The ELRA Newsletter

October - December 2002

Vol.7 n. 4

IN MEMORY OF DON ANGEL MARTIN MUNICIO

Contents

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

Editor in Chief:
Khalid Choukri

Editors:
Khalid Choukri
Valérie Mapelli
Magali Jeanmaire

Layout:
Magali Jeanmaire

Contributors:
Sandra Berrebi
Khalid Choukri
Daniel Tapias
Jean-Paul Taravella
Laurie de Temmerman

ISSN: 1026-8200

Translations:
Anna O'Hora-Bimbot

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

Dear Colleagues,

Angel Martin Municio was vice-president of ELRA, and gave many years of dedicated services to the association. We have learnt of his death with extreme sadness, and would like to express sincere condolences to his family and friends. Daniel Tapias, member of the ELRA Board, and Khalid Chourki, ELRA CEO, agreed to contribute an obituary, published in this issue of the ELRA newsletter released in his memory. An article from Angel Martin Municio is also included, entitled “The Origin of Science and Scientific Language”.

The European 6th Framework Programme (FP6) was officially launched last November. The field of Information Society Technologies (IST) is the principal thematic priority of this programme, with a budget of €3625 million. The 1st call for proposals was published mid-December. ELRA/ELDA is planning for strong commitment with FP6, willing to contribute with the two infrastructures set up during the last few years: the one related to language resources, and the other to evaluation of HLT. Within FP6, ELRA/ELDA is willing to assist all projects (integrated projects, network of excellence) through the services it offers with the support of large networks of partners able to handle issues regarding LRs specification, production, validation and distribution, as well as HLT evaluation issues regarding the set up and management of evaluation campaigns, the production and packaging of LRs for evaluation, and all related logistics. The use of ELRA networks (production and validation of LRs, evaluation) ensures an efficient and cost-effective capitalisation of the expertise and know-how of a large set of partners all over Europe and beyond.

ELRA has been carrying the management of such networks for a while and is able to elaborate organisational models for any project within FP6. So please do not hesitate to ask for our assistance for your proposal preparation. ELRA is also working towards the set up of a European Research Area on HLT.

There has been a number of notable achievements at ELRA during the last quarter: major updates have been implemented on our web sites, and our presence in the field of HLT has been further increased, notably by highlighting key aspects which concern the whole HLT community. A directory of the HLT key players and tools in France has been published on the ELDA web site. It is updated on a regular basis, thanks to the information you may wish to provide us, by returning the form available in this section, at the following URL: www.elda.fr/fr/proj/euromap/register.php. This database was developed in the framework of the Euromap Language Technologies project, which has obtained a 2 months extension from the European Commission, until the end of April 2003.

The ELRA web site can now be visited at a new web address, www.elra.info. It has seen a number of changes in the last quarter, notably in its Validation section, updates on the catalogue, etc.

ELRA also maintains the web site set up in the framework of COCOSDA, the international COordination COmmittee for Speech Databases and speech I/O systems Assesment. You will find on this web site, at http://www.cocosda.org, the presentations given during the COCOSDA 2002 workshop, which elaborated on the regional activities conducted in 2002 within COCOSDA, as well as some topics like dialog evaluation, minority languages, multimodal resources, standards, etc.

One of ELRA main concern is the validation of language resources, to which the association devotes large efforts. ELRA Validation Centre for Written Language Resources (VC_WLR) was selected; CST, the Centre for Sprogteknologi, Denmark, will take care of the co-ordination of a network of validation centers for written language resources, which will validate the WLR available in ELRA's catalogue, and will participate in the elaboration of new procedures. Information material provided by ELRA VC_WLR can be viewed on the ELRA web site. As for the validation of SLR, SPEX, the SPeech EXpertise centre, the Netherlands, had been appointed some time ago as ELRA VC_SLR. Several Quick Quality Checks have already been produced. These documents aim at describing the quality of the SLR which are distributed by ELDA, and should be made available in the near future on our web site, along with the description of the concerned resource. Last but not least concerning the issue of validation, a new prize will be awarded to the best SLR bug reporter. The bug report service was launched by ELRA at the beginning of year 2002. Two prizes (PDAs) have already been awarded this year. The call for a 2002 third prize will be open until end of February 2003. You still have time to submit bug report forms, which can be completed and submitted from the Bug Report Service section on the ELRA web site, http://www.elra.info. A similar form will be implemented shortly for WLR.

ELRA is also much involved in the evaluation of HLT. We are now expanding our activities; new staff members have already joined us to work in this area, and others should start by early Spring 2003.

During this quarter, we have been very active and successful regarding the language resources' collection: 24 new resources, including 6 SpeechDat-Car databases, can be found in our catalogue (ELRA references and descriptions at the end).

ELRA production activity has been boosted, with its participation in several projects: the first recordings have started for OrienTel, a European project which aims at designing and developing multilingual interactive communication services for the Mediterranean languages (26 languages will be covered). Within Speecon, where voice driven interfaces for consumer applications will be produced for 18 languages., we were responsible for three languages speech recordings and transcriptions (French, Swedish and Italian) which are now finished. More information on http://www.speecon.com. Lastly, within the SALA II project, which aims at creating speech databases covering the languages spoken in Americas (South, Central, and North America), ELDA is responsible for collecting the speech data in American English. You will find more at http://www.sala2.org.

Apart from these projects we participate in for the production of resources, a number of issues deserve to be mentioned, in the framework of other French, European and international projects. Some of these are mentioned below.

Within Technolangue, the French national programme on language technologies, the set-up and the launch of an HLT portal are being discussed among the project partners, with ELDA as co-ordinator. A draft version of this new web site should be released shortly. In the Technolangue evaluation section, ELDA co-ordinates the Evalda project, which includes 8 sub-projects addressing written and spoken language systems evaluation.

We have been involved in 2 other French projects, PRISME, to set up a technology intelligence watch tool on the Internet dedicated to SMEs, and AVISE, which planned to offer access over the phone to press releases and news extracted from the French financial newspaper Les Echos. Within the latter, we were in charge of evaluating the prototype developed by our partner, Elan Informatique. Both projects are still running, and no results are available yet.

Furthermore, we have launched among ELRA members and partners a survey about the industrial requirements with respect to language resources in the framework of the ENABLER project, which stands for European National Activities for Basic Language Resources. Feel free to contact us if you want to receive more information about this issue.

Now, as for the content of the newsletter, two articles dealing with written language processing are included in this issue. The first paper, written by Laurie de Temmerman and Jean-Paul Taravella, elaborates on the use of ontologies in NLP, illustrated by the ontology used within the European project CLASS (Collaboration in Language and Speech Science). Later on, ELDA plans to compile a single set, combining this with our own ontologies and typologies. In the second article, from Sandra Berrebi, the focus is on the need for collaborative work to improve terminology management, a central issue in the framework of multilingual environments. A useful application, which allows several users to manage the terminological information simultaneously in the framework of a specific project, is presented: LexSyn.

The new resources for which you will find the detailed descriptions in the last section of this issue are: British English SpeechDat-Car (S0131), Danish SpeechDat-Car (S0132), Finnish SpeechDat-Car (S0133), Concise Oxford Dictionary - Audio files (S0134), French SpeechDat-Car (S0135), SmartKom Multimodal Corpus SKP 1.0 (S0136), TAXI Multilingual Telephone Dialog database (S0137), Cantonese SpeechDat-like database MDB-2000 (S0138), Flemish/Dutch SpeechDat-Car (S0139), Spanish SpeechDat-Car (S0140), SALA Spanish Venezuelan database (S0141), Austrian SpeechDat (AT) FDB-1000 (S0142), Austrian SpeechDat (AT) MDB-1000 (S0143), New Oxford Dictionary of English 2nd Edition (L0045), NODE + DIMAP (L0046), New Oxford Thesaurus of English (L0047), Oxford Paperback Thesaurus (L0048), Oxford French Minidictionary (M0027), Concise Oxford Duden German Dictionary (M0028), Pocket Oxford Italian Dictionary (M0029), Concise Oxford Spanish Dictionary (M0030), Oxford Business French Dictionary (M0031), Oxford Business Spanish Dictionary (M0032), and Telecommunications Dictionary (T0367).

We wish you a happy new year, and have a pleasant reading.

Don Ángel Martín Municio, Vice President of ELRA, passed away in November

In this new issue of the ELRA newsletter, we have the painful mission of informing you of the very sad news that, on November 23rd, Professor Don Ángel Martín Municio, passed away. He was president of the Spanish Royal Academy of Exact Sciences, Physics and Natural Sciences, as well as a full member of the Royal Spanish Academy and vice president of ELRA.

