The ELRA Newsletter

January - March 99

Vol.4 n.1

Contents

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

ISSN: 1026-8200

ELRA/ELDA
CEO: Khalid Choukri
Assistant: Rébecca Jaffrain

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 jaffrain@elda.fr
WWW: http://www.icp.grenet.fr/ELRA/home.html

Dear Members,

During the last quarter of 1998, we organised our Annual General Assembly. The Board reported on the activities of 1998 and the General Assembly accepted both the Management and Financial reports. In our proposal for future activities, we emphasized the need to pursue our "user requirements" surveys and to focus on market analysis.

The preliminary outcome of these tasks is our project to co-produce Language Resources. The surveys helped us draw preference lists in the Speech, Written, and Multimedia/Multimodal areas. Therefore, we initiated and posted a Call for Proposals early February. This Call aims at encouraging owners of existing resources to package them for a use by a large audience. It also aims at stimulating the production of new resources identified as the most demanded ones in the surveys mentioned above.

The details of this Call are given in this issue of the newsletter. We would like to draw your attention to the key dates: deadline to receive the proposals: 19 March 1999; the notification of acceptance, no later than 5 April 1999. We would like to fund several projects. Although part of the funding comes from the European Commission, the projects will be funded by ELRA under terms and conditions to be negiotiated.

As an EU funded project, ELRA has to go through annual peer reviews. The final review took place in Luxembourg on the 10th of February and allowed us to describe our activities during the three years of the project. The comments of the reviewers were very useful and will be considered in our future plans. Their general conclusion was that ELRA successfully achieved its objectives. A general request made by the reviewers and other partners is related to market analysis and corresponding facts/figures. You may have conducted such surveys; if you would like to share such information with us, please keep us informed as we are prepared to derive (or help you derive) an executive summary from the reports you may have. We strongly believe that disseminating such information will stimulate the market and its results will be a benefit to all.

During this last quarter 1998, we have finalised the first phase of a study carried out by Lancaster University, UK (Tony Mc Enery). This study aims at learning more about the features you expect to see in our catalogue of Language Resources. The outcome is a set of essential, desirable, and not required features. We are starting to discuss the implementation of such recommendations with our providers.

We have also been involved in the AURORA project which is described next page. ELRA distributes the CDROM with the corrupted speech databases and the baseline HTK scripts to all interested parties.

This issue of the ELRA newsletter starts with a brief profile of Jeff Allen who joined us in December 1998 to be in charge of the Written and Terminology technical issues. To continue our evaluation tour, and after Bente Maegaard's paper about Evaluation Methodologies (see Vol.3 n4), Patrick Paroubek (LIMSI-CNRS) elaborates on the contrastive evaluation in NLP and its vital importance for the Industry sector. The second paper from Jean Véronis (Université de Provence, France), concerns an evaluation exercise, ARCADE, and reports on the evaluation of multilingual parallel text alignment systems.

A short summary regarding the Finnish National Language Engineering efforts, by Manne Miettinen, will give you an idea of what is happening in Finland. There is also a paper of Lisa Decrozant (University of Maryland), and Clare Voss (Army Research Laboratory). This one elaborates on the different necessary steps to obtain on-line Language Resources for "the lesser-commonly taught languages" essential to MT evaluation experiments.

As always, the final pages are devoted to the new resources for which we have obtained distribution rights. These are:

- Fixed1itDesign (textual material of the Italian SpeechDat database).
- Colombian Spanish Speech Database
- BREF-120 - A large corpus of French read speech
- Spanish SpeechDat(M) - (Phonetically rich sentences & application oriented utterances such as keywords, digits, etc.)
- Portuguese SpeechDat(M) - (Phonetically rich sentences & application oriented utterances such as keywords, digits, etc.)

Last but not least, we would like to remind you that Roberto Cencioni (head of E4 of DG XIII) recently announced that the first call for proposals, within the Human Language Technologies (HLT) actions, will be launched on the 16th of March 1999. The two areas that will be open to proposals concern Multilinguality and Natural Interactivity. So watch for the official Website at: http://www.linglink.lu

Antonio Zampolli, President Khalid Choukri, CEO

ELDA Profile

Jeff Allen, ELDA Technical Director

Born in Portland, Oregon (USA) in 1966, Jeff Allen completed 2 undergraduate degrees in literature and French in the USA before spending 5 years at the Université Lyon 2 for master's and doctoral studies in linguistics with two theses in Creole linguistics. He joined Caterpillar Inc. in 1995 as a trainer of translation systems and controlled language technical writing. In early 1997, he took a position as Research Linguist at the Language Technologies Institute/Center for Machine Translation of Carnegie Mellon University where he worked until joining ELRA in December 1998. His previous teaching posts include: French at Portland State University (1988-1992), the Ecole Supérieure de Commerce de Lyon (1992) and Indiana University (1994); Ethnology, Communication and English at the Université Lyon 2 (1991-1993); English and French at Executive Language Services in Paris (1993-1994); Sociolinguistics at the Société Internationale de Linguistique where he was visiting head of the department and also taught general linguistics (1992).

He has worked in the areas of controlled language, knowledge- and example-based machine translation, translation terminology databases, translation memory, multilingual text and speech database compilation, speech recognition, speech synthesis, Optical Character Recognition, SGML, and Workflow.

Evaluation of Distributed Speech Recognition - The AURORA project

The Aurora project was originally set up to establish a worldwide standard for the speech feature extraction software which forms the core of the front-end of a DSR (Distributed Speech Recognition) system. Last year, ETSI formally adopted this activity as work items 007 and 008. The two work items within ETSI are:

- ETSI DES/STQ WI007: Distributed Speech Recognition - Front-End Feature Extraction Algorithm & Compression Algorithm
- ETSI DES/STQ WI008: Distributed Speech Recognition - Advanced Feature Extraction Algorithm

The Aurora project is now entering the operational work of WI008. For this purpose, interested parties are invited to contribute to the standardisation activity. In order to initiate the process, the Aurora project has established an experimental framework to enable DSR front-ends to be evaluated. The framework makes use of a speech database based on TI digits from LDC with artificially added noise over a range of SNR's and Entropic's HTK (with a particular configuration) as the "standard" HMM recognizer.

For more information, please see http://www.icp.grenet.fr/ELRA/aurora.html

Conference on co-operation in the field of terminology in Europe, 17, 18 and 19 May 1999

Most importantly, the discipline of terminology deals with all aspects of the communication process. Its products are indispensable tools in the transfer of knowledge.

It is becoming more and more important that terminology is recognised as a scientific discipline in its own right and that a Pan-European infrastructure is established which allows for discussions, the definition of concrete actions and the implementation of transnational forms of co-operation between terminologists and European specialists, which is the main objective of the "Conference on co-operation in the field of terminology in Europe". This Conference is organised at the initiative of the European Association for Terminology (EAFT) in co-operation with the following terminology associations: AETER, ELETO, BriTerm, DTT, TermRom-Bucarest, TermRom-Moldova, DANTERM, Termip, Ass.I.Term, NL-TERM and Pro-TLS.

On December 6, 1998, the members of the Board of the EAFT and representatives, or Presidents, of the above-listed associations met in Paris to prepare the organisation of the Conference. They discussed the objectives of the event, as well as its possible results.

In order to obtain maximum results of the Conference, an inquiry is being conducted among the members of the European national terminology associations. Its analysis will help to carefully determine the topics of the various thematic sessions. In these sessions, experts will address problems that are specific to the activities of terminologists. Simultaneously, demonstrations of terminological tools, websites, etc., will be organised.

It was decided that this meeting will take place on May 17, 18 and 19, 1999, in Paris, or in the region of Paris. The local organiser is the Union Latine. The morning of May 19 will be dedicated to a round table presentation of the conclusions of the Conference. This session will be followed in the afternoon by the annual General Assembly of the EAFT.

The purpose of the Conference is not meant to discuss scientific matters. Its objectives are manifold: among other things, the meeting will deal with problems that terminologists and experts encounter during their work in the field of terminology; to find solutions for these problems and to discuss the establishment of a terminological infrastructure in Europe.

The Conference will be concluded by the drafting of a Plan of Action of decisions to be taken with regard to the requirements defined by terminologists and to establish various forms of co-operation.

In the near future, a Call for Papers will be distributed. The papers of the Conference will be published in the Proceedings.

