The ELRA Newsletter

July - September 2002

Vol.7 n.3

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:

Philip Edmonds
Christian Galinski
Richard Walter

ISSN: 1026-8200

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

Dear Colleagues,

This is the 3rd issue of the ELRA newsletter for the year 2002. The previous one was a special issue devoted to the LREC 2002 conference. The proceedings of the main conference, as well as the workshops' proceedings, are now available from the ELDA offices. To place an order, please download the order form from the LREC web site, www.lrec-conf.org, or contact Valérie Raymond, <raymond@elda.fr>.

Late September, the LangTech 2002 conference took place in Berlin (Germany). Over 300 participants attended this first European industrial forum for language technology, where key players in the field of HLT, notably the areas of voice, multilinguality and knowledge management, gave oral presentations on topics like in-vehicle spoken dialog, large-scale translation memory and machine translation systems, intranet search in cross-lingual environments, etc.; along with the presentations given in the three parallel sessions, LangTech focused on the commercial aspects of language technologies, with the demonstration of products and applications, the introduction of start-up companies for venture capital purposes, and an exhibition, which gathered 20 companies. The next edition, LangTech 2003, for which the dates and the location are still to be confirmed, will be held in Paris (France). You may contact us, langtech2003@elda.fr, if you are interested in participating in LangTech 2003, if you would like to get information or if you would like to express suggestions.

During these summer months, ELRA & ELDA performed a number of tasks for some of the European, international and French projects we are involved in. In the framework of the *Euromap Language Technologies* project, a seminar was organised in Paris at the beginning of July, with over 100 participants. Major objectives consisted of promoting HLT among French players, and of drawing a map of language engineering in France and in Europe. Representatives from the French Ministry of Research presented TechnoLangue, a national programme on language technologies supported by the French Ministries in charge of Industry, Research and Culture, which are co-funding the programme. TechnoLangue consists of 4 actions: development and reinforcement of language resources, creation of an infrastructure for the evaluation of language technologies, better accessibility to norms and standards, and setting up of an intelligence watch network in HLT. A representative from the European Commission (DG InfoSo), Brian Macklin, presented the 6th Framework Programme (FP6). You can find the presentations and more information on the Euromap LT web pages on the ELDA web site, at the following address: http://www.elda.fr/fr/proj/euromap/seminar.html. General information on Euromap Language Technologies can be found at http://www.hltcentral.org/. The work is in progress for the EC-FP5 *OrienTel* project. The partners of the consortium met to finalise the specifications both for the corpus and for the transcriptions to be used for the Arabic speech data. A set of 21 databases will be collected to unable the design and development of multilingual interactive communication services for Mediterranean and Middle East countries, ranging from Morocco in the West to the Gulf states in the East, including Turkey and Cyprus. The languages covered include e.g. standard and colloquial Arabic, French, English, Turkish, Hebrew, Greek, etc. Within the EC-FP5 *CLEF* (Cross-Language Evaluation Forum) project, the 2002 workshop was organised late September, where the partners met to discuss the results of the 2002 campaign and the planning of the next evaluation campaign. In addition to the evaluation of cross-language retrieval systems, the next campaign should include exploratory tasks such as e.g. the evaluation of question-answering systems or image retrieval. Within *TechnoLangue*, ELDA plans to set up a long-lasting HLT evaluation infrastructure. This action is based on an initiative launched by ELRA two years ago to establish such an evaluation infrastructure for Europe, in order to fulfil the requirements of technology developers and integrators regarding the field of HLT evaluation. To strengthen its involvement in the evaluation activity, ELDA is currently seeking to complete its evaluation team, to take care of every aspect of the evaluation activity, in the framework of European and international projects. A job announcement was disseminated recently (see page 13).

As far as ELRA's internal activities are concerned, the network of validation centres is under completion, with the set up of another node, a Validation Centre for Written Language Resources (VC_WLR). Its tasks and missions will be similar to those of the Validation Centre for Spoken Language Resources (VC_SLR). The validation of the resources available in our catalogue is an issue to which we devote much effort, in order to offer resources of good quality, thus ensuring that our partners have all the information needed to assess how the data may fulfil their needs. Already a number of Quick Quality Check reports drawn by our validation centre SPEX (Speech EXpertise centre) are available for some of the spoken resources we distribute. In the near future, we intend to offer that kind of quality check for all the resources, both written and spoken.

This issue of the ELRA newsletter presents 3 articles. These 3 articles deal with different aspects of the processing of written language resources: the first one, written by Christian Galinski, from InfoTerm, the international information centre for terminology and member of the ISO/TC37 committee, explores the information and knowledge society with regard to the management and standardisation of language resources. Richard Walter, who works for CNRS, the French Research council, shows how they managed to create a large electronic corpus of human sciences based on the French magazine Hermes, from its original digitalisation to its final formatting, to get an exploitable version, now available via ELDA. The third article, from Philip Edmonds from Sharp Laboratories in Oxford provides an overview of the Senseval 2 evaluation campaign and word sense disambiguation.

You will also find in this newsletter a page dedicated to the SCALLA project, also known as SCiLaHLT for Sharing Capability in Localisation and Human Language Technologies. A few pictures from LREC 2002 and LangTech 2002 are also included.

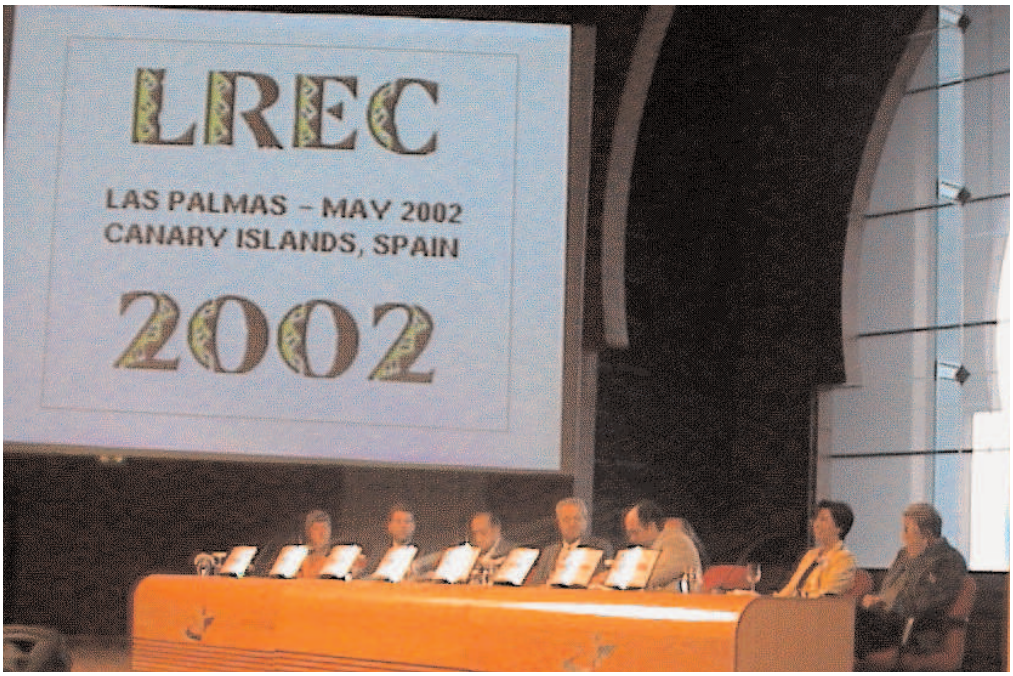
Last but not least, the last section presents the language resources catalogued during this last quarter: These new resources cover an Asian language which was not yet represented in our catalogue, the Korean language. Three sets of speech databases in Korean, as well as 5 written language resources, either monolingual or multilingual (a Korean lexicon, two English-Korean terminology databases in computer science and biology, a Korean annotated corpus, and a multilingual parallel corpus in Korean, Chinese and English) have been added.

