Spanish to English. Appropriate language resources were specified for these languages and the process for their production and validation started immediately. In the meantime, baseline systems for speech translation and automatic speech recognition were developed by exploiting publicly available language resources. To create the baseline voices and to support the research tasks, the specifications of language resources for speech synthesis were also defined.

An evaluation of baseline systems for ASR and SLT took place during September and October 2004 while during March 2005 the first TC-STAR evaluation campaign was carried out. The aim was to measure progress made in automatic speech recognition and spoken language translation. For this purpose training, development, and evaluation data sets for the EPPS task were made available to the TC-STAR partners as well as to external participants. Components for speech recognition and translation were also evaluated on a Broadcast News (BN) translation task with translation from Mandarin Chinese into English. In addition to single component evaluations for ASR and SLT, full systems that perform speech recognition and translation were evaluated.

In addition to TC-STAR partners, two external sites participated in the first evaluation campaign. Results of this evaluation campaign were presented and discussed in the first TC-STAR Evaluation Workshop held in Trento (Italy) in April 2005.

With respect to speech synthesis, the evaluation criteria were defined. Their aim is to evaluate the speech synthesis component as a whole and also its single modules. Furthermore, some evaluation tests were defined for specific research tasks.

Initial prototyping work explored different implementation and scenario aspects for the TC-STAR technological infrastructure. The findings have been used to develop the design and requirement documents to cover the two TC-STAR scenarios, namely for automated competitive evaluation and showcases. The TC-STAR infrastructure implementation will be based on the Unstructured Information Management Architecture (UIMA).

In conclusion, the most significant results of the first year of the project were:

- Significant improvement of the state-of-the-art in speech-translation for an unrestricted real-life task (EPPS).
- Development of an evaluation package for speech translation (available for the speech and translation community).
- Demonstration that the competitive research paradigm adopted is effective in the context of a European project.

Given the results achieved, future activity for the project will continue using the adopted strategy based on "competitive and cooperative" research among partners.

Gianni Lazzari
Centro per la Ricerca e Tecnologica
(ITC-IRST)
lazzari@itc.it
www.itc.it

# *The DIINAR.1 - « معـالي » Arabic Lexical Resource, an outline of contents and methodology*

*Joseph Dichy; Mohamed Hassoun*

## *Introduction*

This paper aims at presenting the Arabic monolingual lexical resource DIINAR.1("DIctionnaire INformatisé de l'ARabe, version 1") - Arabic acronym Ma'âlî (Mu'jam al-'Arabiyya l-'âlî, " معـالي – معجم العربية الآلي), now available to researchers and developers as generated lexica presented in portable format under Microsoft Excel, through ELRA/ELDA (www.elda.org; also joseph.dichy@univ-lyon2.fr).

The language resource has been completed through close cooperation in two sites: in Tunisia at IRSIT ("Institut de recherche en sciences de l'informatique et des télécommunications", now IT.COM - Pr Abdelfattah Braham and Pr Salem Ghazali - Tunis), and in France at ENSSIB (M. Hassoun) and the Lumière-Lyon 2 University (J. Dichy).

Nabil Gader and Malek Ghenima participated, as doctoral students under supervision of J. Dichy (linguistics aspects) and M. Hassoun (informatics and information systems), in the elaboration of the database structure of DIINAR.1 and the completion of the user-friendly interfaces used for the input and outdating of lexical information.

An outline of the presentation, number of entries and purpose of the resource (section II) is followed by a brief historical and methodological survey (section III). Basic concepts and underlying representations are then summarized (section IV) and followed by a word on future prospects.

## *An outline of the resource presentation and format*

### A.Source program and generated lexica: why present the resource as an organised set of tables?

The source program of DIINAR.1 is structured as a database, and is of course much more compact than the set of Excel files that have been made available through ELDA. The main motive behind the choice of Excel tables is that the source program is much less transparent and portable. We have found that information included in the source program could only be freely and extensively accessed by a new researcher after a training session of at least one week. Two main reasons account for this difficulty:

a) The first reason is related to linguistic engineering. Accessing lexical resources used in NLP software (analysers, machine-translation, spelling or grammar checker, etc.) is usually rendered difficult because these resources are, to variable extents, embedded in the software that draws information from them. Lexical resources are, in addition, strongly dependent on the formalisms adopted in the applications they have been built for. This is naturally the case of the source program of DIINAR.1, which is dependent on the morphological analysers and/or generators developed for research purposes at ENSSIB and the Lyon 2 University.

b) Then comes the amount of original linguistic observation included. The structure of DIINAR.1 reflects to quite an extent the complexity of Arabic morpho-lexical structures. These structures are represented in a linguistic formalism that is not easily understood by researchers and developers, even with a good knowledge of Arabic and traditional grammar. We have come to develop analyses accounting for sets of morpho-lexical relations, some of which had not been highlighted so far in the description of Arabic. This makes them sometimes difficult to grasp in the sole light of traditional Arabic grammar.

In order to avoid the above hinders, we have come to the conclusion that Excel files and folders would allow research and development teams easy importing of linguistic information into the application or software they chose to elaborate. These files can, in addition, actually be read. From a linguistic standpoint, this feature allows extensive checking of explicit data.