For more information, please contact:

Ms. Helmi Sonneveld
President
A. van Duinkerkenlaan 39
NL-1187 WD Amstelveen - The Netherlands
Tel: +31 20 685 11 94 - Fax: +31 20 453 75 83
E-mail: topterm@euronet.nl

Contrastive Technology Evaluation in NLP, an asset for Science and Industry

Patrick Paroubek, LIMSI - CNRS

Introduction

With the 5th framework programme of the European Community about to take off, and the first calls for proposals due to be issued within the next few months, the issue of evaluation in NLP is being raised by many whose interest has been renewed. In this contribution to the ongoing debate, we endorse the point of view on evaluation which has been developed within the scope of the ELSE preparatory action (LE-4). ELSE's aims are to propose a generic infrastructure for NLP evaluation in Europe. (see http://www.limsi.fr/TLP/ELSE).

We then comment on the need for the large-scale deployment of a contrastive approach based on technology evaluation, in the field of NLP. We show why such paradigm is needed, what its benefits for the research and industry community are, and how it relates to different paradigms of evaluation. We conclude by presenting a few key issues associated with its practical implementation in the Europe.

The paradigm

The evaluation paradigm which lies at the centre of our work consists of the following elements:

- assembling a group of actors around common technological issues;
- organising an evaluation campaign on common data using common metrics (these may need to be defined for this very purpose) and using a standard formalism, or developing one if none exists;
- organising a workshop where results are discussed and methods openly presented and compared by participants;
- lastly, performing an impact study to gauge the effects that the campaign has had on the field.

Requirements

To support such a paradigm successfully at the European level, we think that the infrastructure needs to provide a common platform where actors from both research and industry find enough items from their respective agendas addressed that they are willing to participate. In our opinion, a possible way to achieve this objective is by having an infrastructure that reinforces comparative and collaborative aspects, is task or application independent, relies as much as possible on automatic procedures (to yield reproducible results), uses a quantitative black box approach, applies to both text and speech, and, of course, provides an answer for multilinguality.

Expectations

Given this, the consequences of deploying the infrastructure would be at least to have clear and unambiguous information about existing technologies available, and a better view of their various pros and cons. In turn, this would lead to a reinforcement of the development and use of standards, an increase in the amount of high quality validated linguistic resources and the availability of new and validated evaluation toolkits, a lowering of the cultural barriers between different application domains, and, last but not least, an acceleration of the technological transfer both from research to industry and from industry to the market. At the same time, the paradigm of evaluation could allow the funding agencies to measure the level of a given technology and to assess the possibility of using it.

Why language engineering?

As in many disciplines, the activities in language engineering are based on empiricism, since it is the only operational paradigm available. Furthermore, these activities are basically data-processing oriented. Hypotheses are tested against the reality found in native material. Progress assessment and alternative selections are done in the same manner most of the time. Thus, a contrastive and quantitative methodology (yielding reproducible results) lies at the core of field activities. But contrary to other fields, language engineering is paradoxical in the sense that for many domains products already exist on the market, although the technology has barely reached a sufficient stage of maturity (e.g. speech recognition and machine translation).

Evaluation in the development lifecycle

Looking at the complete development lifecycle of a product, from the first expression of its underlying concept up to the point where several companies are mass marketing the product, there are a number of crucial transition stages, each associated to a particular type of evaluation. In fact, we can distinguish:

1) basic research evaluation for the assessment of novel ideas;

2) technology evaluation for testing how a given technology performs against a particular problem;

3) user-oriented evaluation for testing how well the implementation of a given technology (an application) performs when used by a user;

4) impact evaluation for gauging the socio-economic consequences of the use of a particular application or technology;

5) programme evaluation for determining how worthwhile a funding programme has been.

The main difference between technology evaluation and user-oriented evaluation lies in the presence or absence of a distinction between end-user considerations and core technology considerations in the evaluation process. Technology evaluation tries to answer the question of which technology is best suited for performing a task, while user-oriented evaluation is more concerned with usability criteria in the deployment environment. Both kind of evaluations are complementary, but technology evaluation appears earlier in the development lifecycle. Impact evaluation appears later in the cycle and its relationship with the other types of evaluation is more diffuse. It tries to combine the results of past technology and user-oriented evaluation with other current socio-economic indicators of the field to produce an analysis of past trends or prospective assessments. Programme evaluation can be seen as a sort of sum of all the other types of evaluation. The ELSE consortium identified technology evaluation as the right tool to support progress in language engineering.

Evaluation in the USA

The USA has a long history of large-scale evaluation programmes spanning several years, which have generated growing interest and started to inspire similar efforts in Europe (e.g. SQALE, GRACE, Aupelf/ARC, SENSEVAL/ROMANSEVAL). For text data, the first MUC evaluation took place in 1987 and the TIPSTER programme started later in 1991. For speech data, the first large scale campaign (Continuous Speech Recognition and Large Vocabulary Continuous Recognition) also date back to 1987. DARPA and NIST were the two funding agencies behind these campaigns. The American government provided the important budget that was linked with evaluation objectives often strongly influenced by military or geo-political considerations. Using a very rough picture, we could say that as far as evaluation based programmes are concerned, the USA followed a top-down approach, backed by a permanent infrastructure (DARPA, NIST, NSF) and supported by LDC for data production and distribution. All this happened in a very homogenous environment and in a single country.

Evaluation in Europe

In Europe the picture is less homogeneous for several reasons. First, the amount of resources devoted to evaluation up till now, is much less and comes from many different sources: the EC sponsors, and among others, the projects EAGLES, DiET, DISC, TSNLP, TEMAA and SPARKLE. The various national initiatives include: in Germany, the Morpholympics and Verbmobil; in France, GRACE and the Aupelf ARCs; in the UK, SENSEVAL/ROMANSEVAL co-sponsored by several EC projects, ELSNET, ELRA and the British government. The diversity of goals and infrastructures behind the different evaluation efforts in Europe is an extra factor of heterogeneity. In Europe unifying factors for evaluation are more likely to be of a scientific or economic nature. The general picture is then more the one of a bottom-up strategy, supporting projects of relatively small scale in the heterogeneous environment of 15 countries. Although ELRA now exists as a central focal point for data collection and distribution, for the time being we still lack a long-term or permanent central institution for supporting evaluation.

The need for data

To progress, language engineering needs high quality, easily available, validated resources. Right now it is obvious that for languages other than American English we lack annotated corpora (of every kind: PoS-tagged, tree banks, semantically disambiguated, etc.), ontologies, lexica and large corpora of speech transcriptions. Most of these could be produced at low cost through the deployment of the paradigm of evaluation that we have been describing so far.

Multilingualism

Apart from the straightforward but unpractical solution of running several instances of the same evaluation campaign in different languages, ELSE has identified two other means of addressing multilinguality (there are at least 11 working languages in Europe):

1) explicit cross-language functionality requirements,

2) each evaluation should be performed by all participants in at least two languages, one common to all the participants (and possibly to all the campaigns of the programme, (a strong candidate for such being English), and one specific to each participant.

Neither is an ideal solution, the first poses problem for inherently monolingual tasks (e.g. speech synthesis), the other does not provide a clear answer for results generalisation. Multilinguality is and will remain a difficult problem for evaluation in Europe.

Possible path to deploy Evaluation along

As a start, coverage of most of the domain, and addressing the current preoccupations as identified by the community - dialog management, translation and information retrieval [MLIM98] on the one hand, multilinguality, natural interactivity and active assimilation, and use of digital content [HLT98] on the other - could be achieved by launching 6 evaluation campaigns for the following control tasks:

1) broadcast news transcription,

2) cross lingual information retrieval/ extraction,

3) text to speech synthesis,

4) text summarisation,

5) language model evaluation,

6) text annotation (PoS, lemmas, syntactic functional relations and word sense).

Many more possibilities exist to choose from, for instance ELSE listed 30 different control tasks that could be used for supporting evaluation. Of course, the latter only holds if evaluation is deployed using large-scale campaigns following a pro-active strategy (for which the topics addressed are defined beforehand). In fact, evaluation could also be implemented on a large-scale in Europe using a re-active scheme based on the technologies present in the selected projects and organised in technological clusters in the same fashion as they are now organised around market-opportunity clusters. But in the opinion of the ELSE consortium, the latter seems to offer fewer benefits and raises more infrastructural problems (long term management, calls organisations, participants selection, topic selection, project clustering itself etc.). Of course, any combination of these two extremes is possible. As for the budget required for such programme, a very rough estimate yields a total of euro 3.6 m for 6 tasks over 4 years, an impressive amount compared to what has been spent on evaluation in the past in Europe (an Aupelf ARC campaign of 2 years supporting 1 language is estimated at 167 K€), nevertheless a small sum compared to an estimated $20 m per year devoted by DARPA to finance its programme.