Enjoy your reading, and please do not hesitate to contact us for any comments and suggestions to improve the ELRA newsletter.

Joseph Mariani, President — Khalid Choukri, CEO

LREC 2002, a few pictures

No photograph was included in the previous issue of the ELRA newsletter, yet a special issue dedicated to the LREC 2002 conference, which took place in Las Palmas last Spring 2002.
So to make up for lost time, you will find here two illustrations of this successful event.



Opening Ceremony, at the Alfredo Kraus auditorium, from left to right - Bente Maegaard , Daniel Tapias, Angel Martin Municio, Antonio Zampolli, Joseph Mariani, Khalid Choukri, Nicoletta Calzolari, and Harald Höge.

Gala Diner, at the Santa Catalina hotel, from left to right - Daniel Tapias, Bente Maegaard, George Doddington, Joseph Mariani, Antonio Zampolli.

LangTech 2002, Overview

LangTech 2002, the first European forum for language technology, took place in Berlin on 26th and 27th September 2002.

LangTech 2002 was dedicated to the technological and commercial aspects of language technologies in developement in Europe, or existing ones.

Over 300 European and non-European participants from 30 nations gathered for this first edition.

Two keynote speakers were invited:

- Bill Dolan, Head of Natural Language Processing Unit, Microsoft Corporation (USA), whose keynote speech dealt with "Language Technology in Consumer Software".

- Professor Wolfgang Wahlster, DFKI GmbH (Research institute for artificial intelligence, Germany), who presented the "Language Technologies for the Mobile Internet Era".

The participants could attend the presentations made by a number of key players involved in the various fields of Human Language Technologies (HLT), both for research or for industry, and especially in the areas of voice, multilinguality, and knowledge management. Speakers include e.g. Francis Charpentier, from Telisma (France), with a presentation entitled "The Contribution of Speech Technologies to the Next Generations of Telecom Services", Matthias Heyn, from Trados (Belgium), with "The Value of Language Technologies in Companies", Paul Heisterkamp, from Daimler-Chrysler (Germany), with "In-vehicle Spoken Dialog: Safety and functionality", professor Jun-ichi Tsujii, from the department of Information Science (University of Tokyo, Japan), with a presentation about "Machine Translation and Multilingual Systems in Japan and Asia".

Along with the three parallel sessions, a revolutionary aspect of LangTech 2002 consisted of SME presentations: 23 companies and start-ups, from 12 European and non-European countries were given the opportunity to introduce themselves and tell about their activity, to attract funding from venture capitals.

Last but not least, another feature which deserves to be mentioned is the exhibition which was organised on this occasion. 20 companies could take advantage of LangTech 2002, to present and promote their applications, products, services and/or research prototypes.

To learn more about LangTech 2002, and appreciate how succcessful, and fruitful this first edition has proven to be, please visit **http://www.lang-tech.org**

Next year, *LangTech 2003* will be organised in *Paris* (France). Please contact us to get more information about LangTech 2003, **langtech2003@elda.fr**

Bill Dolan (Microsoft Corporation, USA)

From left to right - Bente Maegaard (Center for Sprogteknologi, Denmark), Bill Dolan (Microsoft Corporation, USA), Hans Uskoreit (DFKI, Germany), Uwe Thomas (State Secretary of the Federal Ministry of Education and Research, Germany)

Uwe Thomas (State Secretary of the Federal Ministry of Education and Research, Germany) and Joseph Mariani (ELRA president, Direction de la Technologie, Ministère Délégué à la Recherche et aux nouvelles technologies)

Prof. Wolgang Wahlster (DFKI,Germany)

Exhibition - Aculab's booth (UK)

Natural Language - Language Resources - Semantic Web

Christian Galinski

Recently more and more aspects of the 'economics of language' (viz. primarily the costs of the use of language in specialized/professional communication) are identified. As communication consumes time or transaction efforts in some way or other, costs are incurred continuously. Some are not yet measurable, other have become measurable. This applies to:

- 'natural' inter-personal communication;
  - whether in oral form or in written form;
  - whether in general purpose language (GPL) or in special purpose language (SPL);
- man-machine communication;
- communication in language between computers.

Of course, the objective is not to avoid communication in view of these costs, but to render communication more efficient and effective at places, in environments, at times, where and when it is necessary or useful. Here methodology unification/standardisation/harmonisation provides the most important clues for cost reduction, and at the same time for the improved quality of communication.

This refers in particular to the unification/standardisation/harmonisation of methods concerning language resources (LRs) for the sake of content management, and may in some cases also refer to the data as well as data structures themselves. During the last couple of years the Technical Committee ISO/TC 37 "Terminology and other language resources" of the International Organization for Standardization (ISO) has opened its scope towards language resources in general. This was due among others to the following considerations:

- terminology is widely - especially in speech and text - embedded in or combined with LRs,
- new information and communication technology (ICT) developments - especially mobile content, e-business, mobile commerce, etc. - increasingly require the integration or combination of all kinds of content (incl. LRs),
- LRs (including terminology), therefore, increasingly have to be treated as multilingual, multimedia and multimodal from the outset.

Content

Everything which is representing information or knowledge for whatever purpose is content. At present the creation of those kinds of content, which are based on LRs, is still too slow, too expensive, mostly not good enough and rarely with a guarantee for its correctness. By using the Internet more effectively - e.g. by using it for net-based distributed co-operative content creation with new methods of content management, by involving many more experts and even users as potential creators of content - the cost of content creation can be decreased dramatically, while at the same time improving considerably the quality of the content thus created. ISO/TC 37 is contributing to this development by preparing standards and other documents with rules as well as guidelines for:

- harmonised metadata
- unified principles and methods for data modelling
- standardised meta-models

The Semantic Web

In a letter to "Business Week" (April 8, 2002) Tim Berners-Lee (MIT, the father of the "semantic web" conception) denies that the World Wide Web will be replaced by the Semantic Web, with the following arguments:

"The WWW contains the documents intended for human consumption, and those intended for machine processing. The Semantic Web will enhance the latter. The Semantic Web will not understand human language ... The Semantic Web is about machine languages: well-defined, mathematical, boring, but processable. Data, not poetry."

thus indicating that he is widely misunderstood or misinterpreted.