### B. Number of entries

The total number of lemma-entries in DIINAR.1 is currently: 119,693. In the near future, 445 tool-words (e.g.: prepositions, conjunctions, etc.) and the prototype of a proper names database of 1,384 entries will be added. The lexical resource comprises:

| | |
|---|---:|
| Nouns, including adjectives | 29,534 |
| [Broken plural forms – جموع التكسير] | [9,565] |
| Verbs | 19,457 |
| Deverbals (مشتقات اسمية):<br>- infinitive forms (مصدر) | 23,274 |
| - active participles (اسم الفاعل) | 17,904 |
| - passive participles (اسم المفعول) | 13,373 |
| - ‘analogous adjectives’ (صفة مشبهة) | 5,781 |
| - ‘nouns of place & time’ (اسم المكان والزمان) | 10,370 |
| [Total number of deverbals] | [70,702] |
| Subtotal of lemma-entries | 119,693 |

TABLE 1: Number of lemmas-entries belonging to main major lexical categories in DIINAR.1

### C. Information associated to entries and general purpose of the resource

What has DIINAR.1 been built for? Although the resource has been shown to support a good level of syntactic analysis [Ouersighni, 2001], it has been devised to operate in the range of the word-form, i.e., in morphological analysis or generation.

Each entry has been associated with morphosyntactic specifiers allowing **morphological analysis** to perform processing of entries in standard unvowelled script, and **morphological generation** to produce fully, partly, or un-vowelled word-forms, on demand. Morphosyntactic specifiers belong to finite sets, but allow exhaustive processing of data, according to a widely original approach, which we will now summarize.

### *Brief history and underlying methodology of the project*

Broad lines of the methodology underlying the DIINAR.1 lexical resource and related morphological processors were first presented in [Desclés, dir., 1983]. It was subsequently extended and developed in Lyon (Université Lumière-Lyon 2 and ENSSIB) in the frame of the SAMIA project ("Synthèse et Analyse Morphosyntaxiques Informatisées de l'Arabe") [Dichy, 1984, 1987], [Hassoun, 1987], [Dichy & Hassoun, eds., 1989].

DIINAR.1 is based on a formal representation of written word-forms in Arabic [Dichy, 1990], and a comprehensive Word-Formatives Grammar (WFG), which includes rules explicitly written for either one of the two asymmetrical processes of generation or recognition (cf. letters "S" and "A" in the acronym SAMIA [Bouché, Dichy & Hassoun, 1984]; in English: [Dichy, 1987, 2000, 2001]). One major contribution of the SAMIA project is the highlighting, from the early 1980ies onwards, of the crucial relations between the lexical nucleus and other word-formatives, and of the subsequent need for morphosyntactic specifiers [Hassoun, 1987], [Dichy, 1984, 1990, 1997]. Specifiers, in very short words, are finite sets of morphosyntactic features associated to the entries of a lexical resource, and accounting for grammar-lexis rules and relations. Each lexical entry is associated with W-specifiers (operating at word-form level, as opposed to sentence-level S-specifiers), which allow morphological software to yield correct outputs in either analysis or generation. W-specifiers have been shown to belong to finite and exhaustive sets [Dichy, 1997, 2000]. They also include links between morphologically related items such as verb↔ deverbal(s) or singular↔'broken' plural nouns, etc.

The above research conducted in Lyon has resulted during the 1990ies in the elaboration of DIINAR.1 in close collaboration with a leading Tunisian institution, IRSIT. Research included elaboration of sophisticated and user-friendly interfaces for the input and updating of lexical information (Lyon 2 and ENSSIB), and their subsequent experimenting in common with IRSIT, where information included in the database was selected, discussed and entered [Gader, 1992], [Ghenima, 1998], [Braham & Ghazali, 1998], [Dichy; Braham, Ghazali & Hassoun, 2002].

Further developments resulted in the coordination by J. Dichy (Université Lumière-Lyon 2) of the DIINAR-MBC Euro-Mediterranean project (E.C., INCO-DC program, project n° 961 791 - Feb. 1998-Dec.2000 - www.univ-lyon2.fr/langues/promodiinar/Accueil.htm). Main results are: a high level morphosyntactic analyser ([Ouersighni, 2001], based on [Ditters, 1992]), a set of software and multilingual lexica (Arabic, English and French), including procedures for the processing and indexation of Arabic corpora. A corpus of contemporary Arabic texts of around 10 million words has been compiled in Nimegen and Tunis. One later offshoot of the project was the Kalimât Arabic modern technical vocabulary [Guidère, ed., 2003].

DIINAR.1 has also been in use in the following related developments:

-A comprehensive analysis of Arabic conjugation [Dichy, 1993] resulted in the systematic tables and associated database of 10,000 Arabic verbs, each of which is related to its conjugation model. The work was published by Hatier in the famous Bescherelle series [Ammar & Dichy, 1999a and b].

- The object of Riadh Zaafrani's doctoral dissertation under supervision of the authors was the elaboration of software for the cognitive learning Arabic as a foreign language [Zaafrani, 2002]. The program drew 'intelligent' information from DIINAR.1. A crucial experiment with French students was directly concerned with the learning of Arabic lexical structures, in relation with the reading process.

- A concordance software drawing on DIINAR.1 has been elaborated by Ramzi Abbès, and presented in his doctoral dissertation under supervision of the authors [Abbès, 2004]. Frequency indexing of Arabic texts and the identification of word-forms have become, as a result, proficient enough to allow (a) compiling lexical frequency lists, and (b) the building of context-based dictionaries of Arabic.

- Jonathan Grainger, head of the Laboratoire de Psychologie Cognitive (Université de Provence/CNRS) conducted in 2003 a psycholinguistic experiment on the recognition of Arabic written words, in cooperation with researchers from Lyon (J. Dichy and R. Abbès), using word frequency variation extracted through DIINAR.1 and the above concordance software. Results showed evidence supporting the idea that 3-consonant Semitic roots play a relevant part in the recognition of Arabic words [Grainger et al., 2003].