Professor Don Ángel had just returned to Madrid from a long and exhausting trip to China, where he had participated in the organisation of an International Conference of the Scientific Academies to be held in Spain. Previously, he had made a trip to Costa Rica to participate in a meeting of the Royal Spanish Academies.

As he arrived in Madrid, he felt very tired, and suffered a heart attack a few hours later that caused his death.

We are shocked by this news, but we also feel grateful for having had the opportunity to collaborate and share long working hours with a person like him.

We will always remember his capacity to work, his willingness to help and his advice and support despite his age (he was 78 years old) and the large amount of responsibilities he had, but mainly, we will remember and miss his humanity and friendship, his sense of humour and his capacity to enjoy life.

Dear Don Ángel, Rest in Peace.

Obituary by Daniel Tapias

Don Ángel was born in Haro, a small village in the region of "La Rioja" (Spain), on November 30th, 1923. His father, who was a judge, was appointed to Salamanca, where Don Ángel studied Chemistry, receiving his degree in 1946. He later studied at the University of Santiago de Compostela, where he received the Pharmacy University degree in 1948.

Don Ángel started his professional career in 1948 as organic chemistry assistant professor at the Faculty of Sciences of the University of Madrid, and continued in this position until 1951. During this period, he received a PhD in Chemistry at the Central University of Madrid.

From 1951 to 1954, while Don Ángel was a fellow of the Consejo Superior de Investigaciones Científicas (CSIC), he received a grant from CSIC to work at the Rijks University of Utrecht (Holland). He also worked at the Max-Planck Institute in Heidelberg (Germany), and received a PhD in Pharmacy at the Central University of Madrid.

In 1955, he received a grant from the Spanish Ministry of Foreign Affairs to study in the Medical Research Council of Mill Hill (London, UK). Later, in 1963, he conducted research for the Twyford Mill Laboratories (London, UK). In 1965, he worked at the Organic Chemistry Department of the University of Newcastle. From 1955 until 1966, he was Scientific Researcher and Biochemistry Section Manager at the Chemistry Institute (Spain) and from 1960 to 1966, he was representative for Spain at the OCDE.

During the period 1965-72, Don Ángel was director of the Biology Department of the Juan March Foundation. From 1967 to 1989, he was professor of Biochemistry and Molecular Biology in the Faculties of Biology and Chemistry at the Complutensis University (Madrid, Spain). He was also appointed director of the Biochemistry department at the Complutensis University.
In 1969, he worked in the Medical Research Council (Cambridge, UK) and became the first Spanish member of the European Molecular Biology Organisation (EMBO). Also, on February 11th, 1969, he became a full member of the Royal Academy of Exact Sciences, Physics and Natural Sciences (Madrid, Spain), in the area of Natural Sciences. From 1985 until his death, he was the president of this institution. In 1990, he received the IV Leonardo Torres Quevedo award on behalf of the Royal Academy of the Exact Sciences, Physics and Natural Sciences.
From 1962 until 1990, Don Ángel was representative for Spain at the European Molecular Biology Conference and, from 1982 until 1990, he was also vice president of the European Biology Conference. During the period 1982-86, he was vice principal for Research and International Relations of the Complutensis University. While he worked in this position, he was the advocate of Dr. John Kendrew as "Honoris Causa" Doctor by the Complutensis University and of Her Royal Highness Queen Elisabeth II when she was granted the Gold Medal of the University. Additionally, he was director of the Basic Research Department of the Spanish Health Ministry for Toxic Syndromes (1983-84) and director of the Research Service of the Regional Plan of Oncology of the region of Madrid (1982-84).
From 1983 until 1997, he was member of the Board of the Ortega y Gasset Institute. In 1982, he was elected full member of the Royal Spanish Academy (RAE), and in 1984, he became a full member of this institution. Later, from 1992 until 2000, he was the vice director of the Royal Spanish Academy. Don Ángel was one of the most active members of the scientific terminology seminar at the Royal Spanish Academy, which published the first volume of the Scientific and Technical Dictionary. He also played an important role in many other areas, like the approval of the statutes of the institution.
Don Ángel was a member of the Board of the following foundations: "Ideas and Historic Research Foundation" (from 1987), "Bilbao Vizcaya Bank Foundation" (1990-94), "Central Hispano Bank Foundation" (from 1994), "Environment, Companies and Environment Foundation" (from 1996), "Scientific Foundation of the Spanish Cancer Association" (1989-2001), "San Millán de la Cogolla Foundation" (from 1998) and "Antonio de Nebrija University Foundation" (from 1998).
From 1996, and until his death, Don Ángel was vice president of the Board of the European Language Resources Association (ELRA). He played an important role both in the organisation of the Language Resources and Evaluation Conferences (LREC) and in the management of the Association. The last LREC conference in which he was involved took place in Las Palmas de Gran Canaria in May 2002 and was, without doubt, the most successful, both in terms of the number of participants (more than 700) and the number of published papers.
Don Ángel was also a member of the Board of the Cervantes Institute from 1992, and became a member of its Board of Directors in 1996. He was a member of the Colegio Libre de Eméritos (Emeritus College) from 1990, president of the Scientific Board of the Foundation to Support the National Science and Technology Museum, a full member of the European Academy of Arts, Sciences and Humanities from 1992, full member of the Academia Scientiarum et Artium Europaea from 1997 and vice president of the Spanish section from 1999, correspondent member of the Academy of Physical Sciences, Mathematics and Natural Sciences of Venezuela from 1993, honorary member of the inter-American Medical and Health Association (USA) from 1993, distinguished lecturer in Biomedicine, granted by the School of Medicine of Louisiana State University (USA) 1993, correspondent member of the Academy of Sciences of Colombia (1994), honorary member of the Academy of Sciences of the Dominican Republic (1994), correspondent member of the Academy of Sciences of Russia (1996), honorary member of the Academy of Spanish Language of Colombia (1997) and correspondent member of the Academy of Sciences of Latin America (2001).
Don Ángel had a very great scientific reputation. He published several hundreds of scientific papers in international journals within the areas of Biochemistry and Molecular Biology. He was also author and editor of books within these specialities, director and author of the Scientific and Technical Vocabulary of the Royal Academy of Sciences (issues 1983, 1988 and 1992), of the Dictionary of Essential Sciences (Royal Academy of Sciences, issues 2000 and 2001) and of the Spanish Energy Dictionary (Royal Academy of Sciences, first issue, 2002), director of the Programme to Promote Scientific and Technological Culture (Royal Academy of Sciences, 1998, 1999, 2000 and 2001), director of 53 PhD theses, author of essays, scientific publications and literary review of scientific works, attendant to hundreds of conferences and scientific meetings, and a director of courses and lecturer at many academic and research institutions.
Don Ángel was also awarded the Medal of "Medalla al Mérito Investigador of the Spanish Royal Society of Physics and Chemistry", the Cross of "Alfonso X El Sabio", the Great Cross of the Military Merit, La Rioja Gold Medal, the Honorific Medal of Invention Sponsorship, the Medal of the Complutensis University, the Medal for the Merit of the Colombian Government and the Gold Medal of the Spanish Cancer Association.

Obituary by Khalid Choukri

I met Don Angel when I first started to work for ELRA, at a meeting with A. Zampolli, B. Quemmada, and B. Oakly. The purpose of the meeting was to review the international partnerships ELRA should seek to set up. I was impressed by his knowledge of the HLT area but also of areas such as biology, physics, etc. I was impressed by his open views on what should be done for international co-operation. Later on, he invited us to the launching ceremony of the Scientific and Technical Dictionary (Vocabulario Cientifico y Tecnico), in the presence of the King of Spain, and had the sympathy to introduce ELRA to his Majesty Juan Carlos.

A few weeks later, we started to discuss the possibility to establish a conference that would focus on both language resources and evaluation, and which would bring together all involved experts. As usual, he took an active role in our debates and committed himself on several aspects, not only suggesting Granada as the site to host the event, but also using his impressive network of partners to secure sponsorship and patronage for the event.
He did even more, despite his so many professional obligations. He took an active role in the programme committee discussions and contributed the establishment of the final programme. He continued to work with us on the second LREC, that took place in Athens, and again, when discussions arose about the third LREC, he was ready to help and assist us. He had already anticipated our expectations and contacted local authorities in Las Palmas de Gran Canaria to obtain their support. He did his outmost to ensure an efficient media coverage and the patronage of the highest Spanish authorities. These are but a few memories I wanted to share with you. I do not think words are enough to convey all what I appreciate in Don Angel. His quick-wittedness combined with his enthusiasm for life made him a unique person and many of us will miss him.
The last time I met Don Angel was in Las Palmas, and again I appreciate how he took care of tiny details to make the event successful. Don Angel, rest in Peace.

