

# The ELRA Newsletter



October - December  
2006

*Vol.11 n.4*

## *Contents*

<i>Letter from the President and the CEO</i> .....	<i>Page 2</i>
<i>EAFT 10<sup>th</sup> anniversary - Achievements and Objectives</i> <i>Annelise Grinsted</i> .....	<i>Page 3</i>
<i>Introducing The Union Latine</i> <i>Daniel Prado</i> .....	<i>Page 4</i>
<i>ECESS, a HLT Network on Speech Synthesis</i> <i>Harald Höge, Zdravko Kacic, Imre Kiss</i> .....	<i>Page 6</i>
<i>Forty Years: Machine Translation Evaluation Today and Yesterday</i> <i>Margaret King</i> .....	<i>Page 8</i>
<i>New Resources</i> .....	<i>Page 11</i>

*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:***  
Victoria Arranz  
Valérie Mapelli  
Hélène Mazo

***Layout:***  
Martine Chollet  
Valérie Mapelli

***Contributors:***  
Annelise Grinsted  
Harald Höge  
Zdravko Kacic  
Margaret King  
Imre Kiss  
Daniel Prado

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](mailto:choukri@elda.org)  
Web sites:  
<http://www.elra.info> or  
<http://www.elda.org>

## *Dear Colleagues,*

This is the last issue of 2006 and we would like to highlight a number of topics on which ELRA focussed during this year.

The major event for ELRA has been the organisation of LREC 2006, the fifth edition in the series of the Language Resources and Evaluation Conference launched by ELRA with the support of a large number of active players in the field. Organised in Genoa (Italy), LREC 2006 was a very successful event, which attracted more than 800 participants with over 500 papers and 18 workshops and tutorials. LREC is also the opportunity to honour the memory of our founding President Antonio Zampolli through the Zampolli Prize. The 2006 Antonio Zampolli Prize was awarded to Christiane Fellbaum and George Miller for their outstanding contribution to NLP through the work done on WordNet.

The organisation of LREC 2008 is in its preparatory phase. The LREC Programme Committee and the Board of ELRA decided to investigate the possibility to hold LREC 2008 in Marrakech (Morocco). More information will be provided in the next issue.

ELRA has secured a number of new resources for distribution through new partnerships (e.g., Speech Resources from Beijing Haitian Ruisheng Science Technology Ltd) and through French and European Projects (e.g. NEOLOGOS, EURADIC, CHIL Evaluation Packages). All resources are announced in the last section of this newsletter and consist of:

- S0226-01 IDIOLOGOS "Bootstrap" (NEOLOGOS Project)
- S0226-02 IDIOLOGOS "Eigenspeakers" (NEOLOGOS Project)
- S0227 PAIDIALOGOS (NEOLOGOS Project)
- S0228-01 to S0228-57 and W0045-01 to W0045-08 Speech Resources from Beijing Haitian Ruisheng Science Technology Ltd
- L0049 SCIPER-FR-EURADIC French Monolingual Dictionary
- L0050 SCIPER-AN-EURADIC English Monolingual Dictionary
- L0051 SCIPER-AL-EURADIC German Monolingual Dictionary
- L0052 SCIPER-ES-EURADIC Spanish Monolingual Dictionary
- L0053 SCIPER-IT-EURADIC Italian Monolingual Dictionary
- L0073 DIINAR.1 - Arabic Lexical Resource
- E0009 CHIL 2004 Evaluation Package
- E0010 CHIL 2005 Evaluation Package
- E0011 TC-STAR 2006 Evaluation Package - ASR English
- E0012 TC-STAR 2006 Evaluation Package - ASR Spanish
- E0013 TC-STAR 2006 Evaluation Package - ASR Mandarin Chinese
- E0014 TC-STAR 2006 Evaluation Package - SLT English-to-Spanish
- E0015 TC-STAR 2006 Evaluation Package - SLT Spanish-to-English
- E0016 TC-STAR 2006 Evaluation Package - SLT Chinese-to-English
- M0033 SCI-FRAN-EURADIC French-English Bilingual Dictionary
- M0034 SCI-FRAL-EURADIC French-German Bilingual Dictionary
- M0035 SCI-FRES-EURADIC French-Spanish Bilingual Dictionary
- M0036 SCI-FRIT-EURADIC French-Italian Bilingual Dictionary

During this quarter, ELRA and ELDA have continued to play an active part in a number of European and international projects, in particular in the preparation of the evaluation campaigns due to start early 2007 for both TC-STAR and CHIL projects, and also the speech databases collection in the LILA project.

As for this newsletter, it contains 3 articles on Language Technologies-related associations and a paper on evaluation:

- In "EAFT 10th Anniversary - Achievements and Perspectives", Annelise Grinstead gives an overview of what has been accomplished in the European Association for Terminology over the past 10 years.
- Daniel Prado presents the Union Latine's activities with a focus on the Dtil Division in charge of Terminology and Language Industries, in "Introducing the Union Latine".
- ECESS is a consortium dedicated to Speech Synthesis whose current and future activities are described in "ECESS, a HLT Network on Speech Synthesis".
- In "Forty Years on: Machine Translation Evaluation Today and Yesterday", Maghi King gives an overview of the latest developments in the field of MT evaluation.

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

Bente Maegaard, President

Khalid Choukri, CEO

## EAFI 10th anniversary - Achievements and Objectives

Annelise Grinsted

Ten years after its foundation, the EAFI (European Association for Terminology) organised the Third Terminology Summit in Brussels (Belgium) on the 13th and 14th of November 2006. On this occasion, Ms Annelise Grinsted gave a speech where she elaborated on the fundamental developments achieved during her presidency and the direction that should be taken by the association. First president of the EAFI in 1996-1997, Ms Grinsted has been a Member of the Advisory Council of the European Association for Terminology since 1998 until now.

### EAFI 10th ANNIVERSARY

#### Speech given on 14th November 2006

I would first and foremost like to thank the Board for inviting me to address two questions at this 10th anniversary for the European Association for Terminology:

- What were the fundamental developments / accomplishments during my presidency?

and

- In which direction should the EAFI proceed?

As an individual - president or not - I did not accomplish anything, but I participated as one in a diverse group of people to establish what we felt was needed at that moment. And we took the POINTER Final Report from 1995 as point of departure in which was proposed "a broad-based professional membership organization for facilitating future activities, an anchor for special interest groups on specific problems and topics, and a vehicle for promoting the profession and awareness of it".

POINTER which stands for Proposals for an Operational Infrastructure for Terminology in Europe was a project co-funded by the European Commission, DG XIII-E as part of its Multilingual Action Plan (MLAP) to create a set of concrete recommendations for activities leading to a coordinated but flexible terminology infrastructure for Europe.

The First General Assembly of the future European Association for Terminology was held on 3rd of October

1996 at Southern Denmark Business School. And it was not an easy birth.

I have gone back into my archives from that period and will quote some comments forwarded to me before the assembly which reflect some of the difficulties at the time. The comments are of the most diverse nature. I have, of course, made anonymous the comments as the identity of the persons or organisations are irrelevant today:

1. There has been a lot discussion as to who can and should become members of this new association. There are two schools of thought:

a. EAFI should become an umbrella organisation as a sort of European International Federation of Translators and only associations should be able to join.

b. An association which would accept natural persons as members.

We are all agreed on the fact that a need exists for coordination at European level. The grass-roots activities, however, are conducted by the national associations on the spot.

After a careful and extensive reflection WE have come to the conclusion that membership of EAFI will only be possible if EAFI restricts its scope to that of a coordinating umbrella organisation.

2. What is the association's philosophy about diversity? Should it really promote it (multilingualism, market sensitivity of the vocabulary, etc.) or should it refrain from it as a source of misunderstanding (by developing standardization of terminographical descriptions, of vocabulary, of concept systems, etc.). It is important to mention that there are different approaches, that terminologists don't all agree on what has to be done and so that the association will be a place to deal with this kind of opposite trends.

3. It should be mentioned that the general assemblies will be held in different countries each time.

4. (translated from French and a reply to a letter I - as organizer - had written in English)

I have received your letter which has caught my attention.

I call your attention to the fact that we cannot participate in an association which will not give equal importance to the official languages of the European Union and gives a *de facto* monopoly to English: where is, I ask, the European character of such an organisation? And that, despite the fact that it is much easier for a terminology organisation to practice multilingualism than other organisations. I therefore ask that you give me all your assurances in regard to your project's European multilingual character.

5.(translated from French)

I hope that - even though we "southerners" have not been well informed about the existence of the organisation until now - there will also be room for us in the organisation.