- OPTAR ('Optique Arabe'), a database of 5,000 Arabic terms and phrases in the domain of optics, with corresponding terms in French and English, has been collected by Xavier Lelubre (Université Lyon 2). The idea of a hyperbase connexion associating OPTAR with DIINAR.1 has been presented in [Labed & Lelubre, 1997]. The work includes the extension of the concept and methodology of morphosyntactic specifiers to 'terminological specifiers' [Lelubre, 2001, 2002].

### *Understanding DIINAR.1: basic concepts and representations*

This last section recalls two basic concepts and representations shortly mentioned above. Information may not be new to readers familiar with the published work of the authors and other DIINAR.1 contributors, although we have endeavoured to be more explicit on a number of points.

#### A.The structure of the word-form in Arabic (a short recall)

Word-forms in Arabic can be described on the whole as consisting of a nucleus formative (NF) to which extension formatives (EF) are added, either to the left or to the right [Dichy, 1997].The NF, usually called stem, can be represented in terms of prosodic or non-concatenative morphology (after J. McCarthy's original and much discussed insights). In Semitic morphology, stems are considered, according to a somewhat recent, but very widely followed tradition, as a compound of root and pattern. (This view has been held in what can be described as over-powerful terms, and should be limited - see, among others, [Dichy, 2003], with many references).

Arabic word-forms ([Cohen, 1961], [Desclés, ed., 1983], [Dichy & Hassoun, eds. 1989]) consist of:

- proclitics (PCL), which include mono-consonantal conjunctions, i.e. wa-, 'and' , li-, 'in order to', or prepositions, i.e. bi-, 'in, at' or 'by', etc.;
- a prefix (PRF). The category, after [Cohen, 1961], only includes the prefixes of the imperfective, e.g., ya+, prefixed morpheme of the 3rd person;
- a stem, which can be represented in terms of:

(a) a ROOT (henceforth in small capitals), i.e., in the context of Semitic languages such as Arabic, Hebrew, Aramaic, etc., an ordered triple of consonants, or, by extension of the system, a quadruple, and

(b) a PATTERN (also in small capitals), i.e., roughly, a template of syllables, the consonants of which are the triple of the ROOT, to which vowels and mono-consonantal affixes are added. After [Cohen, 1961] and in partial accordance with traditional Arabic grammar, pre-ROOT, and some post-ROOT elements are conventionally included in PATTERNS (although conventions defining PATTERNS are slightly more intricate than indicated here). E.g.: the stem 'ista'mal, 'to make use of', consists of 3-consonant ROOT /'-m-l/ and of PATTERN /'istaR1R2aR3/, where R1, R2 and R3 respectively stand for 'radical consonant 1, 2, 3', and are instantiated by the triple of the ROOT (R1=', R2=m, R3=l). Consonants s and t belong to the PATTERN. In traditional Arabic grammar, /'istaR1R2aR3/ is represented as 'istaf'ala, using the 'meta-ROOT' /f-'-l/ to refer to the triple above (both representations can be considered equivalent, with a few conventional adjustments);

- suffixes (SUF), such as verbal desinences, nominal cases, the nominal feminine ending +a&, etc.;
- enclitics (ECL). In Arabic, enclitics are complement pronouns.

The table below gives two apparently equivalent representations of the structure of Arabic word-forms, although (2) emphasises relations between nucleus and extension formatives (NF and EF-s), featuring a triangle (ante- and post-positioned EF-s are abbreviated as aEF-s and pEF-s). The rules of the word formatives grammar are distributed along these three relations. Needless to say, a great number of them are related to the lexical nucleus, and have to rely on grammar-lexis relations. Both representations are valid. The first one revisits that presented in [Cohen, 1961], after [Desclés, ed., 1983]. The second reconsiders information included in (1) at a higher level of abstraction. Table II thus shows a representation operating at two levels of abstractedness. Level (2) emphasises nucleus-formatives relations, and retains by heritage analytic information appearing at level (1). In the next paragraph, we focus on paradigmatic nucleus ↔ nucleus derivational links, and on syntagmatic nucleus → extension formatives rules and relations.

| | |
|---|---|
| **(1)**<br>**Now traditional representation of the word-form** | maximal<br>_______ word-form_______<br>\|                                        \|<br>minimal<br>__word-form__<br>\|                    \|<br>##PCL # PRF +STEM+ SUF # ECL##<br>STEM = <ROOT**PATTERN>, in all verbs and deverbals, but in only a subset of nouns |
| **(2)**<br>**Nucleus-extensions representation (reconsidering (1) at a higher level)** | **NF**<br>/   \\<br>**aEF — pEF**<br>/   \\      /   \\<br>PCL  PRF  SUF  ECL |

TABLE II. The structure of the word-form in Arabic

## B. Morphosyntactic specifiers operating at word-form level (W-Specifiers)

W-Specifiers, which account for grammar-lexis rules and relations within the boundaries of the word-form, belong to two general types, one can describe, respectively, as 'paradigmatic' and 'syntagmatic'.