The article below was presented by Don Angel Martin Municio on 17th October 2001 at the 2nd international Conference of the Spanish Language (16-19 October 2001), during the plenary session entitled "The Scientific and Technical Spanish Language". The complete paper (in Spanish) can be viewed at: http://cvc.cervantes.es/obref/congresos/valladolid/plenarias/martin_a.htm.

We would like to thank the Cervantes Institute for authorising ELRA to publish in its newsletter extracts from this paper.

The Origin of Science and Scientific Language

It is of not too much importance whether the first signs of an incipient scientific communication were the simple algorithms of the small clay boards in the paleobabylonian period, or the volume computation in the Egyptian papyrus of the second millennium before our era, or the Mesopotamian sexigesimal system.

This is not relevant because mathematical communication was really born when, hundreds of thousands of years earlier, the synaptic connections of the evolving brain allowed human beings to count, in coincidence with the origin of language and in coincidence with the origin of his own nature, so that later by interacting with thought, abstract reasoning starts. And it is not in vain that philosophers, linguists and anthropologists coincide with one another to recognise that if there were no signs, we would not be able to recognise ideas. The thought in itself would be like a nebula where nothing is necessarily delimited, and where nothing would be different before the appearance of language. Therefore, thought would be impossible without language; and, moreover, as it was stated by the great naturalist Buffon in the 18th century, "man speaks because he has the power of reason". And, because of this, every human being thinks in his own native language, and language identifies with his imagination and feelings.

In this sense, we should not forget that language is the first science that human beings own. Language is a first classification of the world, and it shows us the organisation of reality. But this initial scientific description, by means of natural language, hardly serves to describe certain kinds of scientific realities. The development of science and the continuous appearance of new domains are accompanied by the need to improve natural language. Natural language serves, however, as a canvas on top of which, the specific terms of scientific language are integrated, with more or less pretensions of universality, and even the symbolic systems themselves with total ambition of universality. It is like a special level belonging to general language; it is about a modified language, a sign system with less ambiguity, which is used together with the general language in variable proportions. In other words, between natural language and logic-mathematical language with a larger degree of symbolism, a gradient of scientism, that tends towards abstraction and a better adjustment to the structure of reality, exists. However, every scientific domain tries to create a symbolic language which is appropriate to its object. Objectivity and quantification move away from the usual modes of language, while it tries to adapt in order to reach the same goals.

In this way, language always served to express concerns of thinking about the origin and nature of human beings and the universe. And an expression of these concerns was the literary mythic-religious creations in all languages; which would lead to the artistic exaltation of myths and, in parallel, to philosophic and mathematical reasoning. Surely, this has also something to do with the fact that Greek mathematics was born in perfect unity with philosophy; and from this identification, the axiomatic-deductive method, which is still in use for proving the truths established by theorems, was born.

From this time onward, the books by Archimedes, Euclides and Apolonio de Parga, and their systematisation, were valid until the Renaissance. And, among them, Euclides' Elements 13 books which, together with the Bible, are the two books that have seen more editions and are among those which have influenced culture in the history of civilisation the most, which compile definitions, postulates, axioms and propositions perfectly. Later, Rey Pastor asserted: "if you tried to add or remove something you would immediately recognise that you move away from science and approach error and ignorance".

The Middle Ages, in which the ten Arabic and Latin centuries played an important role in the origins of European science, then had to be crossed. Not in vain, calculus - arithmetic, algebra and trigonometry - , the sciences of the concrete and the practical, owe more to oriental than to Greek science; and, in these routes towards a Europe of mathematics and Greek-oriental science, Spain was, without doubt, the main route. The role of the Arabic world was more than a simple intermediary and it is mainly represented by the algebra of Al-Khawarizmi and the spherical trigonometry of Ibn al-Haytham, or Alhazen, author of a treatise on optics which served as a guide for the knowledge of light and vision during the Middle Ages in western Europe.

Precisely, Right at the end of the first millennium, one of the interactions between western European and Arabic science centres in the peninsula was the one represented by the treatise "De Astrolabia", from the Bishop of Reims, Gerberto d'Aurillac, who later became Pope Silvestre II.

At the same time, the Canon of Medicine from the physician and philosopher Ibn Sina, or Avicena, consolidated the medical knowledge accumulated by the Greeks, Romans and Arabs. The "Óptica", from Alhazen and the "Canon" from Avicena have traditionally served as masterpieces, framing the state of universal scientific communication in the transition to the second millennium.

On the other hand, this meant the transcendental beginning of Castilian Spanish, when, in the 10th century, the scriptorium of San Millán, nerve centre of its library, was able to participate in the future of the new language with the "Comentarios a los Salmos" (Comments to the Psalms), the copy of the "Ciudad de Dios" (City of God) from San Agustín and, principally, the famous "Códice 46", encyclopaedia of that epoch including the vocabulary, culture and thought at the Middle Ages. And, before these initial documents, before the linguistic innovations and hesitations of a millennium ago, we cannot but recognise together with Marañón the efforts of those who preceded us in past centuries because "those who inherit great riches are not as aware of it as the ones who had to endeavour to obtain it. We are privileged to have learnt the language of Castile. But we have to earn this privilege every day with our effort and love. The treasure of an illustrious language implies a service of being permanently on the alert and of striving for constant perfection". It has been deserved, sure enough, by those who, on the one hand, after monastic babbling, have passed the token of the perfection and the beauty of language up to our days, and on the other hand, those who spread language across seas and continents in military, missionary, colonial and cultural enterprises.

However, five centuries before Castilian Spanish became "universalidad" in Lope, and, in Cervantes, "the beginning of modern times" in the history of mankind; and much before the Spanish language had its norm in the grammar of Nebrija, it already served Alfonso X el Sabio as the language of science and technology in "El Saber de Astronomía" and in "El Lapidario", and as an encyclopaedic language of culture and law in "Las Partidas".

The American Experience

If the transition to the 16th century meant, with Christopher Colombus and Vasco de Gama, the discovery of new worlds, Galileo also discovered new worlds in the transition to the 17th century, when he directed his telescope at the sky: Jupiter's moons, the Venus phases, Sun spots and mountains of the Moon. In this way, if Europe had to start sharing its presence on earth, with America, the geocentric cosmos had to leave the way open for a heliocentric image of the universe, and humanity was shifted from a predominant central position, in the middle of everything, to the peripheral position of a minor planet. And there is no doubt that the medieval social and political changes under the influence of the inventions of technology, the discoveries and descriptions of the new worlds and the relaxation of the scholastic principles, experienced a coalescence that favoured the birth of modern science. The "Principia" from Newton, in 1687, as a model for the exact description of nature, meant the beginning of law and order in the physical world, and the possibility of reaching a description of the human body and mind.

In the mean time, Castilian Spanish had become universal with the "Crónicas de Indias" (Chronicles from the Indies) from the principal discoverers -Colombus, Cortés, Díaz del Castillo, Valdivia, Núñez Cabeza de Vaca, Jiménez de Quesada and Cieza de León-. And if language had been, with the Inca Garcilaso, at the same time both a splendid example of linguistic and cultural transfer and the origin of American literature; the Discovery, its expeditions, travels and navigation were also the reasons for several of the Spanish knowledge adventures in those centuries. It is well known that this was the age of the travels of Magallanes and Elcano, of Pizarro stay in Peru and Cortés in México, and of the opening of the Indies route by Vasco de Gama. All this influenced the outstanding interest in navigation and cartographical applications of physics and mathematical science. And, at the same time, the environment of the Spanish Court favoured the fostering of mathematical pragmatic applications: cosmography, cartography, geodesic measures, astrology, navigation, building and architecture techniques, and military engineering.