Although many stakeholders in the field of terminology had given input to the POINTER report, there was apparently not TOTAL agreement on how to implement the recommendations as reflected in the comments above. So, the first period of time was spent on trying to establish where the largest amount of stakeholders could agree and gain - and where neither linguistic nor national borders were barriers. One important work area was the creation of the Special Interest Groups devoted to specific subject fields and issues. Another area was determining where synergies and mutual gains could be created. Many interesting initiatives had been and were underway that could profit mutually by cooperating:

- The European Language Resources Association (ELRA), established as a non-profit organisation in February, 1995.
- Infoterm and European Network of Terminology Information and Documentation Centres (TDCnet) and the establishment of the European Terminology Information Server (ETIS)
- The Association for Terminology and Knowledge Transfer founded in 1986 and gaining in weight and importance in the 90s.
- The national or regional associations like e.g.

- NORDTERM, an "old" network (1976) and later association (1987).

- Realiter established in June of 1996.

10 years later I can bring up these examples that reflect individual positions and collective concerns, because my conclusion about what was obtained in 1996 is that it was a fundamental accomplishment that a European organization was established DESPITE all the big and small objections and concerns.

A reflection of this is the fact that the past days' EAFT Summit is an integrated part of a terminology week in cooperation with TERMnet and NL Term in which terminology work is reflected in its many aspects. Furthermore, the amount of members, especially from the new European Union member countries confirms the necessity of an organization of this nature.

So today we meet as an organisation that has had 10 good, sometimes stormy, sometimes calm years. And it is always wise to

stop up and reflect - based on the objectives stipulated in the Statutes:

- to further plurilingualism in Europe through terminology;
- to provide a platform at the European level for the promotion and co-ordination of terminological activities and the heightened awareness, improved recognition and continued professionalization of the terminology sector;
- The EAFT has obtained many results on all three objectives. Terminology work is recognized as essential in an increased number of sectors, public and private. And through the educational systems and other entities terminology work has become a profession reflected both in research and practice. We have heard about the advances during the summit.

Organisations have learned to cooperate and work across interests and borders - some stakeholders stay out and the majority is in.

So where do we go from here? In my opinion it is important to continue to consolidate the advances and at the same time look for the synergies with other related areas - not to exclusively look on ourselves as terminologists. We are linguists with a special knowledge that have to deal with a complex world.

Annelise Grinsted  
University of Southern Denmark  
Institute of Business Communication  
and Information Science  
annelise@sitkom.sdu.dk

## Introducing the Union Latine

Daniel Prado

The Union Latine, an intergovernmental organisation, has 37 Member States whose official language is a Neolatin language<sup>1</sup>. It aims at promoting and developing the Neolatin languages, and ensures cultural dissemination from the Latin countries.

In 1984, at the time when the Union Latine was relaunched, Philippe Rossillon, who contributed considerably to tool development for the French language, focussed on the necessary and harmonious development of the other Latin languages, pointing out a decline in the scientific and technical terminologies/vocabularies in Neolatin languages, as well as a poor offer in specialised information for these languages. Further to this statement of fact, the Union Latine developed a "Program of computational terminology and linguistics", which

<sup>1</sup> Andorra, Angola, Argentina\*, Bolivia, Brazil, Cape Verde, Chile, Colombia, Costa Rica, Ivory Coast, Cuba, Ecuador, Spain, France, Guatemala, Guinea-Bissau, Haiti, Honduras, Italy, Mexico, Monaco, Mozambique, Nicaragua, Panama, Paraguay, Peru, Philippines, Portugal, Dominican Republic, Republic of Moldavia, Romania, San Marino, The Holy See\*, São Tomé and Príncipe, Senegal, East Timor, Uruguay, Venezuela, Order of Malta\* (\*permanent observer).

became the Dtil Division, standing for Direction of Terminology and Language Industries, one of the three programmatic divisions of the Organisation. The two other divisions are dedicated to the cultural dissemination (<http://dcc.unilat.org>) and promotion of the Neolatin languages (<http://dpel.unilat.org>).

The Dtil Division has a double mission: encourage the terminology community from the Latin countries to enhance the dissemination of their work and inform on the language engineering innovations, in order to boost the development/expansion of techniques that enable efficient and fast translation of international documentation, from English in particular - but also from any other language - towards the Latin languages.

No one is unaware that, from the second half of the 20th century on, the most widely spoken languages in the world have experienced a significant decline in the field of communication and scientific and technical information, to the benefit of English. Surprisingly enough, English is the mother tongue of less than 10% of the humanity and many studies

and surveys show that only very few people of those who claim to speak English do actually master the language in a way that allows them to express themselves fluently, to negotiate and to be understood in this language.

This situation leads our society to an alternative: make every citizen on Earth bilingual (English/mother tongue) - which is utopian in addition to being detrimental to linguistic diversity - or allow everyone to access correct translation of the technical and scientific literature.

This is why the Union Latine makes it a point of honour to support the activities of the Latin countries in the fields of terminological enrichment and linguistic automation (machine translation, aided-translation, etc).

The following information gives an overall picture of the activities undertaken by the Union Latine in the sector of the specialised communication, since going through all activities in one article would be long and tedious. For more information, we invite you to visit our website: <http://dtil.unilat.org>.

Since its creation, the Dtil Division has organised, co-organised or sponsored

many meetings which have brought together applied and computational linguistics specialists and specialised communication experts. As early as 1984, the Union Latine organised the first International Exhibition of computational terminology and linguistics in Lisbon. Various institutions, public and private, took part in this event, among which the persons in charge of term banks such as Eurodicautom and those supported by the governments of Canada and Quebec. In 1987, the Union Latine renewed the operation in Madrid, which triggered the creation of the first network of terminology for Spanish and Portuguese languages (Riterm - Latin American Network of Terminology, <http://www.riterm.net>).

After that, many steps were taken to support the terminology communities: permanent support or assistance to the creation of several national associations of terminology (Cuban, Spanish, Italian, Mexican, Moldavian, Portuguese, Peruvian, Rumanian, Venezuelan, etc); permanent support of international associations (European Association of Terminology - AET, of whom the Union Latine is a founding member); and especially, creation and permanent support of the Panlatin Terminology Network - Realiter - (<http://www.realiter.net>) which gathers about sixty personalities or institutions of Latin countries active in terminology and carries out, among others, multilingual terminological work in societal fields.

It has certainly been necessary to train thousands of terminologists and specialised translators to give a serious boost to the terminology activity in the Latin countries. To achieve this goal, the Union Latine organised or supported hundreds of seminars, conferences, meetings of all types, during the past 22 years. The Organisation started or coordinated the major international meetings, the most notorious being, in Europe, the "Conference on the co-operation in the field of terminology", the "Conference for a terminology infrastructure in Europe" and the first two "World Summits of Terminology", organised in the name of the European Association of Terminology.

Several databanks (terminology, neology or lexicography) and tens of specialised glossaries, lexica, dictionaries have been built up with the support of the Union

Latine which was also involved in several international projects on terminology (Pointer, TDCNet, Riterm-BD, Terminesp), on minority languages (Linmiter), on access to scientific and technical information (Redalc) or on neology (Antenas neológicas).

Little mention of the actions related to multilingualism will be made, because this goes beyond the scope of this article. Nevertheless, for several years now, the Union Latine has maintained a high-level commitment in favour of linguistic diversity. Most of the organisations or networks supported by the Union Latine in the field of terminology or applied linguistics are now up and running, so the Organisation can focus its activity towards other sectors requiring an equally voluntarist action.

In this way, the Union Latine takes major actions to strengthen the presence of Neolatin languages and support the non-official languages (whether of neolatin origin or not) in Latin States through studies, investigations or inventories in various fields of knowledge, such as the presence of Latin languages on the Web, a study carried out since 1998 ([http://dti1.unilat.org/LI/2005/index\\_fr.htm](http://dti1.unilat.org/LI/2005/index_fr.htm)). Lately, UNESCO requested the Union Latine to carry out two studies: one on multilingualism in the cyberspace (to be released) and the other on the multilingualism specialists and specialised institutions in the cyberspace throughout the world (in progress).

It is important to highlight the common action led within the "Three linguistic spaces" (<http://www.3el.org>) to reassert the value of the Spanish, Portuguese and French languages. This structure gathers the large international Organisations dedicated to these languages, such as the *Organisation internationale de la Francophonie*, *Organización de Estados Iberoamericanos*, *Secretaría General Iberoamericana*, *Comunidade dos Países de língua Portuguesa* and the Union Latine.