1) The paradigmatic type of grammar-lexis relations

The paradigmatic type of grammar-lexis relations is, in many Semitic languages, related to the analysis of stems in ROOTS and PATTERNS. In short words, some basic morphological derivation links (such as, in some nouns, singular ↔ 'broken' plural, or perfective ↔ imperfective alternation in simple verbs, etc.) are characterised by a paradigmatic change in PATTERN, the ROOT remaining constant([Dichy, 1984], [Hassoun, 1987], [Hassoun & Dichy, eds., 1989]). In Arabic, such basic links divide, on the whole, into (simplified presentation):

- noun ↔ noun, such as: singular ↔ 'internal' (or 'broken') plural (جموع التكسير);
- adjective ↔ adjective, in some forms, e.g. in the comparative or elative adjectives (أفعل التفضيل), , masculine ↔ feminine relations, and singular ↔ 'broken' plural alternation (as in nouns);
- verb-form ↔ verb-form links, such as perfective ↔ imperfective alternations (ماض ↔ مضارع) in 'simple' verbs (الفعل المجرّد);
- verb ↔ deverbal links, such as verb ↔ infinitive form (مصدر), or verb ↔ active participle (اسم الفاعل) relations, etc.

W-specifiers accounting for this type of relations are basic derivational link pointers.(in French: 'fléchage dérivationnel de base'; in Arabic: مؤشّرات العلاقات الاشتقاقية الأساسية).

Two points of general interest should be noted:

1) Derivational links are considered here restrictively. There are, in Arabic, many other derivational relations that can be deemed 'internal' to a given ROOT. Such links can be semantic and/or morphological. They concern, e.g., the relation between a 'simple' or 'augmented' verb and a given augmented verb form, or between a noun and a given denominative verb, etc. In spite of their descriptive impact, these links have

not been included in DIINAR.1. For the sake of consistency, only the above restricted subset of derivational relations has been taken into account, because they are directly related to the analysis or the generation of word-forms.
2) Grammar-lexis relations cope with phenomena that cannot be accounted for by grammar rules. In computational morphology, rules can only operate on the basis of identifiable formal markers. Semantic features that do not directly relate to a given morpheme included in the stem or the word-form must subsequently be added to entries as specifiers. We cannot, for lack of space, go through detailed explication of examples. Suffice it to say that derivation links included as specifiers are either not rule-predictable, or dependent on semantic features that cannot be inferred from the form of the stem [Dichy, 1997]. Such phenomena are by no means marginal: close examination of wide numbers of lexical data in Arabic have shown that - unlike what is still often heard or read - the ratio of non-predictable relations of the types above is very high.

2) The 'syntagmatic' or 'context-rule' type of grammar-lexis relations
The syntagmatic type is concerned with morphosyntactic and semantic features associated to the entries of a lexical database, and ensuring contextual relations. It can also be described as the 'context-rule' type of grammar-lexis relations. Let us recall a few examples, directly related to W-specifiers included in DIINAR.1:
a) The [± transitive] feature in verbs is well-known. We have found that corresponding W-specifiers can be formally reduced to three: intransitive (لازم), transitive to non human objects (متعدٍّ إلى غير العُقلاء), or transitive to both human and non human objects (متعدٍّ إلى العُقلاء وغيرهم). The reduction does not claim general semantic value. Transitivity concerns, within the boundaries of Arabic word-forms, enclitic complement pronouns, the structure of which can easily be shown to divide into these three categories.
b)A morphosyntactic rule forbids, in Arabic, co-occurrence, with a given verb, of a 1st person subject and a 1st person complement. The same goes with the 2nd person. This is due to the fact that the system of the language expresses reflexive meanings through other structures (use of nafs followed by an ECL pronoun, e.g. 'arâ nafsî fî l-mir'ât, 'I see myself in the mirror'; use of the mono-consonantal root of the 'echo-morpheme' [Roman, 1990], [Ammar & Dichy, 1999a]). This rule is suppressed with a small number of 'verbs of thought'
(أفعال القلوب), which refer to mental representation. One can say, e.g., 'alâ tarâka taqûlu..., 'don't you see that you say' (word-for-word: 'don't you see-you say...'). Such verbs therefore require an additional W-specifier.

3) Stems featuring both paradigmatic and syntagmatic relations
It is when both paradigmatic and syntagmatic relations or processes are involved that the set of specifiers used in DIINAR.1 shows the most interesting in mapping the lexicon of the Arabic language, on the basis of a finite set of features and rules. Three significant cases are illustrated below, starting from an easy example, and going on to more novel ones:
a) It is well-known that suffix +iyy, i.e. the relative noun or adjective morpheme (ياء النسبة) is likely to modify the noun it is associated with. E.g., madîn+a&, 'town', 'city' is modified into madan+ when combined with +iyy, thus: madan+iyy, 'city dweller', 'civilian' (as opposed to 'askar+iyy, 'military'), 'urbane'. This could have been considered as a rule-based process, were it not for examples where the rule does not apply, e.g. tabî'+a&, 'nature' tabî'+iyy, 'natural' (stable stem). Another example is the recently appeared word, madîn+iyy, 'urban' (differing from meanings associated with madan+iyy), in which the stem madîn has not been modified. A specifier thus has to relate the nominal stem to its modified form with suffix +iyy whenever applicable.
b) In nouns the 'external' (or suffixed) masculine plural form +ûna appears
(جمع مذكر سالم) in most cases when the noun is not associated in the lexicon with a 'broken plural' form (جمع تكسير), e.g. kâdib+ûna, 'liars'. (One must recall that the masculine plural suffix +ûna refers exclusively in Arabic to (male) humans, عُقلاء.)
In much more frequent cases, the 'broken plural' and the masculine 'external' suffixed form refer to different lexical entries. E.g., sâkin, supports either a 'broken' or a suffixed plural form, and subsequently divides into (simplified data):
- a first entry, which admits the suffixed masc. plur. sâkin+ûna, 'living [somewhere]', 'dwelling', and corresponds to the deverbal active (اسم الفاعل) participle of the verb sakana;
- a second, purely nominal entry sâkin, is associated, through a basic derivational link pointer, with the 'broken' plural form sukkân, 'inhabitants', and does not admit the suffixed masculine plural above. It is no longer a deverbal, but a noun built on the deverbal form just mentioned.