These remarks also point in the direction of how language use in the information and knowledge society in general and in future e-business (comprising the whole range of e-commerce, e-procurement, e-content, etc., to m-commerce) will develop: highly harmonised terminology combined with factual data and common language elements need to be provided in a form:

- presumably nearer to human natural language usage in B2C;
- presumably nearer to (Tim Berners-Lee's) machine languages in B2B.

What is new in this connection is that these machine languages will be multilingual like human language use. They will also be multimodal and multimedia from the outset.

Standardisation of LR related aspects

Standardisation as a rule is a highly co-operative endeavour carried out in a very democratic way involving industry experts, public administrators, researchers and consumers. The standardisation of harmonised metadata, unified principles and methods for data modelling, and standardised meta-models, with respect to LRs, will inevitably result in a higher degree of granularity of database design and data modelling at the field level. This probably will also lay the basis for resolving a whole array of existing problems with respect to:

- sources of smallest units of information,
- history of the evolution of individual pieces of information,
- details on whatever kind of usage,
- restrictions on individual applications, etc.,

thus arriving at a higher level of:

- data/information source indication (as a prerequisite for enhanced copyright management),
- automatic or computer-assisted validation (supporting quality management),
- tracing the 'history' of every data (thus coping with diachronic development of content and the intricacies of versioning control),

- data safety and security management,
- monitoring methods for collaborative work (with a view to interactive and dynamic content management and information/knowledge management), etc.

The resulting standards or guidelines mainly aim at improving content re-use and interoperability under a global mark-up, global usability and global design philosophy. The development from an information society into a global knowledge society cannot occur without technical-industrial as well as methodology standards. Parallel to the standardisation efforts, activities are undertaken to establish content infrastructures for content creation and distribution, which is also supporting UNESCO's efforts for the universal availability of knowledge and universal access to information in cyberspace. Combining ICT solutions (some under an open source philosophy) with language and knowledge engineering approaches, as well as with terminological methods would even allow for a symbiosis between the needs of developing communities for advanced methods and tools on the one hand, and the needs of technologically and economically advanced communities for inexpensive knowledge organisation and content creation on the other hand.

The cost of language in the enterprise

Until recently, a concrete method to calculate the cost of 'language', in order to be in a position to argue the usefulness or even the need to invest in 'infrastructural' measures with respect to corporate language in general and in terminology management in particular, was lacking. This usefulness/need to invest in language and knowledge infrastructures does not only concern so-called word workers (such as scientific authors, technical documentalists, technical writers/editors, specialised journalists, specialised translators, localizers, terminologists, etc., who prominently use 'words' in their professional activities based on communication in written form), but to all professionals, who deal with information and knowledge (i.e. any 'knowledge worker') in their work.

The examples for 'catastrophic' consequences of deficient language use abound. But in the eyes of decision makers this 'anecdotic evidence' only creates uneasiness, because these 'negative examples' do not help to find systematic solutions to the underlying problem: how to ensure the quality (especially consistency and coherence) of corporate language and knowledge as part of a 'strategic survival strategy' on the increasingly competitive markets. Beside, they do not show any systemic approach not only with respect to measures to avoid such 'catastrophes' in the future, but also in direction of arriving at a 'measurable' cost-saving effect. Only the latter would turn the negative argument of unavoidable 'effort=investment=cost' into the positive argument of overall 'cost-saving'.

E-business - especially in combination with mobile computing resulting in m-commerce - is probably going to change the organisation and operation of enterprises and their business quite radically in the near future. Enterprises and other organisations/institutions will be forced not only to link hitherto separated systems to each other, but to really 'integrate' all data processing systems of the organisation. Latest at this point the whole degree of variation in language usage within the organisation will become apparent. It is quite clear that this divergence, inconsistency and incoherence not only bears the uncomforting potential for 'catastrophes' due to misunderstandings, but also results in constantly recurring costs in terms of loss of time, etc. The fact that computers will have to talk to each other and understand language between each other via virtual marketplaces in future e-business will aggravate this problem. Therefore, a much higher degree of unambiguity in language usage - and first of all in the terminology used - will be indispensable in the near future.

In order to be able to conceive a calculation method for the cost of language usage in the organisation, it is necessary to analyse language from the point of view of 'language resources', which comprise:

- (marked-up or tagged) text corpora,
- speech corpora,
- grammar models,
- lexicographical data,
- terminological data,

and to identify 'units' occurring in (spoken or written man-man, man-machine and machine-machine) communication which can be put in relation to 'transaction' efforts (consuming time or labour or funds).

This provides a clue for instance to estimate or even calculate the costs of words and terms across all documentation in conjunction with product description in an enterprise. An American consultancy firm and knowledge management software developer arrived at USD 0.23 for a word in every of its occurrences in technical documentation. If a term is used:

- 10 times in a document,
- in documents for 4 models of a product,
- translated into 7 languages,
- in several formats of the same document,
- stored on several media,

this results in costs exceeding USD 160.00. This further multiplies with every:

- additional model developed;
- further media used for storage;
- other language used for localisation.

Unless the enterprise does not have a central directory, register, repository or index for all terms used in all its documentation, the cost for a global exchange of a word or term in an item of a product catalogue e.g. from "fastened by a steel 3-1/2 threaded bolt", to "fastened by an aluminium 3-1/2 threaded bolt", across all documents on 5 related models in 4 languages in 3 formats would cost USD 138.00 compared to USD 9.20 in case of an appropriate information/knowledge system in place. In e-business in Europe today this lack of appropriate tools already sums up to more than 1 billion USD with a tendency to double every year for the years to come.

The above accounts only for the immediately calculable costs for word units in written documentation, not taking into account the positive effects on:

- product liability,
- quality assurance,
- internal training and external user training,
- corporate identity, etc.

which a firmer grip on 'corporate language' and terminology might bring about.

Increasingly, system designers and developers recognise that only more refined data

models (in terms of a higher degree of granularity and a higher degree of international unification and harmonisation) can enable information and knowledge management in the organisation to cope with the above-mentioned cost situation. A higher degree of standardisation of
- metadata,
- data modelling,
- meta models,
i.e. methodology standardisation with respect to LRs, is a prerequisite for achieving satisfactory solutions for information and knowledge management based on content management in the enterprise.

Christian Galinski
International Information Centre for Terminology (Infoterm)
A-1120 Vienna, Aichholzgasse 6/12 (Austria)
Tel.: +43-1-8174488
Fax: +43-1-8174488-44
Email: infopoint@infoterm.at
Web site: http://www.infoterm.org

SENSEVAL: The Evaluation of Word Sense Disambiguation Systems

Philip Edmonds

Word sense disambiguation

Word sense disambiguation (WSD) is the problem of deciding which sense a word has in any given context. The problem of doing WSD by computer is not new; it goes back to the early days of machine translation. But like other areas of computational linguistics, research into WSD has seen a resurgence because of the availability of large corpora. Statistical methods for WSD, especially techniques in machine learning, have proved to be very effective, as SENSEVAL has shown us.