Neither the decline of our mathematics nor the fact that the cause of errors in our charts were due to the lack of scientific knowledge could be hidden from the sagacity of Felipe II. For this reason, and as a reaction to the new discoveries and to assure explorers' success, and the resolution of practical problems, Felipe II signed in Lisbon, on 25th December 1582, a document to found the Royal Academy of Mathematics, of which the architect Juan de Herrera would be the first director. (…)

If in the 16th century and part of the 17th, the mission of the Royal Academy of Mathematics was the scientific and technical knowledge required to make discovery, colonisation, and, in particular, great scientific expeditions (which represented an impressive adventure of knowledge), possible. The scientific expeditions, focusing mainly on botany, mining and metallurgy, brought illustrious names - Alonso Barba, José Celestino Mutis, the Fausto brothers and Juan José Elhuyar, among others - who started many social, scientific and philosophical initiatives, including universities, under the protection of the language. Boundless initiatives, experiences and enthusiasm of scientists, which strengthened the cultural personality of the American Kingdoms. And it was in this environment, when the Spaniards achieved one of the best Spanish contributions to the history of chemistry; since this period was the only one in history when Spanish names were inscribed in the most famous, universal and permanent scientific charts: the periodic table of the elements. There are three chemical elements - Wolfram, Vanadium and the Platinum - in whose history of discovery appear, respectively, the Elhuyar brothers, Andrés Manuel del Río and the famous sailor Antonio de Ulloa. (…)

And Rodríguez Carracido, director of the Madrid University and President of the Royal Academy of Sciences, referred to the news and scientific communications which arrived from America in the following: "The culture imported by the Bourbon dynasty was purely literary in the beginning, but the great value assigned to what was called useful knowledge promoted scientific studies, giving more importance to those which led to the development and increase of natural productions. Our statesmen, influenced by the trends of their century, showed an important interest in having an inventory of the mineral and vegetable riches of the colonies, and with this desire the scientific Hispanic American literature was reborn". The important studies concerning mining and the American flora, and their corresponding scientific communication in Spanish, did not connect with the European science. Neither our politicians, nor the scientists and philosophers knew how to share methodological innovations that implied the science autonomy, among other novelties of social and political life in the 17th century; or how to join the later revolution and use of chemistry, in the 19th century, which achieved the isolation and identifi-

cation of many natural products obtained from plants. The Spanish language, that provided brilliant contributions to American flora, was unable to serve as a means of communication to later science, though it facilitated the way for other European languages. (…).

The modern science. Europe and Spain

Consequently, Spain and its language contributed very poorly to the mission of naming discoveries in the 19th century. It is important, on the other hand, that we remember what nascent European science had reserved. Because this was the century in which Bernard (1813-1878) and Pasteur (1822-1895) lived, and discovered aetiology of illnesses and the first vaccines, and Darwin (1809-1882) published "On the Origin of Species by Means of Natural Selection". These were also the years when Koch (1843-1910) isolated cholera and the tuberculosis bacillus, when Behring (1854-1917) discovered the antibacterial serum. These were the years of the genetics of Mendel (1865), of the chromosome discovery by Fleming (1875) and of the functional centres of the brain by Charcot (1825-1893), of the synthesis of natural products, like indigo, by Bayer (1879), and the great development of the organic synthesis by Berthelot (1860), of the discovery of the x-ray by Röntgen (1895) and of radioactivity by Becquerel (1896). These were the years of the thermodynamics birth (1853), of the valency theory (1858), of the electromagnetic fields by Maxwell (1864), of the gas kinetics theory by Boltzman (1877), of the set theory by Cantor (1883), of the mathematical logic by Frege (1892) and of the algebraic numbers by Hilbert (1897). These were also the years in which the industrial production of aspirin, of aluminium and of the first artificial colouring was carried out; the first oil well was made and the first spark-ignition engine and the first car with a four-stroke fuel engine were designed; the first transatlantic communication cable was laid and Bell invented the telephone; celluloid and artificial silk were manufactured; the first electric locomotive from Siemens and the electric street lighting of New York started to work; and the first public cinema performance took place. (...).

Natural Language Processsing - A Further Ontology

Laurie de Temmerman and Jean-Paul Taravella

In this article, we present the ontology under development in the context of the European Project CLASS, which aims to describe the resources and tools used in the field of Natural Language Processsing (NLP). We propose a brief overview of industrial issues, recall general issues related to ontologies, present a number of existing ontologies which describe the NLP field, and we show how our approach, based on a functional description of objects, enables improved perception and communication concerning NLP objects. This complements harmoniously what already exists.

Introduction

Our experience relies on recent studies and integration work carried out by SchlumbergerSema in the field of NLP:
- *EDF-DRD*: comparison of several lingware[1] search engines.
- *DGA*: audit on the internal search and filtering technologies, comparison of similar lingware, development and evaluation of the prototype and of the final knowledge management systems.
- *France Telecom*: comparison of several lingware search and filtering programs for awareness applications, development of the prototype and of the final system.
- *EUROSTAT*: evaluation of available components for a morpho-syntactic tagger, in 11 European languages, development and evaluation of an application prototype for assisted document classification according to a pre-existing nomenclature.

Each of these studies has involved SchlumbergerSema as a general-purpose integrator together with software (lingware) producers, component providers and even specialised integrators in the field of NLP. Moreover, each of these studies has required several steps of technology evaluation (acquisition of the technology, improvement of the prototype, evaluation of the final system before delivery).

The lack of a structured and stable terminology in the NLP domain creates problems at the industrial level:

1/ The studies mentioned above involve many exchanges between various players and an acute common understanding of the concepts handled.
- Between project members: how to establish fast and efficient communication? The general-purpose integrator has a more applicative and managerial approach to the project, whereas the specialised integrator has a deeper technological understanding of the field. Both types of know-how are needed for satisfactory progress of the project.
- Between the project team and the technology provider: which technology is precisely referred to, and what is its actual functionalities? Today, we observe an ever-increasing terminology complexity from software providers, with the hope to outbid and overcome competitors.
- Between the project team and the client: among several possible technological options, which one should be chosen and validated, given its performance and its cost? The client has to have a minimal knowledge and understanding of the principles, the functioning, the performance and the cost of the various technological alternatives that can fulfil a given requirement.

A common language would greatly facilitate these exchanges. In particular, the field lacks a global synthetic view that would enable a positioning of the various technologies relative to one another.

2/ These projects also involve repeated technology assessment.
- In black-box mode, for the components that need to be purchased externally: which is the most adapted component for the targeted application?
- But also in glass-box mode, for the intermediate versions of the prototype application: which is the lingware module or component which is defective in the overall application system? Applications integrate several lingware modules or components, lingware themselves integrate other components, terminological resources are involved at different levels, etc.

A structured representation of the field would facilitate the set up of such evaluation processes, by guiding the architecture, tuning alternatives and simplifying communication between experts from different fields.

It seemed to us that a standardisation effort of the terminology aspects and the elaboration of a structured representation of the field, under the form of an ontology, was absolutely necessary to respond to these needs.

What is an Ontology?

Ontologies are not compilations of the world. They only conceptualise the elements that compose a field and thus yield a representation of concepts with respect to

the user needs. In fact, "an ontology bears the marks of the particular task for which it has been built, as well as the reasoning process behind it"[2]. Therefore, existing ontologies (and not only in the NLP field) fulfil, according to Marek Obitko[3], a need to represent knowledge semantically, to conceptualise it, to facilitate communication between agents, to speed-up information retrieval or even to improve learning abilities. In the framework of the ACACIA[4] project led by INRIA[5], work in collaboration with Renault has resulted in a prototype called SAMOVAR[6] which is composed, on the one hand, of semi-automatically designed ontologies and, on the other hand, of a project memory. Other ontologies, such as "Onto-PME" (Team Memory Project)[7], achieved by LaRIA, a laboratory in the University Of Picardy, aim at facilitating knowledge communication and exchanges. Also in the framework of the ACACIA project, as well as in the IST-CoMMA project, the learning of ontologies is based on RDF annotation of Web documents. Each concept is defined "in extension, by a collection of resources and its definition in intention generalises the descriptions of the resources extracted from the RDF network". Ontologies also improve the relevance of information retrieval tasks. They can be used to search and extract knowledge from the Web, to refine the content of a page or to standardise requests. The "OntoSeek" system, as developed by Nicola Guarino, Claudio Masolo and Guido Vetere[8], yields collections which it is possible to search through, such as the yellow pages, using linguistic ontologies, in particular WordNet.

Ontologies and NLP

WordNet remains the reference in terms of linguistic ontologies. It was developed at the University of Princeton and offers a dictionary of definitions. However, as for NLP, very few ontologies exist: the "Registry" of the ACL (Association for Computational Linguistics) maintained by the DFKI and the "Activities in NLP" ontology, created by ELRA. The first one presents available products in the field, with a link to the corresponding web sites; the latter, designed for the French Ministry of Education, Research and Technology, aims at listing NLP products. It is thus the purpose and the needs of a particular ontology that constrain its features, especially its coverage of the field.

With respect to what is currently carried out, it seems quite obvious that theses two ontologies fulfil the need to represent and manage knowledge. But it seems to us that neither of them offer the possibility to help and assist decision-making processes, in particular in the context of project management. Therefore, it would appear that the ontology designed for this purpose by LIMSI in the context of the CLASS Project[9] opens the way to a new type of application.

The goal is to provide the possibility for several participants to work simultaneously on research and development projects, when they use the same resources or when they design systems which have the same functions. This is expected to increase performance. It is also a way to represent all projects in a more accurate and formalised way than with any existing ontology, even if the price to pay is a relative complexity of the formalism, which does not make it adapted for other purposes (such as those covered by the ELRA ontology or by the ACL Registry).