A major part is being played by the Union Latine to support languages other than the official languages of the Organisation (Catalan, Spanish,

French, Italian, Portuguese and Rumanian). For example, it supports a network of minority Latin languages (Linmiter) gathering lower-diffusion languages such as Galician, Occitan, Friulan, Ladin or Corsican.

Promoting specialised vocabularies in Amerindian languages (Quechua, Aymara and Guaraní) mainly through the creation of three lexica containing nearly 50,000 terms in the fields of health, food processing industry and biodiversity is another achievement of the Union Latine. More actions are in progress concerning Central America languages, the Haitian and Cape Verdean Creoles and some African languages.

As far as language industries are concerned, it should be mentioned that the Union Latine, as an intergovernmental Organisation, has no authority whatsoever to develop applications on its own or to work out vocabularies. Motivate, justify, encourage, support (despite the weakness of its financial means), design solutions and advise Latin States' authorities in order for their language to benefit from the latest technological developments, are the Union's daily activities.

The Union Latine thus acts by disseminating widely news and developments concerning research and applications likely to help the translator or the writer to obtain specialised documents, of good quality, quickly and at a low cost. For instance, in the past, the Union Latine translated reports that the Latin community should know of and supported the writing or the publishing of scientific documents, or information dissemination on various tools. The Union is also in charge of disseminating any information on these developments to the specialists in the Latin countries, through its former Terminometro bulletin, which has now become a daily information web site on terminology, language industries, scientific and technical translation and associated disciplines. Terminometro is available in five Latin languages and can be visited at <http://www.terminometro.info>.

In addition, information on language industries is disseminated and can be discussed on the Termilat list (<http://www.termilat.info>), which is a forum attended by the linguists of the Latin countries, by the

chatters of the SIIT (<http://www.siit.info>) list, intended for the translators of Latin America and the Iberian Peninsula, and finally by those of the Consortium list, constituted by Rumanian language specialists. A specific page dedicated to the dissemination of any useful information on the applications (and research) in the field of computer-assisted translation (CAT) has been created. Information on machine translation, translation memories, terminology management systems, among others, can be found there.

The Union Latine has trained specialists to the use of various translation-aided tools and terminology and neology management and organised several workshops and demonstrations of products and projects.

It has allowed the development of the first prototype of machine translation system for Rumanian (allowing the translation in other Latin languages and English as both source and target languages), in cooperation with the Atamiri laboratory, which had already developed similar products for other languages (<http://www.atamiri.cc/es/index.html>).

On the basis of this project, the Union Latine currently undertakes a pilot experiment aiming at applying the Atamiri project to various multilingual mailing lists. It has funded the development of a lexico-

graphy and terminology management software (free license) for the minority Neolatin languages and is currently financing a specific application for terminology management in Rumanian, on the basis of a free license software as well. It has provided its technical and financial support to the constitution of several terminology databanks, in particular in Brazil, in Italy, in Romania, in Moldavia or within the framework of Mercosur.

Finally, technology watch and political advice to the Member States are the main focus of the Union Latine which is very active in organising international meetings where authorities and institutions are invited to discuss these subjects. The last international meeting is the "Third Inter-American Language Management" which took place in Rio de Janeiro in June 2006 ([http://dtil.unilat.org/tercer\\_seminario/index\\_fr.htm](http://dtil.unilat.org/tercer_seminario/index_fr.htm)). We can also quote the second edition of this seminar, during which workshops on machine translation were carried out ([http://dtil.unilat.org/segundo\\_seminario/index.htm](http://dtil.unilat.org/segundo_seminario/index.htm)) or the "Conference on the presence of Latin languages in specialised communication". Further to these seminars or meetings, as well as those mentioned at the beginning of

this article, in which specialists mix with political personalities, conclusions or recommendations focusing on the development of language technologies are always broadly disseminated.

Other sites of interest, related to the activities of the Union Latine:

- **Portalingua**, web site dedicated to the linguistic aspects of the languages spoken in the Latin States: <http://www.portalingua.info>
- **Latinosapiens**, web site dedicated to the scientific articles written in Neolatin languages: <http://www.latinosapiens.info>
- **Documentarium**, web site dedicated to the scientific and technical documentation of the Latin countries: <http://www.documentarium.info>
- **Linmiter**, site of the Network of terminology for the minority Neolatin languages: <http://www.linmiter.net>

Daniel Prado  
 Direction de la Terminologie et des  
 Industries de la Langue - DTIL  
 Union Latine  
 131 rue du bac  
 75007 Paris  
 +33 (0)1 45 49 60 62  
[dtil@unilat.org](mailto:dtil@unilat.org)  
<http://dtil.unilat.org>

## ***ECESS, a HLT Network on Speech Synthesis***

*Harald Höge, Zdravko Kacic, Imre Kiss*

### *Introduction*

ECESS (European Center of Excellence on Speech Synthesis) is an open, non-funded consortium for institutions working on speech synthesis and related topics. ECESS was founded in February 2004 on Harald Höge's initiative. Currently ECESS counts 13 active members located in Europe, China and Japan. ECESS aims at building an infrastructure whose goal is to speed up progress in speech synthesis with respect to models, algorithms, and languages. The idea behind the infrastructure is to facilitate the exchange of modules, language resources, and tools needed for speech synthesis and related topics. The exchange is based on the principles of validation and evaluation, which will be described later.

ECESS is organized into 3 colleges:

- College "Modules and Systems" (Coordinator: Harald Höge, Siemens AG)
  - College "Language Resources" (Coordinator: Imre Kiss, Nokia)
  - College "Tools" (Coordinator: Zdravko Kacic, University of Maribor)
- Currently the consortium is in the process of establishing a consortium agreement, which will regulate the duties and rights of the institutional members. The consortium meets twice a year to discuss both organizational and scientific issues.

The following sections describe the working principles, and the current and future activities of the three colleges.

### *College Modules and Systems*

The partners of ECESS have defined the architecture of a speech synthesis system consisting of three modules:

- Text Processing Module (text normalization, tokenization, POS-tagging, transcription).
- Prosody Generation Module (F0 and duration prediction).
- Acoustic Synthesis Module (unit selection, concatenation, prosodic manipulation).

The APIs between the modules are defined by XML-formatted parameters.

Most partners of ECESS work on these modules covering the following languages : UK, CN, SP, JP, SI, DE, PL, EU, FI, TR, GA. A module is license-free for research



use (binary code), if it meets the thresholds of established evaluation criteria. The following evaluation criteria are used:

- Text processing:
  - Text normalization: word error rate.
  - End of sentence (EoS) detection: precision/recall.
  - POS: POS error rate.
  - G2P: Phoneme/word error rate on different domains.
- Prosody: MOS values.
- Acoustic synthesis: Intelligibility, MOS on quality.

The setting of thresholds for evaluation criteria is defined via LSP (Language Specific Peculiarities) provided by the partner, who delivers the module. The thresholds have to be accepted by the college. The evaluation of modules has been performed during the second and third evaluation campaigns organized within TC-STAR for Mandarin Chinese, British English and European Spanish. The setting of thresholds and exchange of modules is one of the next actions within ECESS.

### *College Language Resources*

This college is responsible for coordinating the exchange of language resources (LR) for speech synthesis. Three types of LR correspond to the three pools for exchange: annotated acoustic databases, pronunciation lexica, and text databases for training automatic POS taggers. LR collection is carried out by each partner individually or through subcontracting. After validation by an external validation centre and acceptance by the college participants and project consortium, LRs are offered for exchange.

The main specifications of the LR cover:

- Acoustic databases: the specifications were developed within the TC-STAR project. TC-STAR TTS databases contain about 10 hours of annotated speech by a selection of professional speakers. For more details, refer to the public deliverable D8 available on <http://www.tc-star.org>.
- Pronunciation lexica: the specifications were developed according to the LC-STAR specifications. LC-STAR lexica contain 100,000 entries distributed as fol-

lows: 50K common words, 45K proper names and 5K special application words. For more details, refer to the public deliverables D1.1-2.4 available on <http://www.lc-star.com>.

- The specifications for POS text corpora are not yet finalized. The minimum size of a corpus is expected to be 100,000 tokens on TC-STAR text domains (in line with acoustic data creation). The use of LC-STAR or comparable tag sets is proposed, and 100% of the POS tags has to be manually checked. As for the previous two types of LR, the creation of a LSP document is mandatory. The validation and evaluation details are currently being worked out.

Validated acoustic resources are already available in the college for exchange for UK, CN and SP languages with 6 more languages under preparation. Some commitments for pronunciation lexica exchange have also been made and discussions are ongoing about POS text databases.