In many ways, WSD is similar to part-of-speech tagging. It involves labelling every word in a text with a tag from a pre-specified set of tag possibilities for each word by using features of the context and other information. Like part-of-speech tagging, no one really cares about WSD as a task on its own, but rather as part of a complete application in, for instance, machine translation or information retrieval. Thus, WSD is often fully integrated into applications and cannot be separated out (for instance, in information retrieval, WSD is often not done explicitly but is just by-product of query to document matching). But in order to study and evaluate WSD, researchers have concentrated on standalone, generic systems for WSD. This article is not about methods or uses of WSD, but about evaluation.

SENSEVAL

The success of any project in WSD is clearly tied to the evaluation of WSD systems. SENSEVAL was started in 1997, following a workshop, "Tagging with Lexical Semantics: Why, What, and How?", held at the conference on Applied Natural Language Processing. Its mission is to organise and run evaluation and related activities to test the strengths and weaknesses of WSD systems with respect to different words, different aspects of language, and different languages. Its underlying goal is to further our understanding of lexical semantics and polysemy.

SENSEVAL is run by a small elected committee under the auspices of ACL-SIGLEX (the special interest group on lexicon of the Association for Computational Linguistics). It is independent from other evaluation programmes in the language technology community, such as TREC and MUC, and, as yet, receives no permanent funding.

SENSEVAL held its first evaluation exercise in the summer of 1998, culminating in a workshop at Herstmonceux Castle, England, on September 2-4 (Kilgarriff and Palmer 2000). Following the success of the first workshop, SENSEVAL-2, supported by EURALEX, ELSNET, EPSRC, and ELRA, was organized in 2000-2001. The Second International Workshop on Evaluating Word Sense Disambiguation Systems was held in conjunction with ACL-2001 on July 5-6, 2001 in Toulouse (Preiss and Yarowsky 2001).

The rest of this article describes the SENSEVAL-2 exercise- its tasks, participants, scoring, and results. The article concludes with a short discussion of where SENSEVAL is heading.

SENSEVAL-2: Tasks and participants

The main goal of SENSEVAL-2 was to encourage new languages to participate, and to develop a methodology for all-words evaluation. We were successful: SENSEVAL-2 evaluated WSD systems on three types of task on 12 languages as follows:

All-words: Czech, Dutch, English, Estonian

Lexical sample: Basque, English, Italian, Japanese, Korean, Spanish, Swedish

Translation: Japanese

In the all-words task, systems must tag almost all of the content words in a sample of running text. In the lexical sample task, we first carefully select a sample of words from the lexicon; systems must then tag several instances of the sample words in short extracts of text. The translation task (Japanese only) is a lexical sample task in which word sense is defined according to translation distinction (by contrast, SENSEVAL-1 evaluated systems only on lexical sample tasks in English, French, and Italian.).

Language	Task	N° of Submissions	N° of teams	IAA	Base-line	Best system
Czech	AW	1	1	-	-	.94
Basque	LS	3	2	.75	.65	.76
Estonian	AW	2	2	.72	.85	.67
Italian	LS	2	2	-	-	.39
Korean	LS	2	2	-	.71	.74
Spanish	LS	12	5	.64	.48	.65
Swedish	LS	8	5	.95	-	.70
Japanese	LS	7	3	.86	.72	.78
Japanese	TL	9	8	.81	.37	.79
English	AW	21	12	.75	.57	.69
English	LS	26	15	.86	.51/.16	.64/.40

Table 1 - Submissions to SENSEVAL-2

Table 1 gives a breakdown of the number of submissions and teams who participated in each task. Overall, 93 systems were submitted from 34 different teams. Some teams submitted multiple systems to the same task, and some submitted systems to multiple tasks. Dutch data was also prepared, but was not available in the exercise. Inter-annotator agreement (IAA), and system performance is discussed below.

A task in SENSEVAL consists of three types of data.

1) A sense inventory of word-to-sense mappings, with possibly extra information to explain, define, or distinguish the senses (e.g., WordNet).

2) A corpus of manually tagged text or samples of text that acts as the Gold Standard, and that is split into an optional training corpus and test corpus.

3) An optional sense hierarchy or sense grouping to allow for fine or coarse grained sense distinctions to be used in scoring. General guidelines for designing tasks were issued to ensure common evaluation standards (Edmonds 2000), but each task was designed individually.

WordNet was used for the first time in SENSEVAL, version 1.7, for the English tasks, and versions of EuroWordNet for Spanish, Italian, and Estonian. WordNet was chosen because of its wide availability and broad coverage, despite the often unmotivated demarcation of senses (Wordnet was designed from the point of view of synonymy rather than polysemy). In fact, WordNet 1.7 now includes revisions suggested by the human-tagging exercise for SENSEVAL-2.

The Gold Standard corpus must be replicable; the goal is to have human annotators agree at least 90% of the time. In practice, agreement was lower (see Table 1). At least two human annotators were required to tag every instance of a word, but often more annotators were involved in order to settle disagreements.

SENSEVAL-2: Evaluation procedure and results

Regardless of the type of task, each system is required to tag the words specified in the test corpus with one or more tags in the sense inventory, giving probabilities (or confidence values) if desired. A distinction is made between supervised systems, that use the training corpus, and unsupervised systems, that do not. An orthogonal distinction is made between systems that use just the test corpus (pure unsupervised) and systems that use other knowledge sources, such as dictionaries or corpora, but, in practice, few systems are pure.

The evaluation was run centrally from a single website at the University of Pennsylvania and followed the same procedure as used in SENSEVAL-1. For each task, data was released in three stages: trial data, training data (if available), and test data. Each team registered their system, and then downloaded the required data according to a set schedule. Teams had 21 days to work with the training data and 7 days with the test data. Each team submitted their answers to the website for automatic scoring. The Japanese tasks were handled separately because of copyright issues.

SENSEVAL-1 established a scoring system that was used again in SENSEVAL-2 with little change. Fine-grained scoring was used to score all systems. If the task had a sense hierarchy or grouping, then coarse-grained scoring was also done. In fine-grained scoring, a system had to give at least one of the Gold Standard senses. In coarse-grained scoring, all senses in the answer key and in system output are collapsed to their highest parent or group identifier. For sense hierarchies, mixed-grained scoring was also done: a system is given partial credit for choosing a sense that is a parent of the required sense according to Melamed and Resnik's (1997) scheme.

Systems are not required to tag all instances of a word, or even all words, thus, precision and recall can be used, although the measures are not completely analogous to IR evaluation. Recall (percentage of right answers on all instances in the test set) is the basic measurement of accuracy in this task, because it shows how many correct disambiguations the system achieved overall. Precision (percentage of right answers in the set of answered instances) favours systems that are very accurate if only on a small subset of cases that the system could give answers to. Coverage, the percentage of instances that a system gives any answer to, is also reported.