Principles of the CLASS Ontology

The CLASS ontology has been designed with the use of a document management method using candidate terms coming from various sources: indexes such as the Walloon Office for Language Industries [10] and terminological glossaries, in particular that of the University of Montréal[11]. It provides a distinction between "passive" and "active" linguistic products, so as to differentiate between resources versus tools or applications. It was found to be necessary to define criteria for selecting (or rejecting) terms, and to determine their level in the ontology. The major criterion has been the functionality: in other words, the ontology describes tools and systems and not a particular application domain. Nodes in the tree correspond to an attributed functionality: a new node is created when it is not possible to express the properties of a given tool with reference to the already existing properties of the concepts in the ontology. In this way, each new class of tools is the source of a node.

For what concerns the format chosen to represent the ontology, the experimental language LIFE[12], appeared suitable, because it enables the representation of partially-informed concept hierarchies in a high-level logical format. The RDF format has not been chosen, because the CLASS ontology relies on a unique semantic heritage, that is links of the hypernym-hyponym type. Concepts and attributes cannot be located at the same level. For what concerns the TopicMaps formalism, it is designed to represent knowledge semantically and to exploit it with the help of networks, which is not the purpose of this ontology.

The CLASS ontology also highlights the operations performed by the basic components and the lingware modules. In fact, CLASS also aims at facilitating the reusability of resources and at proposing common evaluation procedures (in the spirit of the recommendations of the European ELSE[13] project). It is indeed possible to implement crossed or "plaited" evaluation (to reuse the expression originally proposed by Richard Crouch, Robert Gaizauskas and Klaus Netter[14]), on tools of different

Excerpts from the CLASS Ontology

```
linguistic_nlp_product
        active_linguistic_product
                basic_component
                        analyser (...)
                                tagger
                                        lexicographic_analyser
                                                morpho_syntactic_tagger | marker
                                                semantic tagger | semantic_marker
                                morphologic_analyser (...)
                        generator
                                morphologic_generator (...)
                                aided_text_production_tool
                lingware | linguistic_software | language_processing_software (...)
                        automatic_summarisation_tool (...)
                        translator
                integrated_application (...)
                        natural_language_processing (...)
        passive_linguistic_product (...)
                resource
                        lexical_resource (...)
                        corpus (...)
```

origins, as long as the data yielded at the output of one system are identical to those needed at the input of the next one. A same evaluation benchmark can then be put into place for both systems: it is precisely the point of interest of this ontology, and hence the importance of expressing the evaluation criteria within this ontology.
Finally, the CLASS ontology can be viewed as a finite model. It corresponds to a given point of view, at a given time, but it is bound to evolve, at least for reaching a pedagogical dimension. Indeed, it would be of great use to complement it with a brief description for each term, for instance in order to fulfil criteria for other needs and other projects. It has not been achieved yet, because of a lack of time, and a number of descriptors are still questionable. For instance, to define the term "translator", should one speak about "translation", "aided-translation" or "generation"? These questions are still pending today, but in the future, external contributions to this ontology may indicate that it is being reused by others, which would constitute a further validation step by the scientific community.

[1] A tool implementing linguistic functionalities
[2] Bourigault Didier, Charlet Jean. *Ontologies et Textes*. IC'2000. http://www.irit.fr/IC2000/ACTES/ontologies-textes.pdf
[3] Obitko marek. *Ontologies, Description and Applications*. Faculty of Electrical Engineering, Czech Technical University (Prague). 2001. ISSN 1213-3000.
[4] Knowledge acquisition for aided design with interaction between agents (ACACIA, Acquisition des Connaissances pour l'Assistance à la Conception par Interaction entre Agents).
[5] ACACIA project, Activity Report. INRIA 2001.
[6] Analysis and modeling system for the validation of Renault vehicles (SAMOVAR, Système d'Analyse et de Modélisation des Validations des Automobiles Renault).
[7] http://www.laria.u-picardie.fr/EQUIPE/ic/demo/onto-pme.html
[8] Content-based access to the web http://www.ladseb.pd.cnr.it/infor/Ontology/Papers/OntoSeek.pdf
[9] http://www.class-tech.org
[10] http://www.owil.org/fr_repertoire.htm
[11] http://www.fas.montreal;ca/ling/lhomme/cgi-bin/traductique.pdf
[12] Aït-Kaci Hassan, Dumant Bruno, Meyer Richard, Podelski Andreas, van Roy Peter. The Wild LIFE Handbook. DEC Corporation, Paris Research Laboratoty. March 1994.
[13] http://www.limsi.fr/TLP/ELSE
[14] Crouch Richard, Gaizauskas Robert, Netter Klaus et al. Interim Report of the Study Group on Assessment and Evaluation, pp35-38. DFKI. April 1995.

Laurie de Temmerman, student
DESS (professional post-graduate degree) in multilingual engineering
INaLCO (Institut national des Langues et Civilisations orientales)
2, rue de Lille
75007 Paris (France)
Email: lauriedet@yahoo.fr
Web site: http://www.crimlangueso.asso.fr

Jean-Paul Taravella
Senior Consultant
SchlumbergerSema, Consulting
Knowledge Management
9, bd. Charles De Gaulle
92126 Montrouge Cedex (France)
Tel.: +33 (0)1 40 92 39 03
Fax: +33 (0)1 40 92 48 23
Email: jptaravella@slb.com
Web site: http://www.slb.com

A Synergetic Software - LexSyn

Sandra Berrebi

The terms used in technical and commercial communication material such as brochures, Web sites, user manuals, etc. are created by companies and evolve over time. The creation of a consistent database of quality terminology is a strategic question for any company or organisation, which ever sector is concerned.
According to Nemesia, a famous French company in the field of knowledge management: "Terminology is the basis of knowledge management as building a terminology is building the *smallest memory* of the company".
In fact, quality is often neglected and consistency is not always guaranteed. How can we avoid the improper usage of terms in a technical form? How can we avoid translating terms in several different ways in a given technical manual, because of a lack of communication between translators? How can we construct a lexicon or a set of terms (using common reflection and intellectual exchange) to improve quality?
The management of terminological knowledge is a psycho-organisational issue. In a document dedicated to terminology management in a large international organisation, the concluding sentence written by one of the specialists summarizes the question:
"Indeed, one of the people interviewed characterized the major problem of terminology management (…) as a problem of how to share knowledge".
The issue is the following: how to enable communication, collaboration and expression of all players involved in the company, while preserving the acquired terminological knowledge which has already been validated?
Practically, in a context where oral communication is necessarily limited (even though it is facilitated when the intervening parties work in the same location), how is it possible to ensure communication and collaboration between these parties, as well as coordination of the lexical and terminological knowledge? Meetings also have their limits. Moreover, because of the (legitimate) fear of damaging or polluting the database with erroneous information, the company is generally bound to restrict the rights of editing the terminology to a few users. The other users are only allowed to consult the database whereas they should be able to participate on-line on its elaboration and evolution.
The lack of communication and interaction of the parties creates not only a number of terminological inconsistencies but also a sub-optimal organisation of the work, as several parties often end up executing the same database search. The example of a large terminology database used at the European level is very revealing of this situation:
"The second (worry), already mentioned and very widely expressed, was that the availability of local systems would lead to a proliferation of scattered small terminology bases - (...), the contents of which were unknown to his colleagues and therefore not easily shareable, with the result

that much work was done several times over and consistency across translators was much harder to achieve (...)".
Babeling, a newly created French company specialised in language and knowledge engineering is precisely dealing with this difficult issue. *LexSyn*, its leading product, is the first collaborative working tool (synergetic software) for on-line multilingual lexicography and terminology management (for intranet, extranet and Internet). It is designed for the creation and the management of quality consistent terminology, thanks to a real-time communication and collaboration of parties and to a coordination of lexicography and terminology knowledge.
The originality of *LexSyn* is also to enable the expression of all parties, while preserving the acquired terminological knowledge previously validated. As a matter of fact, the existence within *LexSyn* of personalised competence profiles and the use of hierarchical written communication, ensures a real communication and collaboration between parties, allowing everybody's expression, while preserving validated information.
A second advantage of the synergetic software lies in the fact that geographically distant players can, wherever they are located, work simultaneously on the same project: they can consult, create, edit and manage information on-line (intranet, extranet, local network, Internet) and in real-time, for a terminological and/or multilingual dictionary.
This type of system has a two-fold interest: on the one hand, *LexSyn* avoids the costly step of terminology consistency check-up and correction. As a matter of fact, the correction of a terminological error would cost 20 times more during the publication or distribution stage than during the editing stage. On the second hand, the communication between parties improves work quality and avoids task redundancy.
LexSyn optimises work and reduces costs.