References

MLIM98:

Eduard Hovy, Nancy Ide, Robert Frederking, Joseph Mariani, Antonio Zampolli, Editors, "Multilingual Information Management - Current Levels and Future Abilities". A study commissioned by the US National Science Foundation and also delivered to the European Commission's Language Engineering Office and the US Defense Advance Research Projects Agency, July 1998. (URL:http://ww.cs.cmu.edu/~ref/mlim/).

HLT98:

European Commission, Human Language Technology, "Proposal Concerning the IS Programme 1998-2002 (excerpts)", COM (98) 305 Final, 13 May 1998.

(URL://http:/www.linglink.lu/le/ist/ist/excerpts_ist_pgme.htm)

Patrick Paroubek
LIMSI-CNRS
BP 133
91403 Orsay Cedex
France
Tel: +33 (0) 1 69 85 81 91
Fax: +33 (0) 1 69 85 80 88
Email: paroubek@limsi.fr

ARCADE project: Evaluation of multilingual parallel text systems

Jean Véronis, Université de Provence

1. Introduction

The ARCADE project is one of the Actions de recherche concertées (ARC, Strategic Research Actions) financed by AUPELF-UREF ("Association of French-speaking universities") in the field of the language engineering. This project aims at evaluating multilingual parallel text alignment systems, i.e., texts that are parallel translations of one another. Other ARCs treat the evaluation of other aspects of Natural Language Processing (natural language access to textual data , automated extraction of terminological and semantic databases, message understanding, speech dictation, voice dialog, speech synthesis, cf. Mariani, 1998).

The project lasts for 4 years (1996-1999) and consists of a friendly race between systems developed in different countries. There are two main tracks: the first is sentence alignment and the second is word and phrase alignment. This report describes the progress on evaluation at the two-thirds mark of the project (September 1998). The project hosts a Web site[1] where complementary information can be found, as well as an e-mail discussion list[2] to which anyone interested can subscribe. The aligned corpora that have been tested will be made available through ELRA.

2. Aligning parallel texts

Translation is certainly a long-established trade with parallel texts that can be traced back to ancient times. For example, bilingual inscriptions could be found on the gravestones of Elepantin during the third millenium BC. The idea of exploiting these texts in a particular way is a fairly recent activity. The most well-known example is that of the Rosetta stone that was discovered by Napoleon's soldiers in 1799. This stone, containing a text written in two languages, produced in three types of writing (hieroglyphics, demotic and Greek) was the key that Jean-François Champollion used to decipher the hieroglyphic writing system.

It was not until the 1980s that parallel texts were exploited in a systematic manner within the framework of computational linguistics. A few attempts were made during the 1950s, yet the issues of memory and processing of the computers at the time did not allow for decent processing of the data. The first alignment method was developed by Martin Kay starting in 1984, after which many methods for textual alignment were created for differents levels of alignment: paragraphs, sentences, words and phrases. The types of applications are quite diverse, including the creation of translation memories, automatic term extraction, bilingual terminology glossary extraction, compilation of Computer-Assisted Language Learning examples, knowledge extraction for cross-lingual information retrieval, etc. Given the increasing importance of multilingualism in the language engineering industry, pushed by globalization efforts for information and other markets, the processing of parallel text corpora appears to have a promising future.

Table 1: Participating teams

Laboratory	Location	System
Sentence alignment		
CTT&LIA	Stockholm, SW	APA1
IRMC	Tunis, TU	IRMC
ISSCO	Genève, CH	ISSCO
LORIA	Nancy, FR	LORIA
RALI	Montreal, CA	SALIGN, JACAL
CEA	Gif-sur-Yvette, FR	CEA
CTT&LIA	Stockholm, SW	APA2
West Group	Eagan, MN, USA	GSA, GSA+
LILLA	Nice, FR	LILLA
RALI	Montreal, CA	SFI
Word alignment		
CEA	Gif-sur-Yvette, FR	CEA
LILLA	Nice, FR	LILA
Linköping Univ.	Linköping, SW	LWA
RALI	Montréal, CA	RALI
XEROX	Grenoble, FR	XEROX

3. Sentence alignment

The evaluation was conducted on an aligned bilingual test corpus (French-English) that was compiled by the team at RALI (University of Montreal) and my team at the University of Provence. This corpus contains approximately 800,000 words per language and contains different types of texts including: institutional texts, scientific articles, technical manuals, literature. Certain texts presented structural obstacles: missing segments, word order differences (in glossaries for example), etc.

Twelve systems have been evaluated up until now (Table 1), using the same protocol: the participating teams, having received non-aligned texts that were broken up into individual sentences, were required to return the texts aligned at the sentence level by a set deadline. The analysis was conducted according to general principles of precision and recall. Given that certain systems are more favorable to precision, and others to recall[3], a global efficiency measurement, combining the two measurements, was used for the final results (F measure, i.e., the average of precision and recall[4]).

The results showed a large range in performance by different systems. However, the best systems that were tested attained a coefficient of F between 97,8 and 99,7% on "normal" texts that did not contain any significant structural differences. These results fall below 92% on texts that contain such differences (although the systems with the best performance are not the same on the two different text types). The style of texts seems to have little influence on performance with respect to structural differences.

4. Word alignment

Word (and phrase) alignment is undoubtedly a more difficult problem, and alignment techniques for this are currently underdeveloped. Beyond this technical difficulty, word alignment finds itself confronted with some theoretical difficulties: from a linguistic perspective, it is not easy to match each word in a sentence with the exacts equivalent words in the translated version of the text. Grammatical morphemes are particularly one of the most difficult problems for word matching.

In order to make the evaluation possible, the project decided to adopt a simpler task than full alignment of all the words in the two texts. This task involved translation spotting of a given test group of words. Such a task can be done more easily than full alignment as one is able to eliminate the problematic examples of grammatical translation spotting. It is also useful, as such, for a wide range of applications (e.g., translation aids, lexical compilation, terminology extraction, identifying poor translations, etc.). For example, Table 2 gives results of translation spotting for English equivalents of the French word apporte. One can see that a single word often corresponds to a complex phrase (and this potentially being a split construction) in the other language, thus indicating the difficulties encountered in alignment tasks.

The test set of French words used in ARCADE consisted of 20 adjectives, 20 nouns and 20 verbs. These words were selected using a specific methodology in relation with the ROMANSEVAL[5] lexical disambiguation project

(Véronis, 1988) in such a way as to allow for a future study on the relation between ambiguity and translation. The test corpus consisted of the French and English parts of the JOC corpus, containing questions asked by European Commission parliamentary officials on varying subjects (environment, industry, education, international politics , etc.) and the respective answers. The corpus contains approximately 1.1 million words per language and the test words are found in slightly more than 3700 example tokens.

These 3700 examples were manually aligned with their translation equivalents by 2 different human annotators. Due to the fact that decisions are often difficult to make, an annotation manual had been prepared to help annotators to deal with the main difficult points in a consistent manner. Inter-annotator agreement was fairly good: depending on the category, there was 93 - 98 % for the source language (French) and 84 - 93% pour the target language (English[6]). The category of verbs caused the majority of difficulties due to the fact that verbs are often embedded in complex phrases that correspond to each other as a whole.

Table 3: F-measure

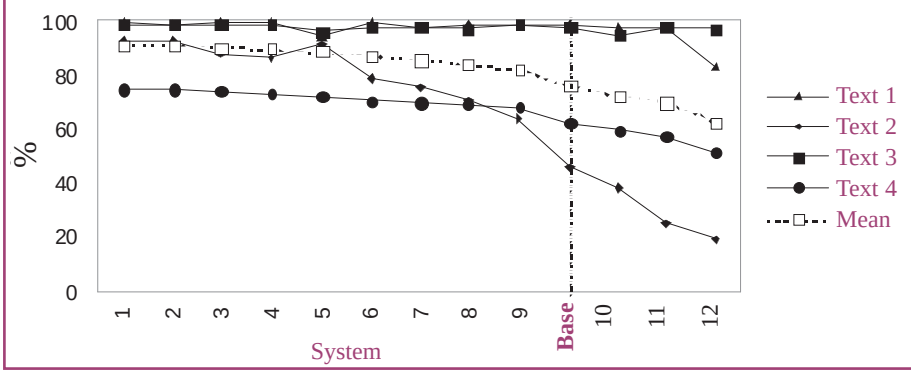

Table 2. Examples of translation spotting

French	English
La BERD apporte une contribution supplémentaire […]	The EBRD brings an additional contribution […]
Le même numéro de cette revue apporte de nouvelles précisions sur des initiatives prises par des entreprises japonaises […]	The very same issue, however, contains new information regarding initiatives by Japanese companies […]
La Communauté européenne apporte une aide aux réfugiés palestiniens depuis 1972.	The European Community has been providing assistance to Palestinian refugees since 1972.
La Commission pourrait-elle apporter des éclaircissements sur sa position […]	Can the Commission clarify its position […]
Une réunion, qui s'est tenue à Bruxelles […] a permis d'accentuer l'effort pour apporter des éléments concrets de réponse aux préoccupations exprimées par l'honorable parlementaire.	A meeting held in Brussels […] went a long way towards meeting the concerns expressed by the Honourable Member.