Table 1 gives an overview of the results, as reported in Preiss and Yarowsky (2001). Inter-annotator agreement (generally, the percentage of cases where two human annotators agree on a sense, but this varies depending on the task), is shown. Baseline performance is generated in different ways, but usually as most frequent sense in the tagged corpus. The recall of the best system with perfect or near-perfect coverage is given for each task. For the English lexical sample task, scores for supervised and unsupervised systems are separated by a slash.

Notably, the results in SENSEVAL-2 were about 14 percentage points lower than in SENSEVAL-1 (for the English lexical sample), even though the same evaluation methodology was used and many of systems were improved versions of the same systems that participated in SENSEVAL-1. This can be seen as evidence that WordNet sense distinctions are indeed not well-motivated, but more research is required to confirm this.

Edmonds (2001) gives a more complete account of SENSEVAL-2 evaluation methodology. Almost all data and results of SENSEVAL is in the public domain. Visit the web site to download it.

Where next?

SENSEVAL-2 was very successful in opening up new avenues for research into WSD and polysemy. It's clear that the current best systems achieve their high performance by using supervised machine learning. Research is now ongoing to explore how feature selection for the machine learning algorithms affects the performance on different types of polysemy. Indeed, it is hoped that we can now identify different types of polysemy on the basis of how easy or difficult the words are to disambiguate with different features and methods. Another result of SENSEVAL-2 was to underline the importance of a well-motivated sense inventory with the right level of granularity of sense distinction. If humans cannot reliably disambiguate a word based on the information in the sense inventory, then there is no meaningful way of evaluating a system. Efforts are ongoing to design new methodologies for building sense inventories and for annotating large corpora, which will inform research in lexicographics and lexical semantics. In particular, researchers are investigating methods

to form well-motivated groupings of senses. Finally, the task of WSD set up in SENSEVAL is very divorced from real applications. Questions run from whether the sense distinctions in generic resources are useful, in particular applications or domains, to whether a separate WSD module is useful, to whether we need to make explicit sense distinctions at all.
Planning for SENSEVAL-3 is currently underway and the SENSEVAL Committee welcomes proposals for tasks to be run as part of exercise. Any task that can test a word sense disambiguation (WSD) system, be it application dependent or independent, will be considered. The committee especially encourages tasks for different languages, cross-lingual tasks, and tasks that are relevant to particular NLP applications such as MT and IR. It also encourages tasks for areas related to WSD such as semantic tagging and domain classification.
Visit http://www.senseval.org/ for more details.

References

Philip Edmonds (2000). *Designing a task for SENSEVAL-2. Technical Note*. Senseval-2 website.
Philip Edmonds and Scott Cotton (2001). *SENSEVAL-2: Overview*. In Preiss and Yarowksy (2001), pages 1-5.
Adam Kilgarriff and Martha Palmer (2000). Guest editors. *Special Issue on SENSEVAL: Evaluating Word Sense Disambiguation Programs*. Computers and the Humanities 34(1-2).
Dan Melamed and Phil Resnik (2000). *Tagger evaluation given hierarchical tag sets*. Computers and the Humanities 34(1-2).
Judita Preiss and David Yarowsky (2001). Editors. *The Proceedings of SENSEVAL-2*: Second International Workshop on Evaluating Word Sense Disambiguation Systems.
SENSEVAL Website:
http://www.itri.bton.ac.uk/events/senseval
SENSEVAL-2 Website:
www.sle.sharp.co.uk/senseval2

Philip Edmonds
Sharp Laboratories of Europe
Oxford Science Park
Oxford OX4 4GB (United Kingdom)
Email: phil@sharp.co.uk
Tel.: +44 1865 747711
Fax: +44 1865 714170

Creation of a new Electronic Corpus: the Hermes Journal on Humanities

Richard Walter

Following a call issued by ELRA/DGLF in 2000, on the enrichment of contemporary French, we have worked on the creation of a corpus of language resources in electronic form, based on the digitisation of the original paper version of the Hermes journal. In this article, we analyse the first steps of the process and we explain how the digitisation of a text effects the format, the quality and the time required to build the language resource. Indeed, the problems met during the digitisation process have influenced the mark-up step. Although we do not claim to be exhaustive on the matter, we describe a few problems, which can be viewed as a first step towards a state-of-the-art on the question.

In order to extend the French part of the Parole corpus, we have digitised the first 10 issues of the Hermes journal, edited by CNRS-Éditions. This set extends the 10 issues that were formerly integrated in the Parole corpus. Hermes is a reference journal in the field of Humanities. For its first 10 issues, there was no electronic version available, which made it necessary to digitise, verify and reformat the output of the digitisation. The entire process was divided into three steps, using various software and formats (i.e. Image, Word and HTML). The result is a structured corpus, with standard (SGML) mark-up, and a total size of 10.7 Gb.

The first obstacle came from the fact that this corpus was edited over a 10-year period. Through the years, the editorial chart evolved. The typology of the articles was gradually modified, especially in what concerns the “paratext” : title, author, notes, references, abstract, etc... This heterogeneity must be taken into account in the editorial principles during the formatting process. For instance, half way through the 10 years, an English abstract was introduced, as well as the mention of the author's affiliation. Given that these elements are missing at the beginning of the 10-year period, should they be ignored when they appear? We think that they should not. The “physical” aspect of the document must be rendered as accurate as possible in terms of logical indications. At the same time, the encoding must be homogeneous across the corpus. It is thus necessary to find the right balance between, on the one hand, the closeness to the original edition and, on the other hand, the need for an easy electronic mark-up process. Therefore, it is necessary to preserve the presentation and the evolution of the original content but this must not go against a minimal structure and mark-up of the resource.

Each page has been digitised in the Tiff 300 dpi format. This relatively high resolution was chosen, even though it requires more time-consuming processing and larger storage files. A 72 dpi resolution would have been sufficient for legibility, but the quality of the text output have been too poor. With 300 dpi, the results were better and the image file was of adequate quality for the verification step.

The images were converted into text using the OCR (Optical Character Recognition) software Omnipage 7.0. This software is able to process the text detected in the image, to compensate for a possible rotation or curvature, and to convert the character patterns into letters and words. The performance of the conversion turned out to be variable but globally correct. The high recognition rate can be explained by the fact that the texts are recent, with modern fonts, and printed on a good quality paper. The performance would certainly be lower for older material.

The principles on which the OCR software is based, namely the reuse of a pattern that was detected previously, have both advantages and drawbacks. In fact, as soon as the software recognizes a particular pattern, this pattern is stored and, when a similar one is detected later, they are merged with each other. If there is no error in the initial recognition, this approach saves a lot of time. However, in the opposite case (for instance if “the” is transformed into “the”), it is necessary to post-process the document and carry out a number of “replace” operations. In practice, we have decided not to use the spell-checker provided with the OCR software, and we have carried out the verification step in a semi-manual mode with another software, in order also to preserve the few existing mistakes in the original document.

To our knowledge, the OCR software is not multilingual. We have used it with French as the recognition language, since the vast majority of texts were in French. It is therefore normal that the software was less efficient on non-French orthographic forms, especially on Greek characters. A multilingual OCR software program would be desirable.