Sandra Berrebi
Director R& D Department
Babeling
85 ter, av. Foch
94100 Saint Maur des Fossés (France)
Tel.: +33 (0)1 42 83 54 91
Email: s.berrebi@babeling.com
Web site: http://www.babeling.com

New Resources

ELRA-S0131 British English SpeechDat-Car

The British English SpeechDat-Car comprises the recordings of 300 British English speakers from 6 different regions (170 males, 130 females), recorded over the GSM telephone network, in a car. The SpeechDat-Car database has been collected by Vocalis. This database is partitioned into 115 CDs (DVDs are also available). The speech databases made within the SpeechDat-Car project were validated by SPEX, the Netherlands, to assess their compliance with the SpeechDat-Car format and content specifications.
The speech data files are in two formats. Four of the microphones were recorded on the computer in the boot of the car. The speech data are stored as sequences of 16 kHz, 16 bit and uncompressed. The fifth microphone was connected to the cell phone, and was recorded on a remote machine. The data are stored as sequences of 8 kHz 8 bit A-law. Each signal file is accompanied by an ASCII SAM label file which contains the relevant descriptive information. Each speaker uttered around 120 read and spontaneous items. The following age distribution has been obtained: 119 speakers are between 16 and 30, 109 speakers are between 31 and 45, 57 speakers are between 46 and 60, and 15 speakers are over 60. A pronunciation lexicon with a phonemic transcription in SAMPA is also included.

	ELRA members	Non-members
For research use	90,000 Euro	120,000 Euro
For commercial use	90,000 Euro	120,000 Euro

ELRA-S0132 Danish SpeechDat-Car

The Danish SpeechDat-Car comprises the recordings of 300 Danish speakers from 5 different regions (162 males, 138 females), recorded over the GSM telephone network, and in a car. The SpeechDat-Car database has been collected at Center for Personkommunikation (University of Aalborg), in collaboration with Sonofon. This database is partitioned into 15 DVDs (53 GB), plus 1 CD-ROM for e.g. non-signal files and documentation. The speech databases made within the SpeechDat-Car project were validated by SPEX, the Netherlands, to assess their compliance with the SpeechDat-Car format and content specifications.
The speech data files are in two formats. Four of the microphones were recorded on the computer in the boot of the car. The speech data are stored as sequences of 16 kHz, 16 bit and uncompressed. The fifth microphone was connected to the cell phone, and was recorded on a remote machine, with compressed data stored as sequences of 8 bit A-law 8.kHz. Each signal file is accompanied by an ASCII SAM label file which contains the relevant descriptive information.
Each speaker uttered around 120 read and spontaneous items.
The following age distribution has been obtained: 84 speakers are between 18 and 30, 99 speakers are between 31 and 45, 98 speakers are between 46 and 60, and 19 speakers are over 60. A pronunciation lexicon with a phonemic transcription in SAMPA is also included.

	ELRA members	Non-members
Whole database:		
For research use	40,000 Euro	48,000 Euro
For commercial use	50,000 Euro	60,000 Euro
GSM recordings only:		
For research use	20,000 Euro	24,000 Euro
For commercial use	25,000 Euro	30,000 Euro
In-car recordings only:		
For research use	28,000 Euro	33,000 Euro
For commercial use	35,000 Euro	42,000 Euro

ELRA-S0133 Finnish SpeechDat-Car

The Finnish SpeechDat-Car comprises the recordings of 302 Finnish speakers (151 males, 151 females) from 3 major dialectal areas (with 13 sub-areas), recorded over the GSM telephone network, and in a car. The SpeechDat-Car database has been collected at Nokia Research Centre, in collaboration with Digital Meida Institute. This database is partitioned into 142 CDs. The speech databases made within the SpeechDat-Car project were validated by SPEX, the Netherlands, to assess their compliance with the SpeechDat-Car format and content specifications.
The speech data files are in two formats. Four of the microphones were recorded on the computer in the boot of the car. The speech data are stored as sequences of 16 kHz, 16 bit and uncompressed. The fifth microphone was connected to the cell phone, and was recorded on a remote machine, with compressed data stored as sequences of 8 bit A-law 8.kHz. Each signal file is accompanied by an ASCII SAM label file which contains the relevant descriptive information.
Each speaker uttered around 120 read and spontaneous items. The following age distribution has been obtained: 138 speakers are between 16 and 30, 89 speakers are between 31 and 45, and 75 speakers are between 46 and 60. No speaker are over 60. A pronunciation lexicon with a phonemic transcription in SAMPA is also included.

	ELRA members	Non-members
For research use	10,000 Euro	10,000 Euro
For commercial use	40,000 Euro	60,000 Euro

ELRA-S0135 French SpeechDat-Car

The French SpeechDat-Car comprises the recordings of 313 French speakers (158 males, 155 females) from 6 different regions, recorded over the GSM telephone network and in a car. The SpeechDat-Car database has been produced by L&H and Scansoft Belgium. This database is partitioned into 16 DVDs. The speech databases made within the SpeechDat-Car project were validated by SPEX, the Netherlands, to assess their compliance with the SpeechDat-Car format and content specifications.
The speech data files are in two formats. Four of the microphones were recorded on the computer in the boot of the car. The speech data are stored as sequences of 16 kHz, 16 bit and uncompressed. The fifth microphone was connected to the GSM phone, and was recorded on a remote machine, with compressed data stored as sequences of 8 bit A-law 8.kHz. Each signal file is accompanied by an ASCII SAM label file which contains the relevant descriptive information.
Each speaker uttered around 120 read and spontaneous items. The following age distribution has been obtained: 208 speakers are between 16 and 30, 78 speakers are between 31 and 45, 25 speakers are between 46 and 60, and 2 speakers are over 60. A pronunciation lexicon with a phonemic transcription in SAMPA is also included.

	ELRA members	Non-members
For research use	104,000 Euro	182,000 Euro
For commercial use	130,000 Euro	182,000 Euro

ELRA-S0136 SmartKom Mutimodal Corpora - SKP 1.0

This multimodal corpus produced by BAS between 1999 and 2003 within the SmartKom project comprises the recordings of 45 speakers, recorded in public places (cinema and restaurant) in the technical setup SmartKom Public, which is comparable to a traditional public phone booth, equipped with additional intelligent communication devices. The recorded modalities are the audio in 10 channels, the video of the face, the video of the upper body from the left, the infrared video of the display area (to capture the 2D gestures) as input to the SIVIT device (Siemens gesture recognizer), the video of the GUI output, the coordinates of graphic tableau (when the pen was used), the coordinates of the SIVIT device (when fingers/hands were used). Annotations files are also included (transliteration, 3D gesture, user states in three modalities, turn segmentation), as well as documentation files. The corpus is structured into volumes that can be selected and purchased separately, and is available either on DVD or an IDE (hard disk).

	ELRA members	Non-members
SKP Total on DVD:		
For research use	9,200 Euro	18,400 Euro
For commercial use	9,200 Euro	18,400 Euro
SKP Total on IDE:		
For research use	7,700 Euro	15,400 Euro
For commercial use	7,700 Euro	15,400 Euro
SKP Single Volume:		
For research use	127.82 Euro	255.65 Euro
For commercial use	127.82 Euro	255.65 Euro

ELRA-S0137 TAXI - Multilingual Telephone Dialog Database

TAXI was produced by BAS, in collaboration with the German research centre for artificial intelligence, DFKI. This speech database contains recordings which consist of dialogues, 94 on the whole (spontaneous speech), between a German speaking cab dispatcher and his clients, who always answered in English. To prevent overlap and to allow automatic segmentation by the recording server, each party pressed a button on his phone to signal the other one that his turn was over. They were recorded over the telephone network. Each dialogue part is translated into the other language. Noise markers are included in the transcripts (not in the translations).The database was annotated following the SpeechDat specifications, and validated to assess its compliance with the SpeechDat format. The files are stored as BAS Partitur Format files.

	ELRA members	Non-members
For research use	127.82 Euro	255.65 Euro
For commercial use	383.03 Euro	511.29 Euro

ELRA-S0134 Concise Oxford Dictionary - Audio Files

This 'acoustic dictionary' contains 60,000 soundfiles from the 9th edition of the Concise Oxford Dictionary. The recordings were made by actors in a studio. It features recordings with British-English pronunciation, with an accurate coverage of different homographs, variant forms and inflections. Full information on parts of speech and subsenses is covered, and the soundfiles are clearly linked to the related phonetic information. The format in use is 22kHz 16-bit WAV.