A total of five systems were evaluated in this test (Table 1), and the results were analyzed with respect to precision and recall measurements as well as the F -measure. In cases of disagreement between the two human annotators, we computed the evaluation score with respect to the manual alignment that offered the best match with the system. Similarly for sentence alignment, one notices a significant level of variation among the systems. System performance also varies according to grammatical categories: the best system obtained a coefficient F of 84% for adjectives 76 % for nouns and 65 % for verbs, with an average of 74% for all three categories (Table 3).

5. Conclusion

Evaluation efforts completed until now for the ARCADE project certainly have a few limitations. In particular, the systems have been evaluated by limited tasks that do not reflect the full capacity of these systems. It is evident that other characteristics (processing speed, ergonomics, robustness, etc.) can also be as important as recall and precision for practical applications. In addition, only one language pair was tested (i.e., French - English) whereas it is well-known that other languages (especially non-European languages) pose specific and difficult problems on this level.

The project has however allowed for an important methodological progress in the field, as much for the strategies of compiling an aligned reference corpus as for evaluation protocols and metrics. In addition, it gives a reliable snap-shot of the technical status of alignment processing. As for sentence alignment, we can say that the techniques are satisfactory for texts whose structure is quite parallel. Although some systems clearly need improvement, the better systems that were tested have higher than 97% accuracy. On the other hand, there is a sharp decrease for texts that do not match perfectly at the structural level (i.e., missing fragments, word order difference, etc.), and it seems to be an important direction to pursue with regard to system robustness. As for word alignment, the evaluation revealed that research in this field is far from perfect: the best system tested only attained 75% accuracy on a fairly simple task (i.e., translation spotting) (Table 3). The high demand for research in this field (creation of multilingual lexica, etc.) should therefore push for rapid progress in the near future.

Bibliography

Mariani, J. (1998). The Aupelf-Uref evaluation-based language engineering actions and related projects. Proceedings of First International Conference on Language Resources and Evaluation (LREC), 28-30 May 1998 (pp. 123-128). Granada, Spain.

van Rijsbergen, C. J. (1979). Information Retrieval. 2nd edition, London: Butterworths.

Véronis, J. (1998). A study of polysemy judgements and inter-annotator agreement. Programme and advanced papers of the Senseval workshop, 2-4 September 1998. Herstmonceux Castle, England.

Notes:

1. http://www.lpl.univ-aix.fr/projects/arcade

2.arcade@lpl.univ-aix.fr (please contact Jean.Veronis@lpl.univ-aix.fr for subscription information)

3. Performed by Philippe Langlais who also coordinated discussions on this subject between the participating teams.

4. F-measure (van Rijsbergen, 1979).

5. ROMANSEVAL is the Romance Language part of the SENSEVAL project.

6. We have applied the Dice coefficient to words suggested by each annotator.

Jean Véronis
Université de Provence
29, av. Robert Schuman
13621 Aix-en-Provence Cedex 1
France
Tel: +33 (0) 4 42 95 31 35
Fax: +33 (0) 4 42 59 50 96
Email: Jean.Veronis@lpl.univ-aix.fr

Call posted:
8 February 1999

ELRA 1999 Ca
ELRA Commissioning Produ

1. Presentation of the Call for Proposals

1.1. Introduction

The European Language Resources Association (ELRA) invites proposals for the first of a series of calls for the (co-)production and packaging of language resources (LRs), open to companies and academic organisations that comply with eligibility conditions provided below.

1.2. Purposes of the Call

ELRA is planning to commission the production, packaging and/or customisation of LRs needed by the Language Engineering (LE) Community, and is inviting applications for production and/or packaging/repackaging projects, which could be eligible for funding from ELRA.

The purpose of the call is to ensure that necessary resources are developed in an acceptable framework (in terms of time and legal conditions) by the LE players. This call is targeted towards projects with short time scales (projects lasting up to one year but preferably shorter) and the size of the funding will be considerably small. The ELRA funding is to be seen as effective and useful for producers being both tactical in their aims for the targeted market, which means that they do know all about the needs on the specific market, and strategic with regard to what to produce in order to fulfil these needs.

The resources selected for funding must be in demand on the market and the resources should preferably be easy to produce, without any technical controversies involved. From recent market monitoring, ELRA has identified several key speech and written resources. ELRA has categorised and prioritised this set of resources as indicated in Annexes 1, 2 and 3. Proposals for other types of corpora will also be considered for funding if the resources are necessary for the development of a class of LE applications. For such cases, sufficient evidence of need of such corpora and a very detailed business/exploitation plan are essential and should be submitted in annex to the proposal.

All proposals will be screened by a review committee that consists of the ELRA Board members, a few appointed external experts, and European Commission (DGXIII - Human Language Technologies sector) representatives.

2. Submission information

2.1. Eligibility requirements

In order to qualify for funding, the institution must have been eligible for funding under the 4th FP of the European Commission. The institution(s) making the proposal must belong to one of the European Union Member States, or be in an associated country.

2.2. Legal Aspects

For proposals that are awarded an LR production contract, please note that contracts established between ELRA and an LR provider grant distribution licenses by the provider to ELRA. In other words, the purpose of the contract is for the provider to supply the LRs and to receive payment, royalties or other compensation in return. ELRA agrees to distribute the LRs and grants its users (i.e., member and non-member customers) the right to use them, in full or in part, for research and/or commercial purposes, at the user's institution or site, as defined in the agreements between ELRA and the provider and between ELRA and the user. The funding and production/packaging may be set up in different ways. In all cases, ELRA will be granted the non-exclusive rights to distribute the data to potential customers. In cases of total ELRA funding for production and packaging, ELRA would become the owner. In cases of packaging/repackaging of existing corpora or co-production or resources, the ownership and royalty payment issues are to be negotiated between ELRA and the LR provider.

The contract between ELRA and the users grants the latter a non-exclusive, non transferable right to use, rework and build on the LRs within the user's institution for the purposes agreed upon between the user and ELRA. To this extent the user is allowed to create derivative works or software from the LRs or any component of them.

2.3. Selection criteria

The items listed below are among those considered as selection criteria for the evaluation of proposals. It is not easy to itemize all possible criteria. The most important factor is the fulfilment of the requirements of the call by proposing the production of LR in an efficient and cost-effective manner. Criteria include, but are not limited to: standardisation and evaluation adherence; quality; documentation and exploitation; cost-effectiveness; class of applications for reuse of data; partnership and collaborative work; project management (i.e., identifiable milestones, project baseline and specifications); evidence of market watch (i.e., strategic focus, added value for the European Union).

2.4. Budget

It is preferred that proposals not exceed 100,000 ECU. ELRA hopes to fund several proposals. For all proposals, please modularize the project by providing detailed information on different self-contained components of the overall product to be created. Each module should contain definite deliverables. In this way, it is possible for the selection committee to consider partial components of excellent proposals that exceed budgetary constraints and/or to consider co-funding options.

3. Submission procedures

3.1. Overview

Candidates should respond to the call by submitting a proposal, written in English or French, that is composed of the following elements:

1. Proposal summary: 2-page maximum
2. Detailed proposal description: 5-page maximum
3. Budget and project planning overview: 2-page maximum
4. Exploitation and Business plan: 10-page maximum

ll for proposals
uction of Language Resources

3.2. Detailed description of sections of the proposal

1. Proposal summary (one to two pages). Preferably in English.

1) complete contact information of candidate;

2) description of the resource to be produced or packaged. (e.g. Is it a new resource, an enhanced resource, packaging or repackaging of an existing resource?);

3) general budget plan;

4) experience of the proposing organization in the field;

5) exploitation plan (how will the LR be used, by whom, how) and Business plan;

6) distribution and ownership conditions.