Such phenomena do not seem to have been explicitly and systematically studied in Arabic. In French, the above example of sâkin can be compared, among many other cases, to the active participle résidant, 'residing', 'dwelling', which supports no plural in French grammar, and the derived noun résident, 'resident', which supports the plural form [Dichy, 2003].

c) An essential issue, in the context of Arabic lexicography and/or the building of lexical resources, is that of the structures and types of lexical entries. It is crucial to note that lexical items (henceforth LI) and stems (or NF-s, nucleus formatives in Table II) do not always coincide. As shown in [Dichy, 1997], there are two types of LI-s:
- **simple lexical items**, which can be formalised as: **LI = <NF>.**

This is the case of all the verbs of the language, and of a subset of nouns and adjectives. Let us consider, e.g., the verb 'istahsana, 'to deem good, nice or beautiful'; the LI /'istahsan/ is equivalent to the NF or stem /'istahsan/ (which is liable in turn, to be represented as a non-concatenative compound of ROOT /h-s-n/ and PATTERN /'istaR1R2aR3/). Other example: zamân, 'time', in which LI /kalâm/ is equivalent to NF or stem /kalâm/ (ROOT: /k-l-m/, PATTERN: /R1aR2âR3 /);
- **morphologically compound lexical items,** formalised as: **LI = <NF + LEF>**, where LEF stands for lexicalized extension-formative. In a considerable number of cases, lexical items include one or more EF-s (extension formatives) that have been submitted to a lexicalization process (the NF/EF sequence is lexically 'frozen'). Let us examine a somewhat complex, but very significant example.

The word-form masrahiyya& can be analysed in two different ways:

(1) In the first analysis, it corresponds to the morphologically compound LI masrahiyya&, '(theatre) play', in which LI = <NF + LEF>, and suffixes:

- +iyy (the relative noun or adjective morpheme) and
- +a& (= 'the thing that', morpheme of the res generalis [Roman, 1990],

are LEF-s, i.e., are (a) associated in the lexicon with the NF or stem, and (b) integrative parts of the lexical entry. In addition, the two suffixes combine into the well-known compound suffix +iyya&, which is found in classical Arabic coinages such as su'ûb+iyya&, 'partisans of the su'ûb, 'peoples' (as opposed to the Arabs of the first few centuries of Islam); as well as in modern-time coinages, e.g. 'istirâkiyya, 'Socialism' (the compound +iyya& no longer refers in Modern Arabic to a group of people considered as a whole). In these examples, NF and LI do not coincide. In addition, the semantic value of the whole is not predictable by merely considering the combination between the elements involved: there is more in +iyya& than in the values of +iyy and +a&, and more in the morphological compound masrahiyya& than in the added meaning of its components [Dichy, 2003].

(2) An interpretation in which the meaning of the whole coincides, compositionally, with the combination of the meaning of its components does exist. The same word-form masrahiyya& also admits the analysis according to which masrah, 'theatre' corresponds to a simple LI (in which LI and Stem coincide), and is followed by suffixes +iyy (the relative noun or adjective morpheme) and +a& (which corresponds, here, to the nominal/adjectival morpheme of the feminine). The resulting word-form is the feminine of the relative adjective 'theatrical', masrah+iyy+a& (masc., masrah+iyy).

Comparison between the compositional meaning of (2) and the non-compositional meaning of (1) shows that there is an added semantic element in the latter.

## *Future prospects: morpho-semantic and syntactico-semantic specifiers*

The inventory of the finite set of W-specifiers in Arabic is a crucial element of the word formatives grammar (WFG), on the basis of which the DIINAR.1 lexical resource was built.

The next step should be the inventory of S-specifiers, which should in turn play a central part in the analysis and generation of sentences. This includes going further down the line of formalising grammar-lexis relations, taking into account the categories and functions of the computational syntax of Arabic [Ditters, 1992], and retaining, in the very wide range of possible semantic features, those which remain within the scope of morphological and syntactic rules and relations [Dichy, 2005]. This, with the help of concordance and corpus processing software drawing on DIINAR.1 [Abbès, 2004], is the next challenge.

## *References:*

[1] Ramzi Abbès. 2004. La conception et la réalisation d'un concordancier électronique pour l'arabe. Thèse de doctorat en sciences de l'information, Lyon, ENSSIB/INSA.

[2] Sam Ammar & Joseph Dichy. 1999a. Les verbes arabes, Paris, Hatier (collection Bescherelle). Monolingual Arabic edition, 1999b. الأفعال العربية (Al-'af'â al-'arabiyya). Same publisher (original Arabic introduction).

[3] Richard Bouché, Joseph Dichy & Mohamed Hassoun. 1984. "Enseignement Assisté par Ordinateur de l'arabe. Simulation à l'aide d'un modèle linguistique, la morphologie". Colloque "E.A.O. 84", (Lyon, 4-5 septembre 1984), Paris, Agence de l'informatique, 1984: 81-96, revisited version, Dichy & Hassoun, eds, 1989: 42-62.