### *College Tools*

This college is responsible for coordinating the development of signal and text processing tools needed for the development of TTS systems. The tools will be arranged into tool pools. Two types of tool pools will be created:

- text processing tool pool,
- signal processing tool pool.

The development of tools will be stimulated by evaluation campaigns. On a regular basis, the college will issue calls for expression of interest in order to initiate evaluation campaigns for selected tools. At least three participants have to express their interest for the campaign to be carried out. The college will provide development and evaluation data for all participants of the evaluation campaign. Tool-specific reference database peculiarities will be defined for the evaluation campaign when needed. In case of language-specific tool evaluation, the tool-specific reference database peculiarities can be proposed by the partner cooperating in the evaluation campaign and have to be approved by the college prior to the campaign. Prior to

each evaluation campaign, the college will issue specifications for the tools to be evaluated and will define the evaluation thresholds. The exchange rights for the tools to be exchanged are scheduled by evaluation campaigns. Each partner is eligible to exchange tools which have been accepted in a particular evaluation campaign where these tools were accepted a priori in the tool pool.

The first evaluation campaign of signal processing tools was carried out in 2006. Pitch marking (PMA) and pitch detection (PDA) tools were evaluated. The following partners were involved in the evaluation:

- Pitch marking (PMA) tool: University of the Basque Country (UBC), Technical university of Dresden (TUD), University of Maribor (UMB), Technical University of Catalonia (UPC), Siemens AG.
- Pitch detection (PDA) tool: University of the Basque Country (UBC), Czech Technical University in Prague (CTU), University of Maribor (UMB).

For further information on the activities of the ECESS network, please visit our web site: <http://www.ECESS.eu>

Harald Höge  
Siemens AG, Corporate Technology IC5,  
OttoHahn Ring 6, 81739 Munich  
Germany  
[harald.hoege@siemens.com](mailto:harald.hoege@siemens.com)

Zdravko Kacic  
University of Maribor, Faculty of  
Electrical Engineering and Computer  
Science  
Smetanova17, 2000 Maribor  
Slovenia  
[kacic@uni-mb.si](mailto:kacic@uni-mb.si)

Imre Kiss  
Nokia Research Center  
PO Box 1000, 33721 Tampere  
Finland  
[imre.kiss@nokia.com](mailto:imre.kiss@nokia.com)

## *Forty Years: Machine Translation Evaluation Today and Yesterday*

Margaret King

### *Introduction*

It is just over forty years since the ALPAC report (1) was published. The mandate of the committee responsible for the report was to consider whether further investment in machine translation research and development was justified, given what had been achieved with the funding already made available. The conclusions of the report were negative, contentious and far-reaching. An unavoidable consequence was to push questions of evaluation methodology into prominence: in fact, there is a strong sense in which one might claim that the ALPAC report created a new discipline of evaluation in natural language processing.

The new discipline gained further impetus with the organisation by DARPA/ARPA of a series of evaluation campaigns, where the basic underlying hypothesis was that regular and systematic evaluation of research systems as they were being developed would contribute to the identification of the most promising approaches, encourage the growth and cohesion of communities of scientists working in the same domain, and thereby serve to further the advancement of core technologies which could then be put to use in numerous practical applications. The campaigns covered (and cover) a wide range of systems in the general domain of human language technology - indeed machine translation itself was a rather late candidate technology in the evaluation campaign paradigm, perhaps somewhat ironically because the conclusions of the ALPAC report had put a brake on work in the area. However, the DARPA/ARPA campaigns themselves triggered a great deal of discussion about how evaluation should be done, much of it pertinent to any application in the human language technology area, and thereby reinforced the importance of the infant discipline. In Europe, formal acknowledgement of the new discipline came with the creation of a working group on evaluation as part of the EU EAGLES initiative (1991) and later with the first LREC conference in 1998 - the first major international conference where evaluation appears in the title of the conference.

Economic forces determined that work on machine translation would continue in the post-ALPAC years, especially with the growth of political communities who adopted multilingualism as a matter of principle and were forced to face up to the practical and economic problems of trying to put their principles into practice (Canada and the European Union are primary examples). From the mid-90s on, the creation and explosive growth of the World Wide Web, coupled with political and practical globalisation, made the creation and successful deployment of machine translation a burning issue. In a context where management and exploitation of information across languages is a key economic and political issue, machine translation has an obvious and almost inescapable role to play. In consequence, machine translation evaluation in its turn became a growth industry, whose outward and visible sign is the number of papers dedicated to the topic in conferences and in journals.

In 2006 I was invited to chair one of the two sessions devoted to machine translation and evaluation in the fifth meeting of LREC. There were seven papers in the session I chaired, which covered a wide range of topics, most of which were also represented in the session I did not chair. Listening to these papers, I found myself reflecting on whether the questions raised by the ALPAC evaluation have found satisfactory answers. In what follows, I shall use the content of the papers in "my" session in order to try to set out why, although my answer to the question is rather depressingly negative, I nonetheless feel that evaluation as a discipline has made great strides.

### *Metrics, metrics and more metrics.*

Many of the major criticisms levelled at the ALPAC evaluation concerned the metrics used. These were based on more or less traditional categories used to discuss translations produced by humans: are the translations faithful to the original,

intelligible to the reader and transparent in the sense that they give the impression of having been written in the target language. First, notice here the implicit assumption that the appropriate way to assess machine translation output is to compare it to translations produced by humans: we shall return to this point later. That basic assumption however was not challenged. The definition of the metrics though gave rise to considerable debate. For each metric, human judges were asked to score candidate translations on a scale, where definitions of points on the scale were given by brief descriptions in English. The definitions of the points were challenged, the number of points on the different scales was challenged, the directionality of the scales was challenged - but most of all the fallibility, or rather the variability, of human judges was challenged. Humans are not robots: they will have different interpretations of the same set of instructions, they will be more jaundiced in their judgements on Mondays than on Tuesdays, they will be influenced more or less subtly by emotional, psychological and physiological factors. In other words, they are not reliable.

One possible way to compensate for the unreliability of humans would be to have a lot of them, and base final judgement on some sort of averaging of their scores. But here we run into another major problem: appropriate humans are scarce and expensive. They have to be paid for their time, and the more of them there are, the more it costs.

Hence, of course, the search for some sort of automated metric, which would be objective in the sense of being independent of human judgement and cheap to administer because, once the initial programming has been done, computers do not cost a lot to run. And of course computerized metrics have the additional advantages of being able to accomplish very large amounts of work very quickly and of being always available.

A number of automated metrics for machine translation have been proposed, but have yet to meet with any very widespread agreement. One problem is that they rely on comparison,



one way or another, with one or more reference translations - translations produced by humans which are held to be the standard by which the machine translation output is to be judged. Whatever is chosen as reference material defines *ipso facto* what the translation ought to be. Given the notorious fact that, except for very short and uncontentious stretches of text no such thing as a correct translation can be said to exist, coupled with the problem in many cases of finding appropriate reference material, this tends to provoke unease. Other problems relate to how the candidate machine translation is compared to the reference translation(s). Another way of stating this latter concern is to ask what is really being measured, and whether what is being measured is in fact an indicator of "good" translation - whatever that might be. Thus, in practice, the claims for validity for most automated metrics are based on the degree to which they correlate with human judgements. The sceptical might be pardoned for thinking that thus we come full circle. But of course this is not necessarily true: if a set of human judgements could be produced and validated - probably at considerable expense - that set could then be used as a yard-stick against which the results of applying other metrics could be judged: the idea of assessing relative to a gold standard which has permeated much work on evaluation methodologies would thereby be extended to the validation of metrics themselves. (There are other problems however, many of them still to do with expense and practicality, which we shall not have time to discuss here).

Two of the papers in the LREC session were directly concerned with these issues.

The first of these, Hamon and Rajman, **X-Score: Automatic Evaluation of Machine Translation Grammaticality**, presented the results of an experiment designed to test the validity of a metric based on measuring the grammaticality of machine translation output. The underlying assumption here is that realistically, machine translation systems will not produce output that could be mistaken for human output: the closer to acceptability the system comes, the less likely it is to produce ungrammatical output. The paper describes the metric itself, the experimental method and the different tests used over the data of CESTA, the French machine translation eval-

uation campaign. The main advantage of the metric is that it is inexpensive, since it requires no reference translation(s). The authors argue that the results are promising, and that they correlate well with human judgements - in other words that grammaticality is in fact a predictor of translation quality, at least in the case of machine translations.