2. Detailed proposal description (up to 5 pages)

1) clear and detailed description of the data to be produced, how it is intended to be used, and by whom it could be used;

2) statement on why the data and the production is of importance for your company/organisation and to the Language Engineering community;

3) if it is to be incorporated in any applications or used for development of any applications;

4) statement justifying why ELRA should support the production;

5) a list of experience and related skills of the participants of the team;

6) detailed description of how the production will be conducted; elements of production and production phases, including detailed time estimates for the entire production process, specifying all different phases of the production;

7) statement on how the LR will adhere to existing validation criteria or will follow other validation criteria (please enumerate). For more information on ELRA Validation manuals, see the following website: http//www.icp.grenet.fr/ELRA/validat.html

3. Budget and project planning overview (up to 2 pages)

1) a breakdown of the costs estimated for the entire production process. Specify the cost effectiveness of the production, estimate the price of the final product and the return on investment;

2) clear milestones and deliverables must be indicated;

3) duration of production project for a maximum of 12 months (preference for 9 months).

4. Exploitation plan and detailed business plan

(how will the LR be used, by whom) (up to 10 pages) -- For proposal themes not listed in Annexes 1, 2, and 3 of this call, the exploitation and business plans should be very detailed.

It is necessary to provide the following information;

1) evidence of market need for the proposed LR (potential buyers);

2) indication of exploitability of LR;

3) indication of portability of LR to various applications.

3.3. Timetable of deadlines

· Circulation of the Call: 8 February 1999

· Submission deadline for proposals: 19 March 1999

· Notification of reception of proposals before 26 March 1999

· Acceptance notifications and negotiations to start on the 5th April 1999

3.4. Additional provisions

* Only complete proposals will be reviewed. Should you have further questions, please contact Jeff ALLEN at the ELDA/ELRA office for details before 1 March 1999.

* All information submitted with proposals will be regarded as confidential and will only be used in the context of this project.

* This call is an initial step towards the production of LRs, and ELRA reserves all rights to select the projects which will be qualified for funding.

4. NO OBLIGATION TO AWARD THE CONTRACT

ELRA shall be under no obligation to award contracts pursuant to this call for proposals. ELRA shall not be liable for any compensation with respect to candidates whose proposals have not been accepted. Nor shall it be so liable in the event of its deciding not to award contracts.

5. RELATIONSHIP BETWEEN ELRA AND THE EUROPEAN COMMISSION

It is of paramount importance to highlight that ELRA is not taking over the role of the European Commission or that of national agencies involved in the strategic, long-term creation of resources and infrastructures. This is not ELRA's mission and its funding does not allow to do so. ELRA's activities fits in the frame of DG XIII actions for revitalising the LE field.

The Community of European Countries (CEC) supports projects with substantial global and generic goals. ELRA will contribute to packaging and customising small sets of key resources that would not be supported in the framework of the LE program, but which nevertheless are crucial and ready to be embedded in LE systems, to help LE players in developing new systems. The fundings from CEC are of substantial size, while ELRA intends to devote only small amounts for this process, a light weight performance in comparison to the heavy weight actions of the CEC. The CEC projects are launched and viewed on set moments in time and the time-scales are usually extended over a period of time which consists of several years. ELRA calls are targeted at short time projects (less than one year).

Contact for enquiries and submission of proposals

Jeff ALLEN c/o ELRA/ELDA
55-57, rue Brillat-Savarin
75013 Paris - FRANCE
Tel: (+33) 1 43 13 33 33 - Fax: (+33) 1 43 13 33 30
Email: jeff@elda.fr
http://www.icp.grenet.fr/ELRA/callpr99.html

ELRA Call for proposals - Preference lists

Speech Language Resources (SLRs)

1. SpeechDat like database

The SLR should contain a language or/and an application area (fixed, mobile, car) not yet covered within the SpeechDat family. Number of speakers 1000-5000. See the SpeechDat website (http://www.phonetik.uni-muenchen.de/SpeechDat.html).

2. Speech database for embedded systems

These are recordings (16kHz sampling) in a 'Handheld' (Handy, PDA, Toy, household) environment. This environment is noisy. As these devices are personal, speaker adaptation techniques could also be applied. Thus, some recordings should be done in order to investigate adaptation techniques (ranging from a total of 500-1000 speakers). Many companies currently active in the telecom area are now also looking for the market of embedded systems, because this is an attractive emerging market. Computer chip manufacturers are also looking in this area.

3. Pronunciation lexica

Pronunciation lexica should be designed for speech recognition and speech synthesis. Two alternatives can be considered:

A. a pronunciation lexicon that covers the most possible extent of proper names (first and last), street and city names (as well as major important location names and places), and covering directory assistance applications.

B. a pronunciation lexicon that adds a phonetic/phonemic layer to the basic lexica produced within PAROLE project.

It is important to consider customization of such pronunciation lexica to include pronunciation variants. These are not only relevant as commercially attractive SLRs but are also important from the perspective of phonological and phonetic research. Variants of phonemic transcriptions should touch dimensions including: speech style (formal, informal), regional accent, and perhaps word context. This type of information will stimulate research into topics such as which pronunciation variant of a word is used under which conditions (e.g., phonetic, phonological, lexical, syntagmatic, semantic, pragmatic, sociolinguistic, etc.). Such rich lexicons allow analyses of large SLRs and scan them for such variants.

4. Dialog corpus

The availability of oral dialog corpora is very important at the present time for conducting dialog studies, as well as oral dialog systems development and evaluation, even if dialog evaluation is still an open issue. Dialog corpora would be of interest for both the speech and NL scientific communities. Annotation could comprise word transcriptions, meaning, dialog acts, even prosodic information. Recommendations for transcription may be found at the MATE (http://mate.mip.ou.dk/) or DISC (http://www.elsnet.org/disc/) sites.

5. Enriching existing SLRs in terms of phonemic segmentation, prosodic annotation, word class annotation (both text and lexicon)

In order to conduct proper research on databases, additional annotation to orthography and background noise is needed most of the time. This additional information serves research into speech production and speech synthesis. For example, a reliable phonemic segmentation in the form of label files is needed for research into durations of speech units, but also for tailoring the durations in a speech synthesis system. Similarly, there is an obvious need for prosodic annotation and word class information for many SLRs to make them valuable for research purposes. This type of enhancement should be distinguished from error correction when updating databases, although updated releases with error corrections could feature such additional information.

6. Multilingual speech synthesis database

A large (few hours) speech database recorded in adequate conditions by a small set of speakers (e.g. 1 male and 1 female) which would be useful for multilingual segmental speech synthesis.

Written Language Resources (WLRs)

1. Large monolingual corpora

ASCII and Unicode text are the basic text type standards. Corpora with SGML , HTML, or XML markup is preferred. Part or all of a given large monolingual corpus should contain Part-of-Speech or other syntactic annotation following recognized standards (see EAGLES - http://www.ilc.pi.cnr.it/EAGLES/home.html). It is possible to have a corpus with different levels of annotation (see MULTEXT-http://www.lpl.univ-aix.fr/projects/multext/). General and/or specific domains as well as single or multi-genre domain corpora will be considered. Preference for production of non-newspaper corpora, except for cases where the given written source in a language has not yet been developed into a distributable LR.

2. Parallel texts

ASCII and Unicode text are the basic text type standards. Corpora with SGML , HTML, or XML markup is preferred. Part or all of a given large monolingual corpus should contain Part-of-Speech or other syntactic annotation following recognized standards (see EAGLES - http://www.ilc.pi.cnr.it/EAGLES/home.html). It is possible to have a corpus with different levels of annotation (see MULTEXT-http://www.lpl.univ-aix.fr/projects/multext/). General and/or specific domains as well as single or multi-genre domain corpora will be considered. Parallel texts should be aligned at various levels for optimum porting from one application to another. Preference for proposals that demonstrate the use of such corpora for multiple applications.

3. Bi/multilingual computational lexica

Such lexica should contain detailed linguistic information about syntactic characteristics (i.e., word class, word-class specific subcategorisation, complement structures) and possibly semantic characteristics (e.g., argument structures). They could also include proper names and proper nouns. Number of lexical entries per language should be comparable to or larger than other existing resources (see http://www.icp.grenet.fr/ELRA/home.html).

Multimedia and Multimodal Language Resources

Multimedia and multimodal corpora are growing in demand for current and future research and development. The following descriptions are examples of potential corpora to be produced and packaged. The examples cited below should not be considered as constituting an exhaustive list of possibilities.