	ELRA members	Non-members
For research use	4,900 Euro	7,000 Euro
For commercial use	ELRA	ELRA

ELRA-S0138 Cantonese SpeechDat-like MDB-2000

The Cantonese SpeechDat-like MDB-2000 database comprises 2000 Cantonese speakers (996 male, 1004 female) recorded over the mobile telephone network in China and Hong Kong. The database was produced by Siemens A.G., Munich, Germany, and Jiao-Tong University, Shanghai, People's Republic of China. The MDB-2000 database is partitioned into 11 CDs in ISO 9660 format. This database was validated by SPEX, the Netherlands, to assess their compliance with the SpeechDat(II) format and content specifications.
Speech samples are stored as sequences of 8-bit 8 kHz A-law. Each prompted utterance is stored in a separate file. Each signal file is accompanied by an ASCII SAM label file which contains the relevant descriptive information.
Each speaker uttered around 45 read and spontaneous items.
The following age distribution has been obtained: 74 speakers are below 16 years old, 953 speakers are between 16 and 30, 636 speakers are between 31 and 45, 328 speakers are between 46 and 60, 9 speakers are over 60. A pronunciation lexicon with a phonemic transcription in SAMPA is also included.

	ELRA members	Non-members
For research use	52,000 Euro	65,000 Euro
For commercial use	65,000 Euro	85,000 Euro

ELRA-S0139 Flemish/Dutch SpeechDat-Car

The Flemish/Dutch SpeechDat-Car comprises the recordings of 302 speakers (154 males, 148 females) from Flanders and The Netherlands. They were recorded over the mobile telephone network and in a car. The database contains recordings both in Flemish and in Dutch as spoken in Flanders (about 1/3 of the speakers), as well as recordings in Dutch as spoken in The Netherlands (about 2/3 of the speakers).
The SpeechDat-Car database has been produced by Lernout & Hauspie Speechproducts. This database is partitioned into 162 CDs. The speech databases made within the SpeechDat-Car project were validated by SPEX, the Netherlands, to assess their compliance with the SpeechDat-Car format and content specifications.
The speech data files are in two formats. Four of the microphones were recorded on the computer in the boot of the car. The speech data are stored as sequences of 16 kHz, 16 bit and uncompressed. The fifth microphone was connected to the GSM phone, and was recorded on a remote machine, with compressed data stored as sequences of 8 bit A-law 8.kHz. Each signal file is accompanied by an ASCII SAM label file which contains the relevant descriptive information. Each speaker uttered 120 read and spontaneous items. The following age distribution has been obtained: 107 speakers are between 16 and 30, 127 speakers are between 31 and 45, 66 speakers are between 46 and 60, and 2 speakers are over 60.
A pronunciation lexicon with a phonemic transcription in SAMPA is also included.

	ELRA members	Non-members
For research use	104,000 Euro	182,000 Euro
For commercial use	130,000 Euro	182,000 Euro

ELRA-S0140 Spanish SpeechDat-Car

The Spanish SpeechDat-Car database comprises the recordings of 306 Spanish speakers from 4 different regions (156 males, 150 females), recorded over the Spanish GSM telephone network, in a car. This database has been collected by the Department of Signal Theory and Communications of the Universidad Politecnica de Catalunya (UPC, Spain), with the collaboration of SEAT and Volkswagen. This database is partitioned into 89 CDs (DVDs are also available). The speech databases made within the SpeechDat-Car project were validated by SPEX, the Netherlands, to assess their compliance with the SpeechDat-Car format and content specifications.
The speech data files are in two formats. Four of the 5 microphones were recorded on the computer in the boot of the car. The speech data are stored as sequences of 16 kHz, 16 bit and uncompressed. The fifth microphone was connected to the cell phone, and was recorded on a remote machine. The data are stored as sequences of 8 kHz 8 bit A-law. Each signal file is accompanied by an ASCII SAM label file which contains the relevant descriptive information.
Each speaker uttered around 120 read and spontaneous items. The following age distribution has been obtained: 160 speakers are between 18 and 30, 80 speakers are between 31 and 45, 65 speakers are between 46 and 60, and 1speaker is over 60. A pronunciation lexicon with a phonemic transcription in SAMPA is also included.

	ELRA members	Non-members
For research use	60,000 Euro	80,000 Euro
For commercial use	60,000 Euro	80,000 Euro

ELRA-S0141 SALA Spanish Venezuelan Database

The SALA Spanish Venezuelan Database comprises 1000 Venezuelan speakers (504 males, 496 females) recorded over the Venezuelan fixed telephone network. The corpus design and the collection of the speech data were performed at the Universidad de los Andes (Venezuela), and the transcriptions and formatting at the Universidad Politècnica de Catalunya (UPC). This database is partitioned into 5 CD-ROMs The speech databases made within the SALA project were validated by SPEX, the Netherlands, to assess their compliance with the SALA format and content specifications.
The speech files are stored as sequences of 8-bit, 8kHz mu-law speech files and are not compressed, according to the specifications of SALA. Each prompt utterance is stored within a separate file and has an accompanying ASCII SAM label file.
Each speaker uttered Each speaker uttered around 50 read and spontaneous items. The following age distribution has been obtained: 7 speakers are under 16 years old, 476 speakers are between 16 and 30, 330 speakers are between 31 and 45, 177 speakers are between 46 and 60, and 10 speakers are over 60.

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

ELRA-S0142 Austrian SpeechDat(AT) FDB-1000 Database

The SpeechDat(AT) FDB-1000 database comprises the recordings of 1000 Austrian speakers (544 males, 456 females) recorded over the Austrian fixed telephone network. The speech collection was carried out by the Telecommunications Research Center Vienna (FTW, Forschungszentrum Telekommunikation Wien) in co-operation with Alcatel, Kapsch AG, Nokia, Connect Austria, Philips and Siemens AG. The database is partitioned into 5 CD-ROMs, in ISO 9660 format.
Speech samples are stored as sequences of 8-bit 8 kHz A-law, uncompressed. Each prompted utterance is stored in a separate file, and each signal file is accompanied by an ASCII SAM label file which contains the relevant descriptive information.
This speech database was validated by SPEX (the Netherlands) to assess its compliance with the SpeechDat format and content specifications.
Each speaker uttered around 60 read and spontaneous items.
The following age distribution has been obtained: 15 speakers are under 16, 444 are between 16 and 30, 328 are between 31 and 45, 184 are between 46 and 60, and 29 speakers are over 60.
A pronunciation lexicon with a phonemic transcription in SAMPA is also included. A package including the Austrian SpeechDat(AT) MDB-1000 database is available at a discount.

	ELRA members	Non-members
For research use	16,000 Euro	20,000 Euro
For commercial use	20,000 Euro	25,000 Euro

ELRA-S0143 Austrian SpeechDat(AT) MDB-1000 Database

The Austrian SpeechDat(AT) MDB-1000 database comprises the recordings of 1000 Austrian speakers (543 males, 457 females) recorded over the Austrian mobile telephone network. The speech collection was carried out by the Telecommunications Research Centre Vienna (FTW, Forschungszentrum Telekommunikation Wien) in co-operation with Alcatel, Kapsch AG, Nokia, Connect Austria, Philips and Siemens AG. The database is partitioned into 5 CD-ROMs, in ISO 9660 format.
Speech samples are stored as sequences of 8-bit 8 kHz A-law, uncompressed. Each prompted utterance is stored in a separate file, and each signal file is accompanied by an ASCII SAM label file which contains the relevant descriptive information.
This speech database, was validated by SPEX (the Netherlands) to assess its compliance with the SpeechDat format and content specifications. The read and spontaneous items uttered by the 1000 speakers who have been recorded over the mobile telephone network are the same as the items mentioned in the description of the speech collection over the fixed telephone network (see resource ELRA S0142, description above). A package including the Austrian SpeechDat(AT) FDB-1000 database is available at a discount.

	ELRA members	Non-members
For research use	24,000 Euro	30,000 Euro
For commercial use	30,000 Euro	35,000 Euro

ELRA-T0367 Telecommunications Dictionary

This specialised dictionary comprises 62,200 pairs of English-French expressions. Sub-domains covered include: PABX (Private Automatic Branch eXchange), public telephone exchange, switch, microwave radio system, satellites, multiplexer, signalling, printed circuit, telephony, etc. Information linked to the entry which are given in both languages include: synonyms, abbreviations, and definitions or possible applications in case of polysemy.
The resource is available in the ASCII format, on a floppy disk medium.

	ELRA members	Non-members
For research use	0.38 Euro/entry	0.48 Euro/entry
For commercial use	0.48 Euro/entry	0.60 Euro/entry

ELDA is happy to announce a collaboration with Oxford University Press for the distribution of a selection of language resources, for which you will find the descriptions below.