Hamon et al, **CESTA: First Conclusions of the Technolugue MT Evaluation Campaign**, is, as its title implies, also concerned with activities within the CESTA campaign. The emphasis in this paper is firmly on the relation between automated metrics and those relying on human judgement. The campaign used five automated metrics, three of them well known and two experimental. Human metrics based on assessing fluency and adequacy were also used, thus providing data for comparison between the results of applying automated and human metrics.

Results of a first round of evaluations are analysed in the paper. A comparison of the rankings of the different systems evaluated produced by each of the metrics is presented and discussed. A second round of evaluation concentrated on texts in the medical domain, in order to observe the impact of domain adaptation. Although detailed results were not available at the time of the conference, the presentation reported on interesting differences in the correlation between specific automated results and human judgement between the two rounds of evaluation.

#### *Changing the assumptions*

The two papers briefly described above both rely on a notion of intrinsic quality in machine translation output that is independent of the use to be made of the machine translation, a hypothesis that is very closely related to ALPAC's assumption that machine translation would, in the ideal case, approximate as closely as possible to human translation. Two further papers illustrated a shift away from this towards a notion of what early EAGLES work (2) called adequacy evaluation: is the output such that it facilitates the accomplishment of a specific task or set of tasks?

Miller and Vanni, **Register-Differentiated Arabic MT Evaluation in the PLATO Paradigm**, reported on an attempt to develop a way of deciding what system might be used for what task. PLATO is based on the observation that different tasks place different requirements on the overall quality required of machine translation output, and sets about exploring the possibility of a predictive relationship between discrete, well-defined metrics and the tasks that can be reliably performed. Scores on PLATO assessments constitute a signature to be correlated with different tasks and with automated metrics. The human based metrics used assess clarity, coherence, morphology, syntax, lexical robustness, named-entity rendering and adequacy.

In the experiment presented in the paper, register was used as a criterion for distinguishing input to five Arabic-English machine translation systems. Modern Standard Arabic (MSA) was distinguished from non-standard Arabic text such as that found in electronic discourse. Analysis of the assessment scores revealed specific areas, such as general lexical robustness, where system performance is comparable on both types of input. Divergent performance occurred on assessments of clarity, name rendering and domain terms. These results suggest that, while systems may be considered reliable regardless of the input register for the lexicon-dependent triage task, register may have an effect on the suitability of the machine translation output for relevance judgement and information extraction tasks, which rely on clarity and proper rendering of named entities.

Macklovitch, **TransType2: The Last Word**, presented the results of an on-site, context-oriented evaluation of the TransType2 system. As a translator is entering the translation, TransType tries to "guess" what will come next, offering the translator one or more completions which are compatible with what the translator has typed up until now. The system was evaluated *in situ* in two separate translation agencies, in five rounds of user trials where senior translators of the two organisations carried out the assessment. The main focus of the evaluation was on whether the translators could increase their productivity by using TransType2, although the asses-

sors were also actively encouraged to record their own reactions whilst using the system.

The results show significant gains in productivity, but the users' comments highlight the fact that in real life use other usability factors can be just as important as gains in productivity. For example, the system was unable to remember what corrections the translator had made to a proposed continuation. Thus, if the text was repetitive, the translator would be called upon to make the same corrections each time the source segment recurred - thus producing a very negative reaction to use of the system despite significant gains in productivity. The author emphasizes the importance of usability factors when machine translation systems are to be deployed.

### *Resources for evaluation*

The last two sections pick up on issues which were fundamental to discussion of the ALPAC report: the question of what constitutes a valid metric and the question of whether machine translation can be evaluated independently of the use to which it is to be put. The next two papers pick up on issues that were not explicitly part of the discussion, but which underlie many decisions about evaluation design, including the design of the experiments reported in ALPAC.

Most metrics, no matter whether the evaluation aims at assessing some sort of intrinsic quality or a system's ability to facilitate a given task, rely on the availability of appropriate data. Producing this data is always a major task, and even more so when what is in question is an evaluation campaign, where the nature and the quality of the data will determine in quite a large measure the acceptability to the participants and sponsors of the results of the evaluation. Indeed, a common claim used to promote evaluation campaigns is that they produce an invaluable resource in the form of re-usable data, just as a common criticism is that the existence of the same data tends to bias future work, just because it is so much cheaper to make use of data that exists than to produce new data.

Two papers reported on particular campaigns and the data used there.

Mostefa, Hamon and Choukri, **Evaluation of Automatic Speech Recognition and Speech Language Translation within T-STAR**, reported on the first evaluation cam-

paigned of what is intended as a long-term effort to advance research in the core technologies of speech-to-speech translation. The campaign took place at the end of the first year of the project.

The first evaluation essentially separated evaluation of automatic speech recognition components from the evaluation of the translation components. Evaluation of the translation components used different kinds of inputs: the output of the automatic speech recognition, verbatim transcriptions of the spoken text and syntactically and semantically correct text input. Perhaps not surprisingly, the translation components found the text input the easiest to deal with.

As a side product, the evaluation also produced six evaluation packages, three for speech recognition and three for translation, which are publicly available.

Strassel et al, **Integrated Linguistic Resources for Language Exploitation Technologies**, reported on the efforts of the Linguistic Data Consortium in the context of the DARPA GALE programme. The goal of the GALE programme is to develop and apply computer software technologies to absorb, analyze and interpret huge volumes of speech and text in multiple languages. Three major language technology engines are involved - transcription, translation and distillation. During development of the technology required for each of these three engines, resources must be created to support evaluation of progress as it is made. These resources include data, annotations of the data, tools, standards and best practices. The presentation described how the various resources will be produced, and how they will be distributed.

The authors emphasize that although the resources are developed within the context of the GALE programme, the LDC Consortium intends to distribute the data more broadly whenever possible. Indeed they hope that their efforts will lead to substantial corpora with durable value to the worldwide Human Language Technology community and to the technology users who benefit from HLT development.

### *Evaluation as a daily development tool.*

The seventh paper was very different from all the others and really falls outside the scope of any comparison with the state of evaluation technology at the time of the ALPAC report, except in the sense that any proposal akin to the one described in this paper would have been quite simply unthinkable in the ALPAC context, simply because of the limitations of computing at the time. Indeed, the paper is based on a common practice in software development outside human language technology, and even in that domain would not have been feasible in the mid 1960s.

Schäfer and Beck, **Automatic Testing and Evaluation of Multilingual Language Technology Resources and Components**, presented an automated evaluation platform used for daily testing of a set of system components under development. The modular based system being developed provides a set of basic components, for example tokenizers or domain specific gazetteers, and can deal with a number of different languages. The components and resources for each module are worked on independently and are frequently modified. A variety of different projects make use of the basic components, for example projects in automatic hyperlinking, opinion mining, question answering and text mining for weather forecasts (but not, so far as I can see, machine translation!).

The evaluation platform first builds a system out of selected (versions of) individual modules by compiling it from the sources defined by a source control system. The "new" system is then used on a known corpus and the results compared automatically to a predefined gold standard. A quick report simply indicates the presence or absence of errors. A more detailed report gives a diagrammatic representation of precision, recall and f-measure results, together with a historical picture of how these scores have developed over time.

### *Conclusion.*

We started by asking whether forty years of work on machine translation evaluation has produced a methodology which is largely accepted and widely practiced, thereby providing a sort of *de facto* standard by which machine translation systems can be judged. The answer has to be no. We are still in a state

where some evaluations take it for granted that there is such a thing as an intrinsic quality of machine translation output which can be judged independently, whilst others base the evaluation design on the tasks to be accomplished with the help of the machine translation. The problem of metrics based on human judgement is still very much with us: a number of interesting automated metrics have been produced, but their validity is still questioned, and, as we have noticed, that validity is most often justified by claims to correlation with human judgements. Most specific evaluations are still (and for very good reasons which we have not gone into here) producing their own data and linguistic resources.

But it would be wrong to conclude from this that we have made no progress. It is

now common practice for scientific papers routinely to include a section on evaluation: we do feel that we have to evaluate and that we have to justify how we choose to do so. We are concerned with the choice, definition and justification of appropriate metrics. Once it would have seemed enough to say that we were measuring intelligibility or fidelity, now we feel obliged to say why and how. It may be that the nature of translation is such that we can never find the magic metric that will solve all our problems, but our understanding of the issues is immeasurably greater than it was forty years ago, and we are becoming ever more ingenious in finding ways around the fundamental problems.