A second aspect that must be taken into account is the recognition of the text structure. Here, the results may be quite variable according to the shift in the page orientation, the uniformity of the background or the homogeneity of the spaces between the lines. As the corpus was recently edited and as the original documents were in good condition, the performance of the structure recognition was satisfactory, even though some passages were occasionally misplaced in the text file. The software performs a block-wise content analysis, based on the structure of the paragraph. A simple “spot” in the picture (for instance some dust on the document or dirt on the scanner's window) or a small variation of the page layout (a shifted line or paragraph) is sufficient to cause the creation of a new block of text, which disturbs the whole structure of the document. Within a paragraph, the software is able to reassemble words that have been hyphenised and it is also able to distinguish between the end of a line and the end of a paragraph. But the results are not always conforming to the truth, especially when a line ends by a weak punctuation (for instance, a comma). It is necessary to review the result in order to check and adjust the paragraph structure.

Some problems arise occasionally, such as difficulties in the recognition of “small” visual characters (for instance, a comma often becomes an apostrophe), double dots, double consonants, words in capital letters, etc. Other problems arise from variations in the original layout of the document: according to the numbering, or the type of apostrophes and inverted commas, the typographic hierarchy of the titles is different (bold, italic, capital letters or a mixture of the three). Typographic rules also vary: the space before a double punctuation sign, the presence or absence of a space between the inverted commas and the quoted text. These variations tend to be reproduced by the OCR software and it is only during the verification and formatting steps that these various modes of presentation can be harmonised, by using a unique and systematic convention.

Finally, the processing of text in italics happens to be delicate and the recognition step can significantly degrade its content. The OCR software performs less well in recognising slanted characters (italics) vs. straight ones (roman). Moreover, the software is not very accurate in locating the zones in italics themselves: quite often, the words following a quotation in italics are in italics themselves in the recognised text, whereas they are in roman font in the original. This happens even more regularly for the first letters of the first word and for the punctuation just after the zone in italics. Conversely, in a quotation in italics, the first words may be recognised as roman... A systematic verification is therefore necessary to correct for these defects. It is imperative to do so, if one wants to respect the original form of the content and the editorial and typographic norms. In order to preserve the various levels of text defined by the author, and to render the “zone-based” structure of the final document, we have followed several principles: italics mark-up is preserved in the final encoding; quotations remain in italics and between inverted commas (a typographic redundancy) if this was the general rule for a particular issue; book titles are always in italics, punctuation marks after a zone in italics are always in roman, etc...

Throughout these examples, one can measure how the time demanded for verification varies a lot from one page to another: it can be done very quickly for pages that have no particular typographic or editorial enrichment (italics, quotations, tables, call for footnotes...) but it takes much more time for “rich” pages. When verification is done without using any a priori knowledge concerning the layout and the mark-up of the document, proper detection may fail, which can bring about a lot of mistakes and a lack of homogeneity. Therefore, it is necessary to study from the beginning the corpus as a whole and to create a list of “principles” to be used during the verification and correction process, according to the targeted level of mark-up and the final objective for which the resource is created.

We have chosen to do the verification step with the Word 7.0 software (Windows 98), with the help of its built-in spell-checker. The conversion of the corpus into the Word format has a certain number of advantages, but also some drawbacks. First of all, it is a user-friendly software program in terms of text readability on a screen, which is a non-negligible advantage for the person in charge of the verification of such a large size corpus. It is also designed for handling various styles or for converting automatically page skips into SGML format. However, we were not able to implement the automatic conversion of footnote calls or titles, because the initial layouts were too heterogeneous, or too similar to the rest of the content. Still, we have benefited from the Word format as a

means to harmonise manually, as much as possible, all these aspects and to prepare their forthcoming mark-up.

Once verified and corrected, the Word files were converted into HTML files, using the Word software. It turned out to be necessary to re-work the output, because the HTML mark-up generated by the software programme was not in standard HTML 4.0, as was recommended by the W3 Consortium. In particular, we had to convert the Greek characters. Otherwise, the paragraph structure was properly handled, as well as the physical and typographic characters, but everything had to be reconsidered for what concerns the text structure (headers, titles, subsections, notes and footnote calls).

The next step was to carry out the SGML mark-up itself. The mark-up has to respect as much as possible the original edition, especially if the resource is to be used in various contexts, including, of course, a paper reprint. We have personalised the mark-up so as to allow for these possibilities, but we have been forced to be slightly unfaithful to the original edition, never in terms of content but in terms of "editorial enrichment". What is the tolerable level of alteration in that case? To answer this question, we detail now these operations in terms of processing and conversion.

Thanks to the SGML structure of the marked up document, some editorial elements can be suppressed: it is useless to repeat the author names and the title at the top of the page or at the end of the article. The type and the size of the fonts were not preserved. This may be questionable, but we did so because the original electronic files were not available. On the contrary, with the help of a word pre-processing, the structure mark-up was rather easy. After having harmonised the indications concerning the author, the title, the beginning and the end of the article, and other peripheral elements, we were able to replace automatically the corresponding localisations into segmentation marks. However, a verification step was necessary because of specific cases on which we had to decide to alter the physical aspect of the content.

In the first issues, the author indication was limited to the name and the surname. Only later was included the author's affiliation. This last piece of information could not be retrieved for previous issues. It was also necessary to process specifically the few texts without author names or signed by the Hermes editorial board. In two issues, the author's biographies were presented at the end of the issue. In order to keep a single structure for all articles, we shifted each biography at the end of the article of the corresponding author, while keeping unchanged the pagination. For all other issues, we did not introduce the corresponding mark-up, because the corresponding field would have been systematically empty.

In some issues, and even sometimes in some articles within an issue, subsections of articles were numbered in different ways, numbers or letters. In order to have a homogeneous structure common to all articles in the corpus, we did not introduce any specific mark for the subsection numbering. However, we kept the indication of the numbering as part of the title of the corresponding subsection.

An other point appeared to be sensitive: the quotations. The choice was made not to complicate the mark-up and to stay as close as possible to the original. Most of the time, the quotations were embedded within a paragraph; apart from the presence of inverted commas or italics, it was not obvious how to distinguish them from the rest of the article. There were only a few cases when the quotation was physically separated from the rest of the text, in a specific paragraph. Moreover, the length of the quotations was very variable - from one single word to several sentences. We were not able to find a criterion so as to mark up a "significant" quotation as a quotation. We have rather dedicated our processing effort to the proper localisation of inverted commas and italics, as being the only signs available to designate a quotation. The indication of a cut within a quotation was harmonised and systematically corrected. In the original corpus, this was notified in various ways: …, (…), […], etc. We chose to group them into a single form: (...). This harmonisation is probably not very important for lexicographic studies, but it can be useful for other purposes, such as the study of the "quotational" system or the re-edition of excerpts from several issues. In the latter, it is preferable that the graphical, typographical and editorial conventions are as consistent as possible.