ELRA-L0045 New Oxford Dictionary of English (NODE), 2nd Edition

This is Oxford University Press's most comprehensive single-volume dictionary, with 170,000 entries covering all varieties of English world-wide. The NODE data set constitutes a fully integrated range of formal data types suitable for language engineering and NLP applications: It is available in XML or SGML.

- Source dictionary data. The NODE data set includes all the information present in the New Oxford Dictionary of English itself, such as definition text, example sentences, grammatical indicators, and encyclopaedic material.
- Morphological data. Each NODE lemma (both headwords and subentries) has a full listing of all possible syntactic forms (e.g. plurals for nouns, inflections for verbs, comparatives and superlatives for adjectives), tagged to show their syntactic relationships. Each form has an IPA pronunciation. Full morphological data is also given for spelling variants (e.g. typical American variants), and a system of links enables straightforward correlation of variant forms to standard forms. The data set thus provides robust support for all look-up routines, and is equally viable for applications dealing with American and British English.
- Phrases and idioms. The NODE data set provides a rich and flexible codification of over 10,000 phrasal verbs and other multi-word phrases. It features comprehensive lexical resources enabling applications to identify a phrase not only in the form listed in the dictionary but also in a range of real-world variations, including alternative wording, variable syntactic patterns, inflected verbs, optional determiners, etc.
- Subject classification. Using a categorisation scheme of 200 key domains, over 80,000 words and senses have been associated with particular subject areas, from aeronautics to zoology. As well as facilitating the extraction of subject-specific sub-lexicons, this also provides an extensive resource for document categorisation and information retrieval.
- Semantic relationships. The relationships between every noun and noun sense in the dictionary are being codified using an extensive semantic taxonomy on the model of the Princeton WordNet project. (Mapping to WordNet 1.7 is supported.) This structure allows elements of the basic lexical database to function as a formal knowledge database, enabling functionality such as sense disambiguation and logical inference.
- Derived from the detailed and authoritative corpus-based research of Oxford University Press's lexicographic team, the NODE data set is a powerful asset for any task dealing with real-world contemporary English usage. By integrating a number of different data types into a single structure, it creates a coherent resource which can be queried along numerous axes, allowing open-ended exploitation by many kinds of language-related applications.

	ELRA members	Non-members
For research use	6,125 Euro	8,750 Euro
For commercial use	ELRA	ELRA

ELRA-L0046 NODE + DIMAP

The DIMAP version of NODE (first edition) is a machine-tractable version of the machine-readable dictionary files in the DIMAP dictionary maintenance programs, adding syntactic and semantic information in the conversion. In addition, DIMAP provides several mechanisms that will allow research into representational formalisms and explorations of the use of these representations in extending the lexical database and in processing text for information extraction, text summarisation, discourse analysis and other LE applications. Using functionality to parse the dictionary definitions, DIMAP has further enhanced NODE through the addition of many semantic links, including hyperonyms, synonyms, and other semantic relations, thus making NODE+DIMAP a semantic network of the English language. For more details on the contents of NODE+DIMAP, you can visit:
http://www.clres.com/node-dimap-contents.html

	ELRA members	Non-members
For research use	7,000 Euro	10,000 Euro
For commercial use	ELRA	ELRA

ELRA-L0047 New Oxford Thesaurus of English (NOTE)

The New Oxford Thesaurus of English is a completely new top-of-the-range thesaurus offering more alternative and opposite words than any of its competitors. The synonyms are arranged in order of 'relevance' to the look-up word, starting with an individually tagged core synonym, and followed by labelled groups with limited currency, such as informal, regional, and technical terms. NOTE also contains a corpus-based example for the vast majority of senses - a total of nearly 38,000.
It contains 628,000 alternative words, including 573,000 synonyms, the rest being antonyms, related terms, combining forms, and hyponyms, and is available in SGML.

	ELRA members	Non-members
For research use	4,900 Euro	7,000 Euro
For commercial use	ELRA	ELRA

ELRA-L0048 Oxford Paperback Thesaurus, 2nd edition

The Second Edition of the Oxford Paperback Thesaurus, derived from NOTE, contains approximately 330,000 alternative words. It has corpus-based examples for each sense of polysemous words (29,000 total). The tagging matches that of the New Oxford Dictionary of English: an additional feature of the tagging is that each synonym has its own field. It is available in SGML.

	ELRA members	Non-members
For research use	5,250 Euro	7,500 Euro
For commercial use	ELRA	ELRA

ELRA-M0027 Oxford French Minidictionary

This revised edition offers thoroughly up-to-date vocabulary (over 100,000 words, phrases and translations), along with many helpful features such as frequently-used words in French and English are given usage notes; it also comprises clear warning symbols for slang and informal language. It is available in SGML.

	ELRA members	Non-members
For research use	3,500 Euro	5,000 Euro
For commercial use	ELRA	ELRA

ELRA-M0028 Concise Oxford Duden German Dictionary

Based on the highly acclaimed Oxford-Duden German Dictionary, recommended by academics, professionals, and students across the world, this dictionary provides learners at intermediate level with an authoritative guide to German and English in a concise and easy-access format. It incorporates the recent German spelling reforms, with changes signalled throughout. It contains 150,000 words and phrases, and 240,000 translations. It is available in XML and SGML.

	ELRA members	Non-members
For research use	4,900 Euro	7,000 Euro
For commercial use	ELRA	ELRA

ELRA-M0029 Pocket Oxford Italian Dictionary

This is a mid-sized bilingual dictionary to cover essential terms and vocabulary, available in XML and SGML. It contains 80,000 words and phrases, and 115,000 translations.

	ELRA members	Non-members
For research use	3,500 Euro	5,000 Euro
For commercial use	ELRA	ELRA

ELRA-M0030 Concise Oxford Spanish Dictionary

Based on the highly-acclaimed Oxford Spanish Dictionary, described by John Butt in the TLS as `indispensable for all serious Hispanists', this dictionary provides an authoritative, up-to-date guide to world Spanish. It is the only Concise Spanish dictionary to present the full wealth of Spanish from both sides of the Atlantic, with coverage of 24 varieties of Spanish as it is written and spoken throughout the Spanish-speaking world There are thousands of real, authentic example sentences carefully selected to illustrate the full range of meanings and typical contexts. It contains 170,000 words and phrases, and 240,000 translations, and is available in SGML and XML.

	ELRA members	Non-members
For research use	4,900 Euro	7,000 Euro
For commercial use	ELRA	ELRA

ELRA-M0031 Oxford Business French Dictionary

This dictionary covers the general language of Business across a range of core areas. It contains over 50,000 words and phrases, and is available in SGML.

	ELRA members	Non-members
For research use	4,900 Euro	7,000 Euro
For commercial use	ELRA	ELRA

ELRA-M0032 Oxford Business Spanish Dictionary

This dictionary covers the general language of Business across a range of core areas. It contains over 50,000 words and phrases, and is available in SGML.

	ELRA members	Non-members
For research use	4,900 Euro	7,000 Euro
For commercial use	ELRA	ELRA

ANNOUNCEMENT

Just published [Hermès publisher, IC2 (Information, Command, Communication) series] *** in French***:

Traitement automatique du langage parlé (Spoken Language Processing) - two volumes - edited by Joseph Mariani

We invite you to visit the following web site to get more information: www.Lavoisier.fr

LangTech 2003

The European Forum for Language Technology

LangTech is a dedicated international forum for people and organisations involved in the development, deployment and exploitation of spoken and written language technologies.

Language technology sectors covered within LangTech include:

- *Voice solutions (customer relationship management, software and machinery voice control, etc.)*
- *Knowledge solutions (information and knowledge management, document authoring, e-learning, etc.)*
- *Multilingual solutions (localisation, translation memories, crosslingual information retrieval, etc.)*

LangTech 2002 was held in Berlin (Germany) on September 26th & 27th, 2002.

This was the first edition of LangTech, a successful and fruitful event.

Over 330 industry participants from 30 European and non-European nations attended the conference and visited the exhibition.

The programme featured keynotes from major industry players and near to 70 international speakers, who made presentations related to new technologies, products, applications and R&D projects in the areas of voice, multilinguality and knowledge management.

LangTech 2002 also gave the opportunity to 23 SMEs and start-ups to introduce themselves and present their activity to attract funding from venture capitals. 19 companies also took advantage of that occasion to participate in the exhibition, which was visited by the participants during the 2-day event.

LangTech 2003 will be organised next Winter in Paris (France):

24th & 25th November 2003

Méridien Montparnasse Hotel
19, rue du Commandant Mouchotte 75014 Paris

To get more information, please contact us at the following email address:

langtech2003@elda.fr

ELRA/ELDA
55/57 rue Brillat Savarin
75013 Paris (France)
Tel.: +33 (0)1 43 13 33 33
Fax: +33 (0)1 43 13 33 30