[4] Abdelfattah Braham & Salem Ghazali. 1998.-32
حصيلة وآفاق – المجلة العربية للعلوم – ع DIINAR
أو مشروع معجم العربية الآلي (معالي
قاعدة البيانات المعجمية العربية

[5] David Cohen. 1961. "Essai d'une analyse automatique de l'arabe", T.A. informations 1961, reprocuded in D. Cohen, Etudes de linguistique sémitique et arabe, Paris, Mouton, 1970: 49-78.

[6] Jean-Pierre Desclés, dir. 1983. By: H. Abaab, J.-P. Desclés, J. Dichy, D.E. Kouloughli, M.S. Ziadah. Conception d'un synthétiseur et d'un analyseur morphologiques de l'arabe, en vue d'une utilisation en Enseignement assisté par Ordinateur, Rapport rédigé pour le Ministère des Affaires étrangères, 1983.

[7] Joseph Dichy. 1984. "Vers un modèle d'analyse automatique du mot graphique non-vocalisé en arabe", 1984, in Dichy & Hassoun, eds, 1989: 92-158.

[8] - 1987. "The SAMIA Research Program, Year Four, Progress and Prospects". Processing Arabic Report 2, T.C.M.O., Nijmegen University:1-26.

[9] - 1990. L'Écriture dans la représentation de la langue : la lettre et le mot en arabe. Thèse d'État (en linguistique), Université Lumière-Lyon 2.

[10] - 1993. "Knowledge-system simulation and the computer-aided learning of Arabic verb-form synthesis and analysis". Processing Arabic Report 6/7, T.C.M.O., Nijmegen University: 67-84, 92-95.

[11] - 1997. "Pour une lexicomatique de l'arabe : l'unité lexicale simple et l'inventaire fini des spécificateurs du domaine du mot". Meta 42, Presses de l'Université de Montréal, Québec, spring 1997: 291-306. www.erudit.org/revue/meta/1997/v42/n2/002564ar.pdf

[12] - 2000. "Morphosyntactic Specifiers to be associated to Arabic Lexical Entries - Methodological and Theoretical Aspects". ACIDA' 2000 (Monastir, Tunisia, 22-24.03.2000), Corpora and Natural Language Processing vol.: 55-60. www.elsnet.org/arabic2001/dichy.pdf

[13] - 2001. "On lemmatization in Arabic. A formal definition of the Arabic entries of multilingual lexical databases". ACL 39th Annual Meeting. Workshop on Arabic Language Processing; Status and Prospect, Toulouse: 23-30. www.elsnet.org/arabic2001/dichy.pdf

[14] - 2003. "Sens des schèmes et sens des racines en arabe : le principe de figement lexical (PFL) et ses effets sur le lexique d'une langue sémitique", in Sylvianne Rémi-Giraud et Louis Panier, dir., La polysémie ou l'empire des sens. Lyon : Presses Universitaires de Lyon (coll. " Linguistique et sémiologie ") : 189-211.

[15] - 2005. "Spécificateurs engendrés par les traits [±animé], [±humain], [±concret] et structures d'arguments en arabe et en français", in Henri Béjoint & François Maniez, eds., De la mesure dans les termes, Colloque en hommage à Philippe Thoiron, université Lumière Lyon 2, 23-25 septembre 2004, Presses Universitaires de Lyon, p. 151-181.

[16] Joseph Dichy, Abdelfattah Braham, Salem Ghazali, M. Hassoun. 2002. "La base de connaissances linguistiques DIINAR.1". In: Abdelfattah Braham. Colloque international sur le traitement automatique de l'arabe, 18-20 avril 2002, La Manouba-Tunis: Université de La Manouba: 45-56.

[17] Joseph Dichy & Mohamed Hassoun, eds. 1989. Simulation de modèles linguistiques et Enseignement Assisté par Ordinateur de l'arabe - Travaux SAMIA I. Paris, Conseil International de la Langue Française.

[18] - 1998. "Some aspects of the DIINAR-MBC research programme". In A. Ubaydly, ed., 1998. Proceedings of the 6th International Conference and Exhibition on Multilingual Computing (ICEMCO 98), Centre of Middle Eastern Studies, University of Cambridge: 2.8.1-6.

[19] Everhard Ditters. 1992. A Formal Approach to Arabic Syntax : The Noun phrase and the Verb Phrase, PhD, Catholic University of Nijmegen.

[20] Nabil Gader. 1992. Conception et réalisation d'un prototype de correcteur orthographique de l'arabe. Mémoire de DEA en Sciences de l'information et de la communication, ENSSIB.

[21] Malek Ghenima. 1998. Analyse morpho-syntaxique en vue de la voyellation assistée par ordinateur des textes écrits en arabe. PhD., ENSSIB/Université Lyon 2.

[22] Jonathan Grainger, Joseph Dichy, Mohamed El-Halfaoui, Mohamed Bamhamed. 2003. "Approche expérimentale de la reconnaissance du mot écrit en arabe", in Jean-Pierre Jaffré, éd., Dynamiques de l'écriture : approches pluridisciplinaires, revue Faits de langue, n°22 : 77-86.

[23] Mathieu Guidère, ed. 2003. Marie-Hélène Avril, Salam Bazzi-Hamzé, Amal El Sabbane, Lynne Franjié, Mathieu Guidère, Xavier Lelubre, Rita Moucannas-Mazen, Hoda Moucannas-Mehio, Manar Rouchdy, Camilia Soubhi, Kalimât .- Le vocabulaire arabe. Ellipses, Paris.

[24] Mohamed Hassoun. 1987. Conception d'un dictionnaire pour le traitement automatique de l'arabe dans différents contextes d'application. Doctorat d'Etat, Université Lyon 1.