The greatest difficulty we met was the processing of the footnote calls within the body of the article. For the footnotes themselves, the journal always had the convention to place them at the end of the article (rather than at the end of the issue). This made it easy to integrate the footnotes in the structure of the article. On the contrary, the mark up of the footnote calls as specific elements was not feasible but semi-manually. The number of notes is very variable across articles (from zero to 60 and more) and so is their position in the text: joined or separated from the preceding term, inside or outside a quotation, as symbols or numbers, sometimes even as numbers between brackets. In order to have a homogeneous system of footnote calls across the whole corpus and so as to facilitate the use of the links between footnotes and footnote calls, we have adopted the same conventions everywhere: a footnote call is separated from the preceding term; in quotations, it is systematically placed outside the inverted commas; it is always a number, possibly followed by a letter (so that symbols could be replaced using this convention without modifying the original numbering system). The localisation and the harmonisation of all the footnote calls in the corpus required a number of systematic operations. It was more efficient to do this at this point rather that during the verification steps.

Finally, we made two modifications to the peripheral information related to articles: the references and the abstracts.

References at the end of articles appear with different names: references, bibliographic notes, bibliographic indications, etc. We kept the same mark for all elements in the references. We have also harmonised the typographic rules (title in italics, first “significant” letter in capital, etc.). Originally, abstracts were only in French, then bilingual French / English. We have systematically marked up the existence of the English version. We also changed the place of the abstracts to the end of the issue, the goal being to have the article as the basic unit. We therefore moved the abstract in the marked up structure of the article, in the last position, still keeping the indication of the original pagination.

These were not the only operations on the corpus, but they seem to us quite representative of the difficulties that were met and the decisions that we had to make in order to solve them. The condition and the age of the corpus is a decisive factor, as well as the various transformation steps that were taken to turn it into a marked-up structured corpus (scanning, optical character recognition, text processing, HTML conversion, SGML marking), and the objectives for which the corpus is designed. Processing time and cost vary enormously according to these choices.

Our final conclusion remains that it is absolutely necessary to take some liberty from the initial format of the content. The task was made more complicated here by the fact that each issue had some editorial peculiarities, which made it difficult to define a standard conversion procedure. However, this liberty has to be controlled. It must not concern the very content of the corpus but rather its structure, its typographic enrichment and its set of notes and quotations. An imperious precaution must be taken by reasoning on the entire corpus and not on a single issue. The creation of acquisition, correction and mark-up grids is thus facilitated. This requires some time to acquire familiarity and to build a global view of the corpus.

Richard Walter
Laboratoire MoDyCo
UMR 7114 (CNRS - Université Paris X)
Université Paris X - Nanterre
Bât L
200, avenue de la République
92001 Nanterre cedex (France)
Tél. 01 40 97 47 34
Fax.: 01 40 97 40 73
richard.walter@u-paris10.fr

EUROSPEECH 2003 SPECIAL EVENT

AURORA: NOISE ROBUST RECOGNITION

Robust Algorithms and a Comparison of their Performance on the “Aurora 2, 3 & 4” Databases

The objective of this special session is for researchers to present leading edge algorithms for noise robustness and their results measured on the same databases. It is hoped that not only will the research community benefit from comparing techniques and reviewing scientific progress but also the process of evaluating on a common database will stimulate new ideas. For this session we have split it into 2 streams:

Stream 1: Small vocabulary: Aurora 2 and Aurora 3

In addition to the Aurora 2 database, researchers are invited to evaluate their algorithms on the set of Aurora 3 databases. While Aurora 2 databases use the controlled addition of noise to clean speech, the Aurora 3 databases are collected in a real-world environment of the car. New baselines will be based on the Advanced DSR front-end & "complex" backend.

Stream 2: Large vocabulary: Aurora 4

This is a new task introduced for Eurospeech 2003. The database has simulated noise addition the 5000 word WSJ task. The large vocabulary adds a further dimension to the evaluations. Also be aware that the processing requirements for this database are substantially larger than for the small vocabulary tasks.

What makes this special session different from the main conference is that each paper will be required to submit results on the evaluation databases. These databases have been prepared within the ETSI standards activity in the Aurora Distributed Speech Recognition working group for the purpose of evaluating the performance of noise robust front-ends. They are available publicly through ELRA at a low price to encourage widespread use.

We invite submissions of papers on noise robust speech recognition including:

- Front-end feature extraction
- Pre-processing techniques

- Noise adaptation
- Noise modelling and compensation
- Missing data techniques
- Combinations of new front-ends with back-end compensation techniques

Important dates:

Submission of full paper for publication: April or May (TBD)

Eurospeech Special Session: Sept 2003

Please send an email to David Pearce <bdp003@motorola.com> in advance if you intend to submit a paper so we can keep you informed of any updated information.

OPEN POSITIONS WITHIN THE HLT EVALUATION DEPARTMENT AT ELDA

Evaluation department director and Evaluation team

ELDA has been strongly expanding its activities related to the evaluation of Human Language Technologies (HLT). The evaluation department at ELDA is intended to promote the HLT evaluation in Europe, and to act as a clearing house for this area with the support of a network of evaluation units based on a large number of European institutes (both public and private ones).

In order to staff this recently set up evaluation department, ELDA is seeking to fill the following positions:

Department director

He/she will be in charge of managing ELDA's activities related to evaluation and co-ordinating the work of the evaluation team and ELRA evaluation network.

Profile:

- Advanced degree in computer science, computational linguistics, library and information science, knowledge management or similar fields;
- Experience and/or good knowledge of the evaluation programs in Europe and the US;
- Experience in project management, including the management of European projects;
- Ability to work independently and in a team, in particular the ability to supervise the work of a multi-disciplinary team;
- Proficiency in English.

Two junior engineers

They will carry out specific activities in evaluation of HLT.

Responsibilities:

Under the supervision of the evaluation department director, the junior engineers will be involved in the evaluation of Human Language Technologies at ELDA, in the framework of collaborative European and international projects.

Profile:

- Advanced degree in computer science, computational linguistics, library and information science, knowledge management or similar fields;
- Good knowledge of the evaluation programs in Europe and the US;
- Experience in project management, including the management of European projects;
- Ability to work independently and in a team;
- Proficiency in English.

If you would like to receive more information about these job offers, we invite you to contact Khalid Choukri <choukri@elda.fr>.

SCALLA
Sharing Capability in Localistaion and Human Language Technologies

SCALLA, previously known as SCiLaHLT, is a European project conducted in the framework of the Asia Information Technology and Communications programme (AsiaIT&C).

SCALLA aims at encouraging the two-way flow of knowledge about HLT and their application in IT&C systems, e.g. through the localisation activity, between South Asia (India, Pakistan, Sri Lanka, Bangladesh, Nepal, Bhutan, Maldives) and Europe.

The main goal is to reduce the linguistic and cultural barriers to the use of information technologies and communications, thus making it easier for the Asian community to access every aspect and feature of the information society.

Three working conferences with experts from South Asia and Europe have been planned. The first working conference has already taken place, where HLT researchers and experts from both areas have met to share and exchange knowledge and experiences in the field, especially in localising.