### References

- 1-ALPAC 1966: Languages and Machines: Computers in Translation and Linguistics. Report of the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council Publication 1416, Washington D.C., 1966.
- 2- EAGLES 1996: EAGLES Evaluation Working Group, EAGLES Evaluation of Natural Language Processing Systems, Final Report. Center for Sprogteknologi, Copenhagen, EAG-EWG-PR-2. ISBN 97-90708-00-8, 1996

Margaret King  
University of Geneva, Switzerland  
maghi.king@gmail.com

## NEW RESOURCES

### ELRA-S0226-01 IDIOLOGOS 1 “Bootstrap” (NEOLOGOS Project)

The IDIOLOGOS 1 “Bootstrap” database was produced within the French national project NEOLOGOS, as part of the Technolangu programme funded by the French Ministry of Research and New Technologies (MRNT). The databases produced in the framework of the NEOLOGOS project are designed for the development and the assessment of French speech or speaker recognizers and speech synthesizers. They consist in:

- 1) the IDIOLOGOS databases are made of adults voices and are available in 2 subsets:
  - the “Bootstrap” database (catalogue ref. ELRA-S0226-01),
  - the “Eingenspeakers” database (catalogue ref. ELRA-S0226-02);
- 2) the PAIDIALOGOS database (catalogue ref. ELRA-S0227) is made of children’s and teenagers’ voices.

The IDIOLOGOS 1 “Bootstrap” database contains the recordings of 1,000 adult French speakers (470 males and 530 females) recorded over the French fixed telephone network. The speakers uttered 45 phonetically rich sentences. The 45 sentences are the same for all speakers.

This database is distributed as 1 DVD-ROM. The speech files are stored as sequences of 8-bit, 8kHz A-law speech files and are not compressed, according to the specifications of NEOLOGOS. Each prompt utterance is stored within a separate file and has an accompanying ASCII SAM label file. This speech database was validated by SPEX (the Netherlands) to assess its compliance with the NEOLOGOS format and content specifications. Each speaker uttered the following items: 1 digit sequence (5+ digits), 1 telephone number (10 digits), 1 credit card number (16 digits), 1 spelling of directory assistance city name, 1 real/artificial for coverage, 45 phonetically rich sentences. The following age distribution has been obtained: 288 speakers are between 18 and 30, 264 speakers are between 31 and 45, 247 speakers are between 46 and 61, and 201 speakers are over 61.

	ELRA members	Non-members
For research use	1,000 Euro	2,000 Euro
For commercial use	10,000 Euro	16,000 Euro

### ELRA-S0226-02 IDIOLOGOS 2 “Eingenspeakers” (NEOLOGOS Project)

The IDIOLOGOS 2 “Eingenspeakers” database was produced within the French national project NEOLOGOS, as part of the Technolangu programme funded by the French Ministry of Research and New Technologies (MRNT). The databases produced in the framework of the NEOLOGOS project are designed for the development and the assessment of French speech or speaker recognizers and speech synthesizers. They consist in:

- 1) the IDIOLOGOS databases are made of adults voices and are available in 2 subsets:
  - the “Bootstrap” database (catalogue ref. ELRA-S0226-01),
  - the “Eingenspeakers” database (catalogue ref. ELRA-S0226-02)
- 2) the PAIDIALOGOS database (catalogue ref. ELRA-S0227) is made of children’s and teenagers’ voices.

The IDIOLOGOS 2 “Eingenspeakers” database contains the recordings of 200 adult French speakers (97 males and 103 females) recorded over the French fixed telephone network. The speakers uttered 45 sentences per call with 10 calls per speaker. The 450 sentences per speaker are common to all speakers. Speakers were selected from the IDIOLOGOS 1 “Bootstrap” (ELRA-S0226-01) database.

This database is distributed as 1 DVD-ROM. The speech files are stored as sequences of 8-bit, 8kHz A-law speech files and are not compressed, according to the specifications of NEOLOGOS. Each prompt utterance is stored within a separate file and has an accompanying ASCII SAM label file.

This speech database was validated by SPEX (the Netherlands) to assess its compliance with the NEOLOGOS format and content specifications.

Each speaker uttered the following items: 1 digit sequence (6 digits), 1 telephone number (10 digits), 1 credit card number (16 digits), 1 spelling of directory assistance city name, 1 real/artificial for coverage, 45 phonetically rich sentences.

The following age distribution has been obtained: 42 speakers are between 18 and 30, 50 speakers are between 31 and 45, 62 speakers are between 46 and 61, and 46 speakers are over 61.

	ELRA members	Non-members
For research use	1,000 Euro	2,000 Euro
For commercial use	15,000 Euro	24,000 Euro

### ELRA-S0227 PAIDIALOGOS (NEOLOGOS Project)

The PAIDIALOGOS database was produced within the French national project NEOLOGOS, as part of the Technolanguage programme funded by the French Ministry of Research and New Technologies (MRNT). The databases produced in the framework of the NEOLOGOS project are designed for the development and the assessment of French speech or speaker recognizers and speech synthesizers. They consist in:

1) the IDIOLOGOS databases are made of adults voices and are available in 2 subsets:

- the "Bootstrap" database (catalogue ref. ELRA-S0226-01),
- the "Eingenspeakers" database (catalogue ref. ELRA-S0226-02);

2) the PAIDIALOGOS database (catalogue ref. ELRA-S0227) is made of children's and teenagers' voices.

The PAIDIALOGOS database contains 37,364 utterances from 1,010 child French speakers (510 males and 500 females) recorded over the French fixed telephone network.

This database is distributed as 1 DVD-ROM. The speech files are stored as sequences of 8-bit, 8kHz A-law speech files and are not compressed, according to the specifications of NEOLOGOS. Each prompt utterance is stored within a separate file and has an accompanying ASCII SAM label file.

This speech database was validated by SPEX (the Netherlands) to assess its compliance with the NEOLOGOS format and content specifications.

Each speaker uttered the following items: 3 application words (set of 42); 4 connected digits: 2 sequence of 3 isolated digits, 1 sheet number (7 digits), 1 telephone number (10 digits); 3 dates (1 spontaneous date e.g. birthday, 1 word style prompted date, 1 relative and general date expression); 2 isolated digits; 3 spelled words (1 surname, 1 directory assistance city name, 1 real/artificial name for coverage); 1 currency money amount; 1 natural number; 4 directory assistance names (1 spontaneous, e.g. own surname, 1 city of where the call is made from, 1 most frequent French city out of a set of 40, 1 "forename surname"); 2 yes/no questions (1 predominantly "yes" question, 1 predominantly "no" question); 6 phonetically rich sentences; 2 time phrases (1 time of call, 1 word style time phrase); 2 phonetically rich words.

The following age distribution has been obtained: 6 speakers are under 7, 541 speakers are between 7 and 11, 308 speakers are between 12 and 14, 154 speakers are between 15 and 16, and 1 speaker is over 16.

	ELRA members	Non-members
For research use	2,000 Euro	4,000 Euro
For commercial use	14,000 Euro	23,000 Euro

### Speech Resources from Beijing Haitian Ruisheng Science Technology Ltd.

In the previous ELRA Newsletter (Vol.11 n.2-3), ELRA announced the signature of a major Language Resources distribution agreement with Beijing Haitian Ruisheng Science Technology Ltd. On behalf of ELRA, ELDA incorporated to the ELRA Language Resources catalogue over 60 new Speech resources designed and collected to boost Speech Synthesis and Speech Recognition. The resources cover mainly Mandarin Chinese with some coverage of Korean and Japanese languages. They consist of:

- Mandarin Chinese Speech Synthesis Corpus (from ELRA-S0228-01 to S0228-03),
- Chinese Telephone Speech Recognition Corpus (from ELRA-S0228-04 to S0228-11, and from S0228-26 to S0228-30),
- Chinese Desktop Speech Recognition Corpus (from ELRA-S0228-12 to S0228-25, and from S0228-31 to S0228-32),
- Mandarin Chinese Speech Recognition Corpus (desktop, in-car, telephone) (from ELRA-S0228-33 to S0228-45),
- Mandarin Chinese high clarity Speech Recognition Corpus (in recording studio) (from ELRA-S0228-46 to S0228-49),
- Korean Mandarin Speech Recognition Corpus (desktop) (from ELRA-S0228-50 to S0228-53),
- Japanese Mandarin Speech Recognition Corpus (desktop) (from ELRA-S0228-54 to S0228-57),
- Original Short-Message Data Collation in Chinese (from ELRA-W0045-01 to W0045-08).