1. Multimedia corpus

A multimedia corpus may contain data corresponding to radio or TV broadcast news, comparable to what is used within the DARPA/NIST Human Language Technology programs (see http://www.itl.nist.gov/div894/894.01/). Transcriptions can be conducted by using, preferably, the Transcriber tool freely available through DGA (France) at: (http://www.etca.fr/English/Projects/Transcriber/). Languages should be distinct from American English. Speech, audio other than speech, text and visual information, if applicable, should be considered.

2. Multimodal corpus

A multimodal corpus should comprise audio speech or textual data together with other kinds of data, such as visual data, or gestual data. Multimodal corpus annotation is still an open issue. However, useful information can be found at the CAVA (Computer Assisted Video Analysis) site (http://www.mpi.nl/world/tg/CAVA/CAVA.html), or at the "Talking Heads" website: http://www.haskins.yale.edu/haskins/heads.html.

Finnish national HLT-programme efforts

Manne Miettinen, CSC

For the past two decades Finnish researchers have been innovative and successful in developing internationally well known methods for processing human language on computers. The best known examples are constraint grammar (CG), self-organising map (SOM) and two-level morphology (TWOL). Enterprises commercialising these methods have also been successful in international competition. Traditionally, the research groups have only had occasional contact to other Finnish research groups. Co-operation has also been difficult because the research groups are scattered in different universities and faculties across the country.

This situation is less than ideal in today's research environment which emphasises large-scale co-operative projects. The dispersion of the human language technology (HLT) research groups also contributes to the rather shallow image of HLT in the eyes of the Finnish business world, funding agencies and the general public.

The Finnish government has taken the decision to increase public funding for R&D to 2.9% of GNP in the year 1999. Multimedia and information society technologies are prominent research areas in on-going large-scale national programmes, but HLT is almost omitted from these programmes. Unlike many other European countries, Finland has not had an HLT research programme yet.

These facts are some of the key arguments in a recently completed project intended to motivate the national funding agencies, notably the Technology Development Centre (funded by the Finnish Ministry of Trade and Industry) and the Academy of Finland (funded by the Finnish Ministry of Education), to launch a large scale national HLT R&D programme in Finland.

The project was carried out by CSC - Center for Scientific Computing - an organisation owned by the Finnish Ministry of Education that specialises in computational science. The project consisted of editing a report on the state of the art of HLT in Finland, organising a one day seminar on HLT in Finland and submitting a concrete proposal for a national HLT programme to the funding agencies.

The project was surveyed by a steering group that was headed by Professor Kimmo Koskenniemi of the University of Helsinki. The other ten members were representatives of the Ministry of Education, the Technology Development Centre, the Academy of Finland, Nokia Research Center, HPY Research Center, Sanoma Oy, Alma-Media Oyj, Tieto Corporation, Helsinki University of Technology and Promentor Solutions Oy.

The report "Kieliteknologia Suomessa" (Human Language Technology in Finland) was published on 11 June, on the occasion of the one day seminar on HLT in Finland, which was attended by over hundred interested people. The guest speaker for the seminar was Giovanni Battista Varile from Language Engineering Unit of the European Commission. The seminar was successful in mobilising HLT researchers, developers and enterprises interested in using HLT in their products. Since the seminar, participants have been kept informed by HLT-dedicated mailing lists and Web site (http://www.csc.fi/kieliteknologia).

After the quiet summer months, the steering group gathered in September to finish the proposal for the establishment of a national HLT programme. The proposal was submitted to the Technology Development Centre on 2 October.

The proposal outlines seven broad research areas:

1) Document management
2) Translator's tools
3) Computer assisted language learning
4) Natural language interfaces
5) Speech signal processing
6) Shared language resources
7) Writer's tools

and suggests a budget that gives highest priority to speech signal processing, which is currently the least studied area of HLT in Finland compared to international activity in this area. The suggested priority ranking for the other areas are: document management, computer assisted language learning, translator's tools, natural language interfaces, shared language resources and writer's tools.

The proposal is currently going through an internal evaluation process within the Technology Development Centre and is being considered as one possible technology programme to be launched in 1999.

Manne Miettinen
CSC - Tieteellinen laskenta Oy
Tietotie 6
PO BOX 405
FIN-02101 Espoo
Finland
Email: Manne.Miettinen@csc.fi

Building a "Tri-Text": Steps in the Conversion of a Hard Copy Document to an On-line Resource

Lisa Hale Decrozant, University of Maryland and Clare R. Voss, Army Research Laboratory

Introduction

As researchers tasked with evaluating machine translation (MT) tools for military linguists in the field, we must often work with "less commonly taught languages" (LCTLs) for which little readily available on-line text exists. While many linguistic resources needed for MT evaluation are commonly found in electronic form for the major languages of commerce (English, French, Japanese, etc.), this is typically not the case for LCTLs[1]. In this brief note, we describe our recent effort transforming hardcopy parallel, sentence-aligned text into on-line form.

The form of the particular document we worked with is uncommon: it contains parallel text in three languages-Haitian Creole, French and English-hence the name "tri-text". For MT evaluation, having all three languages aligned this way has provided us with a way of comparing the strengths of different language pairs on the same MT system platform. We are also able to use the phrasebook's sentences as our test collection both in evaluating different MT engines and in developing language learning tools for on-line use (Decrozant & Voss, 98). In the description that follows however we cover only those steps in working with this linguistic resource that will be relevant to other researchers with similar low-quality, hard-copy paper documents.

Choosing an OCR product

The first step was to choose an OCR pro-

gram which could support the unique character inventory of Haitian Creole. The language of Haiti's population is phonologically similar to French, but does not share its exact character set. Of the languages supported by OCR programs available to us, the French character set provided the closest match to the Haitian character set. An analysis of several Haitian Creole documents showed that one character, the [ò], is the only character that exists in the Haitian Creole writing system, but not the French. Therefore, we knew this character was not in the OCR training set and we expected that its recognition would be problematic. Given the two OCR packages available to us, we tested how each program handled this non-French character in order to decide which one would provide the most consistent, predictable (though incorrect) character recognition (Schlesiger & Decrozant, 98). The results of our OCR pilot experiment showed that one program, Cognitive Technologies Cuneiform, quite consistently turned [ò] to [à], while the other, Caere Omnipage Pro, turned [ò] into either [o], [ô], or [è]. A consistent error is clearly simpler to correct during the ground-truthing phase (i.e. reconciling the OCR output with the original text), therefore we decided to use Cognitive Technologies' product to convert the phrasebook into on-line parallel text.

Pre-OCR document image enhancement

Both the quality of the paper and the printing in our original document was poor. Such low-quality paper is thin and prone to bleed-through, where ink printed on one side of the page permeates the paper and appears unevenly on the other side of the page. The low-quality printing also results in uneven typeset and speckling, where tiny splatterings of ink appear in addition to the letters. These phenomena wreck havoc on OCR programs causing errors throughout converted documents. Therefore, we decided to test different ways of reducing imperfections in the document image during a photocopying step. (Since ground-truthing is so labour-intense and an unpleasant a task, we decided it was worth spending the relatively short amount of time to photocopy the full document and get a less error-full copy.)

We systematically tested several combinations of copier settings to determine which resulting copies OCR-ed with the fewest errors. Surprisingly, by reducing the size of the copy (and thereby the size of the font on the document) by about 25%, we found that the OCR program could recognise characters with greater success[2]. We found that we were also able to "clean up" some of the speckling and bleedthrough by adjusting the copy density: by adjusting the copy density to a lighter setting, we reduced these imperfections. After establishing optimal copier settings for the selected OCR product, we copied the full tri-text document under those settings in preparation for the next processing step.

Overview of process

After photocopying each page of the full tri-text document, the resulting copy was scanned into an image file. Since the individual pages from the phrasebook were printed in three columns, representing the three-way language breakout, we divided the scanned-in page image into thirds, creating language-specific image files. Once run through the OCR program, the resulting "recognised" text was ground-truthed, i.e. thoroughly checked and compared to the original document for recognition errors. Ground-truthing is more efficient when performed by someone who is familiar with the languages involved. This proved to be the most labour-intensive stage of the entire document conversion process.

To recap, from hardcopy to on-line, this process required six steps:

1. Determine best-match OCR program for the low-density language.
2. Determine best copy settings to enhance document image, given the OCR program selected in step 1.
3. Scan document copied in accordance with settings established in step 2.
4. Create language-specific scanned image files.
5. Perform OCR on individual images.
   a. French language OCR for French and Haitian Creole documents.
   b. English language OCR for English documents.
6. Ground-truth all documents for typos and recognition errors.

The result of step 5, an online parallel corpus, required extensive online editing: we encountered many typos and inconsistent spellings in the original document in step 6. This result was to be expected, given the lexical variation in Haitian Creole that has been extensively documented by Allen (1998).