[25] Lamia Labed & Xavier Lelubre. 1997. "DIINAR-TOPT: conception d'une base de données terminologique Arabe/français dans le domaine de l'optique", in JST'97: L'ingénierie de la langue: de la recherche au produit, Avignon, 15-16/04/1997, AUPELF-UREF/FRANCIL: 523-8.

[26] Xavier Lelubre. 2001. "A Scientific Arabic Terms Data Base: Linguistic Approach for a Representation of Lexical and Terminological Features". In ACL 39th Annual Meeting. Workshop on Arabic Language Processing; Status and Prospect, Toulouse: 66-72.

[27] Riadh Ouersighni. 2001. "A major offshoot of the DIINAR-MBC project: AraParse, a morpho-syntactic analyzer of unvowelled Arabic texts". In ACL 39th Annual Meeting. Workshop on Arabic Language Processing: Status and Prospect, Toulouse, pp. 66-72. www.elsnet.org/arabic2001/ouersighni.pd

[28] André Roman. 1990. Grammaire de l'arabe, Paris : P.U.F. (coll. "Que sais-je ?").

[29] Riadh Zaafrani. 2002. Développement d'un environnement interactif d'apprentissage avec ordinateur de l'arabe langue étrangère. PhD, ENSSIB/Université Lyon 2.

Joseph Dichy, Professor of Arabic Linguistics; Université Lumière-Lyon 2, Faculty of Languages and ICAR, "Interactions, Corpus, Apprentissages, Représentations" (UMR 5191, CNRS/Lyon 2), work-group: SILAT, "Systèmes d'information, Ingénierie, Linguistique de l'Arabe et Terminologie" - Lyon, France - joseph.dichy@univ-lyon2.fr

Mohamed Hassoun, Professor of Information Sciences, ENSSIB, "École Nationale Supérieure des Sciences de l'Information et des Bibliothèques", and work-group: SILAT, "Systèmes d'information, Ingénierie, Linguistique de l'Arabe et Terminologie" - Villeurbanne, France - hassoun@enssib.fr

# New Resources

## ELRA-S0175 Mandarin Chinese Speecon database

The Mandarin Chinese Speecon database is divided into 2 sets:

1.The first set comprises 26 DVDs with the recordings of 550 adult Chinese speakers (276 males, 274 females), recorded over 4 microphone channels in 4 recording environments (office, entertainment, car, public place).

2.The second set comprises 3 DVDs with the recordings of 50 child Chinese speakers (26 boys, 24 girls), recorded over 4 microphone channels in 1 recording environment (children room). The database has been collected and is owned by Nokia Research Center (Nokia Group). The database was validated by SPEX, the Netherlands, to assess their compliance with the Speecon format and content specifications.

| | ELRA members | Non-members |
|---|---|---|
| For research use | 50,000 Euro | 60,000 Euro |
| For commercial use6 | 67,000 Euro | 75,000 Euro |

## ELRA-S0176 Finnish Speecon database

The Finnish Speecon database is divided into 2 sets:

1.The first set comprises 22 DVDs with the recordings of 550 adult Finnish speakers (273 males, 277 females), recorded over 4 microphone channels in 4 recording environments (office, entertainment, car, public place).

2.The second set comprises 3 DVDs with the recordings of 50 child Finnish speakers (25 boys, 25 girls), recorded over 4 microphone channels in 1 recording environment (children room).

The database has been collected and is owned by Nokia Research Center (Nokia Group). The databases was validated by SPEX, the Netherlands, to assess their compliance with the Speecon format and content specifications.

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

## ELRA-S0177 Korean Speecon database

The Korean Speecon database is divided into 2 sets:

1.The first set comprises 30 DVDs with the recordings of 568 adult Korean speakers (259 males, 309 females), recorded over 4 microphone channels in 4 recording environments (office, entertainment, car, public place).

2.The second set comprises 4 DVDs with the recordings of 58 child Korean speakers (25 boys, 33 girls), recorded over 4 microphone channels in 1 recording environment (children room). The database has been collected by Appen Pty Ltd and is owned by Toshiba Research Europe Ltd. The databases was validated by SPEX, the Netherlands, to assess their compliance with the Speecon format and content specifications.

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

## ELRA- S0178 Turkish Speecon database

The Turkish Speecon database is divided into 2 sets:

1.The first set comprises 28 DVDs with the recordings of 550 adult Turkish speakers (280 males, 270 females), recorded over 4 microphone channels in 4 recording environments (office, entertainment, car, public place).

2.The second set comprises t4 DVDs with he recordings of 50 child Turkish speakers (25 boys, 25 girls), recorded over 4 microphone channels in 1 recording environment (children room). The database has been collected by Appen Pty Ltd and is owned by Toshiba Research Europe Ltd. The database was validated by SPEX, the Netherlands, to assess their compliance with the Speecon format and content specifications.

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

## ELRA-S0179 Polish Speecon database

The Polish Speecon database is divided into 2 sets:

1.The first set comprises 26 DVDs with the recordings of 550 adult Polish speakers (286 males, 264 females), recorded over 4 microphone channels in 4 recording environments (office, entertainment, car, public place).

2.The second set comprises 3 DVDs with the recordings of 50 child Polish speakers (25 boys, 25 girls), recorded over 4 microphone channels in 1 recording environment (children room). The database has been collected by the Polish Japanese Institute of Information Technology, Poland and is owned by Sony International (Europe) GmbH. The database was validated by SPEX, the Netherlands, to assess their compliance with the Speecon format and content specifications.