Overview of SCALLA 2001 (Bungalore, India, 21-23 November 2001)

7 experts from Europe and 20 from within South Asia were brought together to participate in this first conference, which aimed at drawing a report on the state of the art of HLT and localisation in South Asia. The topics discussed on this occasion were related to:

- Localisation needs and practices: current status, economics of localisation, computer support, etc.
- Writing systems: computerised representation of writing systems, OCR, etc.
- Cultural aspects: calendar systems, colours, person naming, etc.
- Language models: description of languages, differences between the languages across South Asia, etc.
- Language generation: localising software and content only equals to translating messages and texts?
- Lexicography: status, and uses, of dictionaries in South Asia and Europe, etc.
- Speech and literacy: recordings in many languages and dialects, etc.

The second working conference, SCALLA 2002, is going to be more distributed, exploiting the rich variety of conferences available in Europe. Two project members from India attended the LangTech 2002 conference, to obtain a sound overview of language technologies in Europe. Two more people from South Asia will attend the Localisation conference in Dublin in November, and a workshop entitled “Computational Linguistics for South Asian Languages -- Expanding Synergies with Europe” will be organised in Budapest (Hungary) on Sunday, April 14th 2003, at the Agro Hotel.

The third, and last, conference is to be organised in December 2003 or January 2004, and will take place in India or some other neighbouring South Asian country.

Please have a look on the SCALLA web pages to get more information: **http://www.elda.fr/proj/scalla/**

New Resources

ELRA-S0124 Phonetically Balanced Words (1)

Large acoustic corpus of read text in Korean. 2 announcers and 70 native speakers have been recorded (38 males, 32 females), distributed according to 4 age classes. They read two times 452 eojeols (Korean terms), and 2 announcers read one time 2000 eojeols. In these 2000 eojeols, the above 452 eojeols are included. Other information such as the size and the level of studies of the speakers are provided. The recordings took place in a soundproof room. The data are stored in a 8-bit A-law speech file, with a 16 kHz sampling rate. The standard in use is NIST.

	ELRA members	Non-members
For research use	250 Euro	500 Euro
For commercial use	1,000 Euro	2,000 Euro

ELRA-S0125 Phonetically Balanced Words (2)

Large acoustic corpus of read text in Korean produced by Kaist Korterm. Native Korean speakers (males and females) have uttered 36 geographical proper nouns. Information such as the size and the level of studies of the speakers are provided. The recordings took place in a soundproof room. The data are stored in a 8-bit A-law speech file, with a 16 kHz sampling rate. The standard in use is NIST.

	ELRA members	Non-members
For research use	50 Euro	100 Euros
For commercial use	200 Euro	400 Euro

ELRA-S0126 Phonetically Balanced Words (3)

Large acoustic corpus in Korean produced by Kaist Korterm. Two announcers and 70 native speakers (males and females) read 2 times one paragraph. Information such as the size and the level of studies of the speakers are provided. The recordings took place in a soundproof room. The data are stored in a 8-bit A-law speech file, with a 16 kHz sampling rate. The standard in use is NIST.

	ELRA members	Non-members
For research use	63 Euro	125 Euros
For commercial use	250 Euro	500 Euro

ELRA-S0127 Phonetically Balanced Words (4)

Large acoustic corpus in Korean produced by Kaist Korterm. 70 native Korean speakers (males and females) read 4 times 32 cardinal numbers and 9 determinatives of one syllable. Two announcers read these 2 times. Information such as the size and the level of studies of the speakers are provided. The recordings took place in a soundproof room. The data are stored in a 8-bit A-law speech file, with a 16 kHz sampling rate. The standard in use is NIST.

	ELRA members	Non-members
For research use	200 Euro	400 Euros
For commercial use	800 Euro	1,600 Euro

ELRA-S0128 Phonetically Balanced Words (5)

Large acoustic corpus in Korean produced by Kaist Korterm. 70 native Korean speakers (males and females) read 4 times 35 cardinal numbers compounded of 4 single numbers. Two announcers read these only two times. Information such as the size and the level of studies of the speakers are provided. The recordings took place in a soundproof room. The data are stored in a 8-bit A-law speech file, with a 16 kHz sampling rate. The standard in use is NIST.

	ELRA members	Non-members
For research use	250 Euro	500 Euros
For commercial use	1,000 Euro	2,000 Euro

ELRA-S0129 Phonetically Balanced Sentences

Large acoustic corpus in Korean produced by Kaist Korterm. 50 native Korean speakers (males and females) read 1 time 539 sentences and a set of 50 common sentence. Information such as the size and the level of studies of the speakers are provided. The recordings took place in a soundproof room. The data are stored in a 8-bit A-law speech file, with a 16 kHz sampling rate. The standard in use is NIST.

	ELRA members	Non-members
For research use	500 Euro	1000 Euros
For commercial use	2,000 Euro	4,000 Euro

ELRA-S0130 Phonetically Rich Words

Large acoustic corpus in Korean produced by Kaist Korterm. 500 native speakers have been recorded (250 males, 250 females). They have uttered 32 single cardinal numbers, 1620 cardinal numbers compounded of 4 single numbers and 3813 phonetically rich words. The recordings took place in natural environment, by telephone (wire, wireless and mobile phone). The data are stored in a 8-bit A-law speech file, with a 16 kHz sampling rate. The standard in use is NIST.

	ELRA members	Non-members
For research use	313 Euro	625 Euros
For commercial use	1,250 Euro	2,500 Euro

ELRA-T0365 Biology Database

This bilingual terminology database produced by Kaist Korterm consists of 31,884 entries in Korean and English in the field of biology.

	ELRA members	Non-members
For research use	1063 Euro	2126 Euros
For commercial use	6377 Euro	12754 Euro

ELRA-T0366 Computer Science Database

This bilingual terminology database produced by Kaist Korterm consists of 76,272 entries in Korean and in English in the field of computer science.

	ELRA members	Non-members
For research use	3,814 Euro	7,627 Euros
For commercial use	15,524 Euro	30,509 Euro

ELRA-W0034 Qualified POS Tagged Corpus

Monolingual corpus in a .txt format, produced by KAIST KORTERM, containing 1,020,000 eojeols (Korean terms) in Korean. This corpus is morphologically analyzed, POS tagged, and rectified 3 times by specialists.

	ELRA members	Non-members
For research use	667 Euro	1,333 Euros
For commercial use	4,000 Euro	8,000 Euro

ELRA-W0035 Multilingual Corpus

Multilingual parallel corpus produced by Kaist Korterm containing 60,000 expressions in Korean, Chinese and English.

	ELRA members	Non-members
For research use	750 Euro	1,500 Euros
For commercial use	3,000 Euro	6,000 Euro

ELRA-L0044 Korean Lexicon

This monolingual lexicon produced by Kaist Korterm consists of 31,476 compound nouns in Korean.

	ELRA members	Non-members
For research use	1,049 Euro	2,098 Euros
For commercial use	6,295 Euro	12,590 Euro