For prices, visit the ELRA catalogue online: <http://catalogue.elra.info>

### **ELRA-L0049 SCIPER-FR-EURADIC French Monolingual Dictionary**

This French monolingual dictionary was increased and improved within the French national project EurRADic (European and Arabic Dictionaries and Corpora), as part of the Technolanguage programme funded by the French Ministry of Industry. It contains 112,216 lemmas (694,673 inflected forms), with their part of speech and some information related to their inflexion. The data are presented in a table format, where information related to each entry is separated by ";". Other formats and other services may be supplied by the data owner upon request (e.g. conversion into buyer's formalism, selection of subsets of the words missing from your own dictionary).

A description of the project is available at the following address: [http://www.technolanguage.net/article.php?id\\_article=203](http://www.technolanguage.net/article.php?id_article=203) (in French). See also ELRA-L0050, ELRA-L0051, ELRA-L0052, ELRA-L0053, ELRA-M0033, ELRA-M0034, ELRA-M0035, ELRA-M0036.

---

### **ELRA-L0050 SCIPER-AN-EURADIC English Monolingual Dictionary**

This English monolingual dictionary was increased and improved within the French national project EurRADic (European and Arabic Dictionaries and Corpora), as part of the Technolanguage programme funded by the French Ministry of Industry. It contains 171,713 lemmas (365,823 inflected forms), with their part of speech and some information related to their inflexion. The data are presented in a table format, where information related to each entry is separated by ";". Other formats and other services may be supplied by the data owner upon request (e.g. conversion into buyer's formalism, selection of subsets of the words missing from your own dictionary).

A description of the project is available at the following address: [http://www.technolanguage.net/article.php?id\\_article=203](http://www.technolanguage.net/article.php?id_article=203) (in French). See also ELRA-L0049, ELRA-L0051, ELRA-L0052, ELRA-L0053, ELRA-M0033, ELRA-M0034, ELRA-M0035, ELRA-M0036.

---

### **ELRA-L0051 SCIPER-AL-EURADIC German Monolingual Dictionary**

This German monolingual dictionary was developed within the French national project EurRADic (European and Arabic Dictionaries and Corpora), as part of the Technolanguage programme funded by the French Ministry of Industry. It contains 157,810 lemmas (17,634,834 inflected forms), with their part of speech and some information related to their inflexion. The data are presented in a table format, where information related to each entry is separated by ";". Other formats and other services may be supplied by the data owner upon request (e.g. conversion into buyer's formalism, selection of subsets of the words missing from your own dictionary).

A description of the project is available at the following address: [http://www.technolanguage.net/article.php?id\\_article=203](http://www.technolanguage.net/article.php?id_article=203) (in French). See also ELRA-L0049, ELRA-L0050, ELRA-L0052, ELRA-L0053, ELRA-M0033, ELRA-M0034, ELRA-M0035, ELRA-M0036.

---

### **ELRA- L0052 SCIPER-ES-EURADIC Spanish Monolingual Dictionary**

This Spanish monolingual dictionary was increased and improved within the French national project EurRADic (European and Arabic Dictionaries and Corpora), as part of the Technolanguage programme funded by the French Ministry of Industry. It contains 83,952 lemmas (838,391 inflected forms), with their part of speech and some information related to their inflexion. The data are presented in a table format, where information related to each entry is separated by ";". Other formats and other services may be supplied by the data owner upon request (e.g. conversion into buyer's formalism, selection of subsets of the words missing from your own dictionary).

A description of the project is available at the following address: [http://www.technolanguage.net/article.php?id\\_article=203](http://www.technolanguage.net/article.php?id_article=203) (in French). See also ELRA-L0049, ELRA-L0050, ELRA-L0051, ELRA-L0053, ELRA-M0033, ELRA-M0034, ELRA-M0035, ELRA-M0036.

---

### **ELRA-L0053 SCIPER-IT-EURADIC Italian Monolingual Dictionary**

This Italian monolingual dictionary was developed within the French national project EurRADic (European and Arabic Dictionaries and Corpora), as part of the Technolanguage programme funded by the French Ministry of Industry. It contains 70,951 lemmas (557,204 inflected forms), with their part of speech and some information related to their inflexion. The data are presented in a table format, where information related to each entry is separated by ";". Other formats and other services may be supplied by the data owner upon request (e.g. conversion into buyer's formalism, selection of subsets of the words missing from your own dictionary).

A description of the project is available at the following address: [http://www.technolanguage.net/article.php?id\\_article=203](http://www.technolanguage.net/article.php?id_article=203) (in French). See also ELRA-L0049, ELRA-L0050, ELRA-L0051, ELRA-L0052, ELRA-M0033, ELRA-M0034, ELRA-M0035, ELRA-M0036.

## PRICES FOR EURADIC DICTIONARIES

• **For L0049, L0050, L0051, L0052 and L0053:**

	ELRA members	Non-members
For research use	800 Euro	1,600 Euro
For commercial use	6,500 Euro	9,000 Euro

• **For M0033, M0034, M0035, M0036 and M0037:**

	ELRA members	Non-members
For research use	1,000 Euro	1,800 Euro
For commercial use	8,000 Euro	11,000 Euro

• **Discounts are available if you purchase several dictionaries (L0049, L0050, L0051, L0052, L0053, M0033, M0034, M0035, M0036 and M0037):**

- 10% discount for 2 dictionaries,
- 20% for 3 dictionaries,
- 25% discount from 3 dictionaries.

### ELRA-L0073 DIINAR.1 - Arabic Lexical Resource

DIINAR.1 is an Arabic Lexical Resource which was completed thanks to a joint cooperation of IRSIT in Tunisia, ENSSIB and Lumière-Lyon 2 University in France. It includes a total number of 119,693 lemmas, fully vowelled, and distributed as follows:

- 1) Nouns, including adjectives: 29,534 lemmas  
[Broken plural forms (« jumuuC taksiir »): 9,565 lemmas]
- 2) Verbs: 19,457 lemmas
- 3) Deverbals :
  - infinitive forms (« maSaadir »): 23,274 lemmas
  - active participles (« 'asmaa' al-faaCil »): 17,904 lemmas
  - passive participles (« 'asmaa' al-mafCuul »): 13,373 lemmas
  - analogous adjectives (« Sifaat mushabbaha »): 5,781 lemmas
  - nouns of place & time: 10,370 lemmas

[Total number of deverbals: 70,702 lemmas]

- 4) Total number of lemmas: 119,693 lemmas

The data is provided in Excel files and was generated with inflected forms. Each entry has been associated with morpho-syntactic 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. Morpho-syntactic specifiers belong to finite sets, but allow exhaustive processing of data.

	ELRA members	Non-members
For research use	2,000 Euro	3,000 Euro
For commercial use	8,000 Euro	11,000 Euro

### ELRA-M0033 SCI-FRAN-EURADIC French-English Bilingual Dictionary

This bilingual dictionary was increased and improved within the French national project EurRADic (European and Arabic Dictionaries and Corpora), as part of the Technolanguage programme funded by the French Ministry of Industry. It contains 243,539 pairs of French-English terms, with their part of speech. The data are presented in a table format, where information related to each entry is separated by ";". Other formats and other services may be supplied by the data owner upon request (e.g. conversion into buyer's formalism, selection of subsets of the words missing from your own dictionary).

A description of the project is available at the following address: [http://www.technolanguage.net/article.php3?id\\_article=203](http://www.technolanguage.net/article.php3?id_article=203) (in French). See also ELRA-L0049, ELRA-L0050, ELRA-L0051, ELRA-L0052, ELRA-L0053, ELRA-M0034, ELRA-M0035, ELRA-M0036.

### ELRA-M0034 SCI-FRAL-EURADIC French-German Bilingual Dictionary

This bilingual dictionary was developed within the French national project EurRADic (European and Arabic Dictionaries and Corpora), as part of the Technolanguage programme funded by the French Ministry of Industry. It contains 170,967 pairs of French-German terms, with their part of speech. The data are presented in a table format, where information related to each entry is separated by ";". Other formats and other services may be supplied by the data owner upon request (e.g. conversion into buyer's formalism, selection of subsets of the words missing from your own dictionary).

A description of the project is available at the following address: [http://www.technolanguage.net/article.php3?id\\_article=203](http://www.technolanguage.net/article.php3?id_article=203) (in French). See also ELRA-L0049, ELRA-L0050, ELRA-L0051, ELRA-L0052, ELRA-L0053, ELRA-M0033, ELRA-M0035, ELRA-M0036.

### ELRA-M0035 SCI-FRES-EURADIC French-Spanish Bilingual Dictionary

This bilingual dictionary was increased and improved within the French national project EurRADic (European and Arabic Dictionaries and Corpora), as part of the Technolangue programme funded by the French Ministry of Industry. It contains 102,941 pairs of French-Spanish terms, with their part of speech. The data are presented in a table format, where information related to each entry is separated by ";". Other formats and other services may be supplied by the data owner upon request (e.g. conversion into buyer's formalism, selection of subsets of the words missing from your own dictionary).