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

## ELRA-S0180 Portuguese Speecon database

The Portuguese Speecon database is divided into 2 sets:

1.The first set comprises 29 DVDs with the recordings of 553 adult Portuguese speakers (266 males, 287 females), recorded over 4 microphone channels in 4 recording environments (office, entertainment, car, public place).

2.The second set comprises 4 DVDs with the recordings of 52 child Portuguese speakers (19 boys, 33 girls), recorded over 4 microphone channels in 1 recording environment (children room). The database has been collected by Appen Pty Ltd and is owned by Sony International (Europe) GmbH.The database was validated by SPEX, the Netherlands, to assess their compliance with the Speecon format

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

## ELRA-L0056: STO SprogTeknologisk Ordbase (Danish Lexicon for NLP/HLT Applications)

The STO Lexicon is the most comprehensive computational lexicon of Danish comprising approx. 81,530 entry words, and it is well integrated with the European activities in the field of lexicon development building on experience obtained from the PAROLE and SIMPLE projects. The model and descriptive method of the STO lexicon are kept compatible with the architecture and descriptive language of PAROLE/SIMPLE. A number of refinements, adaptations and language-specific extensions to the basic model are implemented in STO

This lexicon is well suited for NLP/HLT monolingual applications, as lexicon component in taggers, parsers, grammar & spell checkers, summarisation tools, web crawlers, computer-aided language learning, as well as multilingual applications; also possibility for linking to other PAROLE/SIMPLE-compatible resources.

| | ELRA members | Non-members |
|---|---|---|
| For research use by academic organisations | 2,000 Euro | 2,500 Euro |
| For research use by commercial organisations | 5,000 Euro | 6,250 Euro |
| For commercial use | 21,000 Euro | 26, 250 Euro |

## ELRA-M0041 Bulgarian WordNet

The Bulgarian WordNet was developed by the Department for Computational Linguistics at the Institute for Bulgarian Language, Bulgarian Academy of Sciences, initially within the framework of the BalkaNet project "Multilingual Semantic Network for the Balkan Languages" (IST-2000-29388) and later on under the scope of the BulNet project, funded at the national level. For more information about the BalkaNet project: http://www.ceid.upatras.gr/Balkanet/, and about the Department for Computational Linguistics: http://dcl.bas.bg.

The Bulgarian WordNet models nouns, verbs, adjectives, and (occasionally) adverbs. It contains 23,715 word senses (synsets), 51,011 literals, 1,863 domain specific synsets, 41,620 lexico-semantic relations, 197 extralinguistic relations.

| | ELRA members | Non-members |
|---|---|---|
| For research use by academic organisations | 237.15 Euro | 474.30 Euro |
| For research use by commercial organisations | 3,557.25 Euro | 7,114.50 Euro |
| For commercial use | 5,928.75 Euro | 11,857.50 Euro |

## *ELRA Membership Fidelity Program*

ELRA is pleased to initiate a fidelity program to reward its loyal members. The principle of the fidelity program is to earn miles by joining and remaining member of our association.

*Miles, what for?*

The awarded miles can be used by members, once earned, for:

- The payment of membership fees
- The payment of registration fees to LREC and other events organized by the association
- The purchase of the Language Resources from the ELRA catalogue with additional discount.

*How many miles?*

This depends on the type of the institution, and therefore on the membership fee. The number of miles per year that can be earned is currently as follows:

| | |
|---|---|
| - Not-for-profit organization: | 200 miles |
| - European small/medium-sized companies (< 50 employees): | 250 miles |
| - European profit making organizations (>= 50 employees): | 375 miles |
| - Non-European profit making organizations: | 1250 miles |

*Rules*

The use and earning of miles is subject to the following rules:

- If membership fee is paid before July 1st, the member gets the annual number of miles immediately.
- If membership fee is paid after July 1st, the member is entitled to keep the miles acquired so far but the member will not earn miles for the current year (all other member benefits still apply, e.g. reduced price on resources, on LREC registration fees, etc.). The rule does not apply to new members who can join anytime and earn miles for that 1st year.
- If membership fee has not been paid by December 31st, all the miles acquired so far are lost.

Miles can be used as soon as they are earned. Exceptionally, for 2005, the 1st July deadline will not apply.

*Examples*

After 2 years (2005 and 2006), a not-for-profit institution will have earned 400 miles (200 each year) and will be able to register 3 students for LREC2006.

After 4 years, any institution member of ELRA will have earned enough miles to get a free membership for one year.

# HLT EVALUATION WORKSHOP

## 1 & 2 DECEMBER 2005

## MALTA

To celebrate ELRA's 10th anniversary, a 2-day workshop dedicated to Human Language Technologies (HLT) Evaluation will be held in Malta on December 1st and 2nd 2005.

In organizing this workshop, ELRA intends to bring together the HLT Evaluation key players to discuss the HLT evaluation from various perspectives: general principles and purposes, technologies, past and on-going evaluation projects, worldwide initiatives, etc. All sectors of HLT will be addressed

**For more information on venue and program, please visit the Events page on www.elra.info**

# LREC 2006

## 22-28 MAY 2006,

## GENOA - ITALY

**IMPORTANT DATES**

Submission of proposals for panels, workshops and tutorials:20 October 2005

Submission of proposals for oral and poster papers: 20 October 2005

Notification of acceptance of panels, workshops and tutorials proposals: 7 November 2005

Notification of acceptance of oral papers, posters:16 January 2006

Final versions for the proceedings: 20 February 2006

**For more information, please visit www.lrec-conf.org/lrec2006**