We are currently using this corpus in two applications: MT system evaluation and Language Training (Decrozant and Voss, 1998, 1999). Our Haitian Creole-English MT system, FALCon, needs to be assessed as we receive new versions of its embedded MEMT engine. The corpus provides us with a test suite for effectiveness (MOE) evaluation in a filtering task. (For a discussion of performance (MOP) evaluation methodology for MEMTs, see Hogan and Frederking (1998).) We have also begun testing STARLing, language maintenance software, to assist in the cross-training of French military linguists working with Haitian-Creole documents. At the users' direction, the concordancer and look-up tools on STARLing retrieve French and Haitian Creole aligned sentences from the tri-text to supplement their understanding of documents translated by FALCon.

The end result of this process is a resource of parallel text in a workable, on-line form. The ability to make changes electronically to the text is important. We found this to be true as we encountered many typos and inconsistent spellings in the original document which we were able to edit. Once in on-line form, parallel text such as this becomes extremely valuable for NLP applications such as MT system evaluation, MT system augmentation and on-line language learning.

References

Decrozant, L. and Voss, C. Cross-linguistic Resources for Language Training and MT Evaluation. In proceedings of Natural Language Processing and Industrial Applications 1998, Moncton, NB Canada. pp 91-95.

Kanungo, T. (1998) Personal Communication.

Schlesiger, C. and Decrozant, L. Comparison of Two French OCR Packages-How They Handle Haitian Creole Text. Unpublished Technical Report, Army Research Laboratory, 1998.

Notes:

1. The LCTL we discuss here is thus a "low-density" or a "low-diffusion" language, in that few linguistic resources are available on-line.

2. We later found out that OCR models are typically trained on images in discrete, incremented sizes and thus perform best when presented with images at those trained sizes. (Kanungo, 1998). Thus, for our document and that OCR model, the closest trained size with the best recognition was smaller than the document's actual font size.

Lisa Decrozant
University of Maryland, Linguistic Dept
College Park, MD USA
Email: ljhale@wam.umd.edu

Clare R. Voss
Intelligent Systems Branch
Army Research Laboratory
Adelphi, MD - USA
voss@arl.mil

New Resources

ELRA-S0062 Fixed1itDesign

With a view of supplying SpeechDat family projects with the textual material used in the Italian speech database, CSELT has decided to produce a CD-ROM with all database specifications including the full list of a designed corpus: a set of phonetically rich sentences and a set of application oriented utterances.

The Italian SpeechDat databases (produced in the framework of SpeechDat(M) and SpeechDat(II)) used this textual material.

The SpeechDat common specification totals 40 utterances per call, comprising a mixture of spontaneous and read speech. The purpose of each telephone call was to record the basic structure of the utterances mentioned below. All utterances are read speech unless marked as spontaneous.

The list of utterances is as follows:

- 3 application words;
- 1 sequence of 10 isolated digits;
- 4 connected digits: 1 sheet number (5 digits), 1 telephone number (9-11 digits), 1 credit card number (14-16 digits), 1 PIN code (6 digits);
- 3 dates: 1 spontaneous date (e.g. birthday), 1 prompted date (word style), 1 relative and general date expression;
- 1 word spotting phrase using an application word (embedded);
- 1 isolated digit;
- 3 spelled-out words (letter sequences): 1 spontaneous (e.g. own forename), 1 spelling of directory city name, 1 real/artificial name for coverage;
- 1 money amount in Lire;
- 1 natural number;
- 5 directory assistance names: 1 own forename (spontaneous), 1 city of birth/home town (spontaneous), 1 most frequent city, 1 most frequent company/agency, 1 "forename surname";
- 2 questions, including "Fuzzy" yes/no: 1 predominantly "yes" question, 1 predominantly "no" question;
- 9 phonetically rich sentences;
- 2 time phrases: 1 time of day (spontaneous), 1 time phrase (word style);
- 4 phonetically rich words.

In the case of the Italian fixed network database, four additional items were added to the one designed in the project:

- 1 telephone area code
- 1 money amount in EURO
- 2 "yes/no" questions

For the Italian SpeechDat corpus, the full list of items is supplied in two different files: the first contains the prompted text read by speakers in the supplied sheet and the second file contains the orthographic transcription.

Statistics are supplied for each corpus, which are computed on the repetition of digits, letters or phonemes (diphones and triphones) depending on the corpus type. These statistics are reported in a separate file for each corpus.

A documentation file aiming at describing the entire corpus design is included on the CD-ROM. It also covers the motivations that lead to that particular design.

Finally, a complete lexicon file (in SpeechDat format) is supplied.

The CD-ROM does not contain any recordings.

KEY FEATURES

Type of resource:	Textual material	Language:	Italian
Domain/Source:	Textual material used within the Italian SpeechDat(M) and SpeechDat(II) databases	File format:	ASCII
		Distribution media:	1 CD-ROM
Related resources:	Italian SpeechDat(M) database (ELRA-S0052 and S0053)		

Price for ELRA members:	for research use: € 2,000	for commercial use: € 3,000
Price for non members:	for research use: € 5,000	for commercial use: € 5,000

ELRA-S0064 Colombian Spanish Speech Database

This database contains speech collected from Colombia. Collection was performed at Siemens Colombia and processed at the Department of Signal Theory and Communications of the Universitat Politècnica de Catalunya (UPC) (Spain).

This database is comprised of telephone recordings from 1,065 speakers (563 males speakers and 502 female speakers) recorded directly over the fixed telephone network using an E-1 interface. The recording platform used an ISDN basic access (BR1) interface.

Speech files are stored as sequences of 8-bit 8 kHz A-law uncompressed speech samples (CCITT G.711 recommendation). Each prompted utterance is stored within a separate file. Each speech file has an accompanying ASCII SAM label file. Speech file format and SAM label files follow the specifications given by the SpeechDat project.

The speakers were mainly recruited from Siemens personnel, students from several Colombian universities, and their relatives. The following sex and age distribution has been obtained: 56 speakers are under 16 years old (38 males, 18 females), 542 speakers are between 16 and 30 (277 males, 265 females), 347 speakers are between 31 and 45 (178 males, 169 females), 99 speakers are between 46 and 60 (59 males, 40 females) and 21 speakers are over 60 (11 males, 10 females).

The transcription included in this database is an orthographic transcription with a few details that represent audible acoustic events (speech and non speech) present in the corresponding waveform files. A lexicon is also provided.

Non-Speech Acoustic Events have been arranged into 4 categories (filled pause, speaker noise, stationary noise and intermittent noise) and are transcribed.

KEY FEATURES

Type of resource:	Speech recordings (Acoustic)	Speech mode:	Read
Recording conditions:	ISDN telephone interface	Language:	Colombian Spanish
Sex and number of speakers:	1,065 speakers (563 males and 502 females)	Linguistic annotation:	Orthographic (+ transcription of audible noises)
File format:	8 bits, A-law	Standard in use:	SAM
Sampling rate (kHz):	8 kHz	Distribution media:	1 CD-ROM
Related resources:	SpeechDat family. Other languages available.		

Price for ELRA members: € 5,000 Price for non members: € 7,500

ELRA-S0067 BREF-120 - A large corpus of French read speech

BREF-120 resulted from the efforts of LIMSI-CNRS researchers under sponsorship from the GDR-PRC CHM, the ACCT (OFIL), the EEC (ESPRIT Polyglot project), and the Aupelf-Uref.

A sub-set of BREF-120 is BREF-80 (ELRA-S0006), which consists of about 50-60 sentences per speaker and recordings conducted only with a Shure microphone. In BREF-80, the sentences were chosen to cover as many prompts as possible.

The BREF-120 corpus was designed to provide read speech data for the development and evaluation of continuous speech recognition systems (both speaker-dependent and speaker-independent), and to provide a large corpus of continuous speech for the acquisition of acoustic-phonetic knowledge of spoken French.

BREF-120 is a large read-speech corpus containing over 100 hours of speech material, from 120 speakers (55 males and 65 females). The text materials were selected verbatim from extracts of the French newspaper "Le Monde". Each of 80 speakers read approximately 10,000 words (about 650 sentences) of text, and another 40 speakers each read about half that amount. Simultaneous recordings were made in a sound-proof room using a Shure SM10 microphone and a Crown PCC160 microphone and were monitored to assure their contents. The speech signal was sampled at 16 kHz and digitised with 16 bits. The BREF-120 corpus contains 28 CDs; numbers 1-13 contain the Shure recorded data and numbers 14-28 contain the Crown recorded data.