A description of the project is available at the following address: [http://www.technolangue.net/article.php3?id\\_article=203](http://www.technolangue.net/article.php3?id_article=203) (in French). See also ELRA-L0049, ELRA-L0050, ELRA-L0051, ELRA-L0052, ELRA-L0053, ELRA-M0033, ELRA-M0034, ELRA-M0036.

### ELRA-M0036 SCI-FRIT-EURADIC French-Italian Bilingual Dictionary

This bilingual dictionary was developed within the French national project EurRADic (European and Arabic Dictionaries and Corpora), as part of the Technolangue programme funded by the French Ministry of Industry. It contains 116,587 pairs of French-Italian terms, with their part of speech. The data are presented in a table format, where information related to each entry is separated by ";". Other formats and other services may be supplied by the data owner upon request (e.g. conversion into buyer's formalism, selection of subsets of the words missing from your own dictionary).

A description of the project is available at the following address: [http://www.technolangue.net/article.php3?id\\_article=203](http://www.technolangue.net/article.php3?id_article=203) (in French). See also ELRA-L0049, ELRA-L0050, ELRA-L0051, ELRA-L0052, ELRA-L0053, ELRA-M0033, ELRA-M0034, ELRA-M0035.

### ELRA-E0009 CHIL 2004 Evaluation Package

The CHIL 2004 Evaluation Package was produced within the CHIL Project (Computers in the Human Interaction Loop), in the framework of an Integrated Project (IP 506909) under the European Commission's Sixth Framework Programme. The objective of this project is to create environments in which computers serve humans who focus on interacting with other humans as opposed to having to attend to and being preoccupied with the machines themselves. Instead of computers operating in an isolated manner, and Humans [thrust] in the loop [of computers] we will put Computers in the Human Interaction Loop (CHIL).

In this context, the CHIL project produced CHIL Seminars. The CHIL Seminars are scientific presentations given by students, faculty members or invited speakers in the field of multimodal interfaces and speech processing. During the talks, videos of the speaker and the audience from 4 fixed cameras, frontal close ups of the speaker, close talking and far-field microphone data of the speaker's voice and ambient sounds were recorded.

The CHIL\_2004 Evaluation Package consists of the following contents:

The whole set of recordings amounts to a total of almost 6 hours of audio recordings and more than 2 hours of video recordings. The language is European English spoken by non native speakers. The recordings comprise the following: videos of the speaker and the audience from 4 fixed cameras, frontal close ups of the speaker, close talking and far-field microphone data of the speaker's voice and background sounds. The database consists of:

- 1) Audio and Video Recordings: 10 seminars (7 seminars recorded from October to December 2003 and 3 seminars recorded in June 2004).
- 2) Annotations: Video annotations done displaying 1 over 10 pictures in sequence, for the 4 cameras.
- 3) Transcriptions: Transcriptions using both TRS and STMUID formats.

	ELRA members	Non-members
For research use	200 Euro	240 Euro
For commercial use	1,500 Euro	1,800 Euro

### ELRA-E0010 CHIL 2005 Evaluation Package

The CHIL 2005 Evaluation Package was produced within the CHIL Project (Computers in the Human Interaction Loop), in the framework of an Integrated Project (IP 506909) under the European Commission's Sixth Framework Programme. The objective of this project is to create environments in which computers serve humans who focus on interacting with other humans as opposed to having to attend to and being preoccupied with the machines themselves. Instead of computers operating in an isolated manner, and Humans [thrust] in the loop [of computers] we will put Computers in the Human Interaction Loop (CHIL).

In this context, the CHIL project produced CHIL Seminars. The CHIL Seminars are scientific presentations given by students, faculty members or invited speakers in the field of multimodal interfaces and speech processing. During the talks, videos of the speaker and the audience from 4 fixed cameras, frontal close ups of the speaker, close talking and far-field microphone data of the speaker's voice and ambient sounds were recorded.

The CHIL\_2005 Evaluation Package consists of the following contents:

The whole set of recordings amounts to a total of almost 6 hours of audio recordings and more than 2 hours of video recordings. The language is European English spoken by non native speakers. The recordings comprise the following: videos of the speaker and the audience from 4 fixed cameras, frontal close ups of the speaker, close talking and far-field microphone data of the speaker's voice and background sounds. The database consists of:

- 1) Contents of the CHIL 2004 Evaluation Package (see catalogue reference ELRA-E0009 for description).
- 2) Audio and Video Recordings: 5 seminars recorded in November 2004).
- 3) Stereo Video Recordings of 10 subjects that move in the camera's field of view while performing pointing gestures.
- 4) Video annotations.
- 5) Transcriptions.

	ELRA members	Non-members
For research use	700 Euro	850 Euro
For commercial use	3,000 Euro	3,600 Euro

### **ELRA-E0011 TC-STAR 2006 Evaluation Package - ASR English**

This package includes the material used for the TC-STAR 2006 Automatic Speech Recognition (ASR) second evaluation campaign for the English language. It includes resources, protocols, scoring tools, results of the official campaign, etc., that were used or produced during the campaign. The aim of these evaluation packages is to enable external players to evaluate their own system and compare their results with those obtained during the campaign itself.

### **ELRA-E0012 TC-STAR 2006 Evaluation Package - ASR Spanish**

This package includes the material used for the TC-STAR 2006 Automatic Speech Recognition (ASR) second evaluation campaign for the Spanish language. It includes resources, protocols, scoring tools, results of the official campaign, etc., that were used or produced during the campaign. The aim of these evaluation packages is to enable external players to evaluate their own system and compare their results with those obtained during the campaign itself.

### **ELRA-E0013 TC-STAR 2006 Evaluation Package - ASR Mandarin Chinese**

This package includes the material used for the TC-STAR 2006 Automatic Speech Recognition (ASR) second evaluation campaign for the Mandarin Chinese language. It includes resources, protocols, scoring tools, results of the official campaign, etc., that were used or produced during the campaign. The aim of these evaluation packages is to enable external players to evaluate their own system and compare their results with those obtained during the campaign itself.

### **ELRA-E0014 TC-STAR 2006 Evaluation Package - SLT English-to-Spanish**

This package includes the material used for the TC-STAR 2006 Spoken Language Translation (SLT) second evaluation campaign for English-to-Spanish translation. It includes resources, protocols, scoring tools, results of the official campaign, etc., that were used or produced during the campaign. The aim of these evaluation packages is to enable external players to evaluate their own system and compare their results with those obtained during the campaign itself.

### **ELRA-E0015 TC-STAR 2006 Evaluation Package - SLT Spanish-to-English**

This package includes the material used for the TC-STAR 2006 Spoken Language Translation (SLT) second evaluation campaign for Spanish-to-English translation. It includes resources, protocols, scoring tools, results of the official campaign, etc., that were used or produced during the campaign. The aim of these evaluation packages is to enable external players to evaluate their own system and compare their results with those obtained during the campaign itself.

### **ELRA-E0016 TC-STAR 2006 Evaluation Package - SLT Chinese-to-English**

This package includes the material used for the TC-STAR 2006 Spoken Language Translation (SLT) second evaluation campaign for Chinese-to-English translation. It includes resources, protocols, scoring tools, results of the official campaign, etc., that were used or produced during the campaign. The aim of these evaluation packages is to enable external players to evaluate their own system and compare their results with those obtained during the campaign itself.

#### **PRICES FOR TC-STAR EVALUATION PACKAGES**

<b>Prices per package</b>	ELRA members	Non-members
For evaluation use	500 Euro	750 Euro
<b>Special prices for a combined purchase of TC-STAR Evaluation Packages:</b>		
• <b>2006 ASR Suite (E0011 + E0012 + E0013):</b>	ELRA members	Non-members
For evaluation use	1,200 Euro	1,800 Euro
• <b>2006 SLT Suite (E0014 + E0015 + E0016):</b>	ELRA members	Non-members
For evaluation use	1,200 Euro	1,800 Euro
• <b>2006 ASR + SLT Suites (E0011 + E0012 + E0013 + E0014 + E0015 + E0016):</b>	ELRA members	Non-members
For evaluation use	2,000 Euro	3,000 Euro
• <b>2005 + 2006 ASR + SLT Suites (E0002 + E0003 + E0004 + E0005 + E0006 + E0007 + E0011 + E0012 + E0013 + E0014 + E0015 + E0016):</b>	ELRA members	Non-members
For evaluation use	3,200 Euro	4,800 Euro