KEY FEATURES

Type of resource:	Speech recordings (Acoustic)	Speech mode:	Read
Recording conditions:	Sound-isolated room		
Microphone/Telephone type:	Two microphones: a Shure SM10 and a Crown PCC160	Language:	French
Domain/Source:	French newspaper "Le Monde"	File format:	16 bits
Sex and number of speakers:	120 speakers (55 males and 65 females)	Linguistic annotation:	Orthographic
Size (hours, vocabulary):	100 hours of speech	Standard in use:	SAM
Sampling rate (kHz):	16 kHz		
Distribution media:	28 CD-ROM; numbers 1-13 contain the Shure recorded data and numbers 14-28 contain the Crown recorded data		
Related resources:	BDLEX (ELRA-S0003 and S0004), BREF-80 (ELRA-S0006), BREF-Polyglot (ELRA-S0007).		

Price for ELRA members:	Research use:	€ 2,500	Commercial use:	€ 10,000
Price for non members:	Research use:	€ 4,000	Commercial use:	€ 15,000

ELRA-S0065 Spanish SpeechDat(M) - DB1

(Phonetically rich sentences & application oriented utterances such as keywords, digits, etc.)

The SpeechDat(M) Spanish database is comprised of telephone recordings from 1002 speakers (508 male speakers and 494 female speakers) recorded directly over the fixed telephone network using an E-1 interface at the recording site. There is also a pronunciation dictionary for the correctly spoken items. It was produced by a collaboration involving Vocalis Ltd and Universitat Politècnica de Catalunya (UPC) within the SpeechDat(M) project. Vocalis had responsibilities for the general Speechdat specification, for the recording site, platform and tools, and overall database production and coordination. UPC was responsible for the detailed content design, speaker selection and coordination, pronunciation dictionary, orthographic transcription of the utterances, and documentation.

It was agreed that the ESPRIT Project SAM standards be followed for speech file storage. Speech files are stored as sequences of 8-bit 8 kHz A-law speech samples (before compression). 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.

All utterances are read speech unless marked as spontaneous. The list of items is as follows:

- 1 isolated digit;
- 4 connected digits and numbers: 4-digit id/sheet number, 9-digit telephone number, 16-digit credit card number, 1 home telephone number (spontaneous), 2 natural numbers;
- 1 natural number with decimal point;
- 2 money amounts: 1 large amount, 1 small amount;
- 3 spelled-out words (7 letter sequences);
- 1 time of day (spontaneous);
- 1 time phrase (prompted, word style);
- 1 date (spontaneous, the speaker's birthday);
- 2 dates (prompted, word style);
- 3 yes/no questions: *Are you calling from the same province?* (as P1), *Do you speak another language fluently?*, *Are you calling from a public phonebox?*;
- 1 place (province of longest residence);
- 6 application keywords (out of a vocabulary of 54 words);*
- 2 additional application keywords (out of a vocabulary of 18 words);*
- 3 embedded application word phrases (from A1-6 vocabulary);*
- 9 read sentences for phonetic coverage.

* lists available on the Web (http://www.icp.grenet.fr/ELRA/home.html)

The set of phonetically balanced sentences was automatically transcribed and manually checked by the Department de Filologia Espanyola of the Universitat Autonoma de Barcelona. Standard Castillian transcription was used. No dialectal variations were considered.

The following age distribution has been obtained: 530 speakers are between 15 and 29 years old, 283 speakers are between 30 and 45, 156 speakers are between 46 and 60, and 23 speakers are over 60; the age of 10 speakers is unknown.

KEY FEATURES

Type of resource:	Speech recordings (Acoustic)	Speech mode:	Read (occasionaly spontaneous)
Recording conditions:	Fixed PSTN telephone network	Microphone/telephone type:	E-1 interface
Language:	Castillian Spanish	Linguistic annotation:	Orthographic
Sex and number of speakers:	1002 speakers (508 males and 494 females)	File format:	8 bits, A-law
Standard in use:	SAM	Sampling rate (kHz):	8 kHz
Distribution media:	3 CD-ROM		
Related resources:	SpeechDat(M) resources for other resources: Danish (ELRA-S0040), English (ELRA-S0011), French (ELRA-S0016), German (ELRA-S0018), Italian (ELRA-S0052), Portuguese (ELRA-S0068).		

Price for ELRA members:	Research use:	€ 11,000	Commercial use:	€ 14,000
Price for non members:	Research use:	€ 20,000	Commercial use:	€ 20,000

ELRA-S0066 Spanish SpeechDat(M) - DB2

(The phonetically rich sentences)

Sub-set of ELRA-S0065 which contains only the phonetically rich sentences without the application oriented utterances.

Price for ELRA members:	Research use:	€ 8,800	Commercial use:	€ 14,000
Price for non members:	Research use:	€ 14,000	Commercial use:	€ 20,000

ELRA-S0068 Portuguese SpeechDat(M) database

The Portuguese SpeechDat(M) database contains the recordings of 1001 calls (453 male speakers and 548 female speakers). This database was collected by Portugal Telecom in the scope of the European SpeechDat Project. The task of designing and post-processing the database (together with the documentation) was subcontracted to INESC. The design of the collection platform and the speech data collection itself was the responsibility of INESCTEL.

Each speaker uttered the following items:

- 3 natural numbers
- 2 money amounts
- 3 spelled-out words
- 1 spontaneous date
- 1 isolated digit
- 2 dates
- 3 word spotting phrases
- 1 spontaneous time
- 1 credit card number
- 1 time
- 9 sentences
- 1 region name
- 1 telephone number
- 6 application words
- 4 yes/no questions

The approach adopted for speaker recruitment involved selecting speakers among the employees of Portugal Telecom (about 20,000) and their relatives. The company has a wide geographical coverage, thus guaranteeing a good representation of many regional accents.

The following age distribution has been obtained: 12 speakers are under 16 years old, 345 speakers are between 17 and 30, 436 speakers are between 31 and 45, 196 speakers are between 46 and 60 and 8 speakers are over 60 (with two speakers to add who did not mention their age and two others who said they were born in 1996).

Speech signals are recorded at 8kHz, 8-bit A-law format. Files are stored according to the file specifications proposed in the "SpeechDat database format specification". The file formats and headers follow the SAM recommendations (header files separated from signal files). A pronunciation dictionary with a phonemic transcription in SAMPA is also included.

KEY FEATURES

Type of resource:	Speech recordings (Acoustic)	Speech mode:	Read (occasionaly spontaneous)
Recording conditions:	ISDN telephone interface	Language:	Portuguese
Domain/Source:	Sentences from the Portuguese daily newspaper PÚBLICO		
Sex and number of speakers:	1001 speakers (453 males and 548 females)	Linguistic annotation:	Orthographic
File format:	8 bits, A-law	Standard in use:	SAM
Sampling rate (kHz):	8 kHz	Distribution media:	3 CD-ROM
Related resources:	SpeechDat(M) resources for other languages: Danish (ELRA-S0040), English (ELRA-S0011), French (ELRA-S0016), German (ELRA-S0018), Italian (ELRA-S0052), Spanish (ELRA-S0065).		

Price for ELRA members	Research use: € 11,000	Commercial use: € 14,000
Price for non members	Research use: € 14,000	Commercial use: € 20,000

ELRA Application form

Organisation ..... Department .....

Name of Designated Representative.....

Address ..... Town ..... Postcode .....

Country ..... Telephone ..... Fax .....

Email: .....Web.....

College ( ) Spoken ( ) Written ( ) Terminology

Category:

( ) Non-profit-making organisations	750 EURO/year
( ) European SME of less than 50 employees	1000 EURO/year
( ) European profit making organisations of more than 50 employees	1500 EURO/year
( ) Non European profit making organisations	5000 EURO/year

( ) I agree to the information above appearing in the ELRA Directory

Signature: Date:

For information, please contact: ELRA Membership Secretariat

55-57 rue Brillat Savarin - 75013 PARIS, FRANCE

Tel : +33 1 43 13 33 33 - Fax : +33 1 43 13 33 30 - Email: jaffrain@elda.fr

Notes

1. An invoice for the membership fee will be sent upon receipt of the completed application form, and should be paid within 30 days.

2. Payment may be made by bank transfer or cheque in EURO, made out in favour of ELRA. Bank : BNP (Luxembourg) S.A, 24, Bd. Royal, L2953 Luxembourg Account n°: 63-114418-57-6102-997. Bank charges to be borne by the subscriber.

3. Membership covers the period from 1 January to 31 December of each year