The ELRA Newsletter

EUROPEAN LANGUAGE RESOURCES ASSOCIATION

October - December 99

Vol.4 n.4

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

Editor:
Jeffrey Allen

Layout:
Audrey Mance

Contributors:
Jeffrey Allen
Hitoshi Isahara
Marilyn Mason
Patrick Paroubek
Reinhard Schäler

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.icp.grenet.fr/ELRA/home.html

Dear Members,

This is the last issue of 1999. In the past, we used to report on our yearly activities in the fourth quarter issue because our former fiscal year was 1 October - 31 September. Due to the shift of our fiscal year in 1999 to align with the 1 January - 31 December calendar year, the next issue (Vol.5 n.1) of the ELRA Newsletter will provide a report on our 1999 activities as well as indicate the date and venue of our Annual General Assembly for 1999 which will very likely take place by mid-March 2000.

- This quarter has seen a steady progress of our project LRs P&P (Language Resources -- Production & Packaging). The New corpus of written Business English has collected the corpus, has converted 70% of the HTML text to text, and is completing authorization agreements with data owners. The PIVOT Project: Sets of bilingual LR dictionaries for English and Russian has produced a report on the status of the development of tools for the Russian texts that have been obtained. The Crater 2 project started in September, so a report has not yet been due, but the project is underway. The Italian Broadcast News Corpus has recently delivered 10 hours of annotated data. The Pronunciation lexicon of British English place-names, surnames and first names has transcribed 2/3 of the entries, has defined the entry formats, and is proofing the completed entries. The Scientific Corpus of Modern French has acquired the full corpus, fully tagged and lemmatized it, and initial mark-up design has started. The German-French Parallel Corpus of 30 Million words has collected 2 million words per language and has spent a significant amount of effort for obtaining permission to use copyrighted data.

- As recently announced, ELDA has conducted a number of surveys. A brief summary of the user survey work is included in this issue. ELRA members can obtain a more detailed report in the Members section of our web site (note: access to that section requires a password).

During this quarter, ELDA submitted two proposals within the European Multilingual Information Society (MLIS) programme, and both projects were accepted:

- NETWORK-DC: The project aims at providing multilingual Language Resources over global networks through the establishment of an efficient collaboration agreement between ELDA and SPEX on the European side, and the Linguistic Data Consortium (LDC) on the US side. This collaborative project includes networking and cross-agreements between these organizations for the production, acquisition, normalization, certification and distribution of spoken and written language data for research, education and technology development. It also aims at taking past and current accomplishments to a new level, by designing and implementing new modes of cooperation between these organizations.

- Gates for Enhanced Multilingual resource Access (GEMA): This project aims at providing a central and organised access point for the linguistic sector and building and developing a linguistic portal with corresponding services. The services will cover the complete range of activities, disciplines and needs of this sector and will include: on-line resource consultation services, on-line resource and tool acquisition services, information services, forum services and other value-added services.

As you know, the next ELRA conference, the Language Resources and Evaluation Conference (LREC-2000), will take place in the prestigious exhibition hall of Athens Zappeion Megaron and is to be organised by the Institute for Language and Speech Processing (ILSP). The programme committee has received over 380 proposals for technical presentations and about 14 proposals for workshops. The conference is scheduled to take place 31 May - 2 June 2000, with satellite presentations and post-workshops in addition. A preliminary programme will be posted at http://www.elda.fr/lrec2000.html as soon as possible. There will also be an exhibition and booth area at LREC-2000 for companies and industrial players to present their products and services. More information about the Exhibition can be obtained from Khalid Choukri (choukri@elda.fr) or Stelios Piperidis (spip@ilsp.gr).

Turning to this issue of the ELRA newsletter, Reinhard Schäler's (University of Limerick) article presents the Transrouter project, aiming at developing a prototype that will help translation managers to decide whether a project should be translated by human translators, translation memories or translation machines. Hitoshi Isahara, of the Communications Research Laboratory of the Japanese Ministry of Posts and Telecommunications, gives a summary of the sessions on Language Resources at the recent Machine Translation Summit VII held in Singapore in September 1999. A third article by Marilyn Mason (Mason Integrated Technologies Ltd.) describes the development of a system for orthography conversion and lexical standardisation. The final article, Patrick Paroubek (LIMSI), presents the MULTITAG project.

At the end of this issue, you will find descriptions of new resources that are now available in the ELRA catalogue. These are new SpeechDat(II) databases for Swedish, Danish, Welsh and British English. Please also note the extension of ELRA-W0015 Le Monde with the year 1998 now available, as well as the updating of the Verbmobil resources (ELRA-S0034).

We would like to take the opportunity to welcome Estelle Neyer (neyer@elda.fr) who has recently joined ELDA as the new CEO assistant.

In conclusion, the ELRA Board and the ELDA team wish all our members and partners a Merry Christmas and a Happy New Year. We look forward to working with you in 2000!

Antonio Zampolli, President Khalid Choukri, CEO

Erratum: *In the article by Xavier Garcias, "Beyond "fuzzy matching" - The Déjà Vu approach to reusing Languages Resources" (Newsletter Vol.4 N.3, p.5), the name of Mr. Garcia's company was misspelt. The correct name is Ampersand traducció Automática. Consequently, his e-mail address should also be corrected as: xavi@ampersandsl.com. We wish to express our apologies to the author and our readers.*

Transrouter - a decision support tool for translation managers

Reinhard Schäler, Localisation Research Centre, University of Limerick

Translation managers in large organisations have to decide how projects should be translated: by a human translator, by translation memory, by MT or a combination of all three. Yet, they often do not have the technical knowledge and experience necessary to make these decisions with confidence. Transrouter is a decision support tool that will provide translation managers with up-to-date information on the features and capabilities of translation technology applications and relevant linguistic resources. It will evaluate specific projects based on user input and an automatic analysis using its component tools. It will suggest possible translation routes for a particular project providing details on time, cost and quality implications.

The problem

In late 1996, Microsoft's vendor manager in Dublin invited a small group of translation technology experts from the localisation industry to discuss the implications of the success of translation memory systems for the localisation process. Translation memory systems, piloted since 1995 by just a handful of highly innovative localisation companies and successfully used first by Softrans-Berlitz and Oracle in a major localisation project, were now being introduced on a large scale. This meant that the knowledge -some call it, maybe more appropriately, *gut feeling* - accumulated by the few translation memory experts around at the time, had to be made available easily and efficiently to the non-technical translation mangers.

Therefore, the main questions discussed around the table at this meeting were:

* What questions are asked and which criteria are used by the experts when they decide whether a translation memory system should be used or not for a particular translation project?
* Can an automated system based on these questions and criteria be developed to make this expert knowledge available to the translation managers thus helping them to decide whether they should use human translators or a translation technology application?

The solution

Following a number of meetings, a translation technology workshop, plus many hours of coding and reviewing, the LRC developed the first pre-prototype of a system called ETAT. The purpose of this development was to get feed-back from potential users on the usefulness of a decision support tool.

The reaction of the localisation managers consulted was so positive that it was decided to explore ways to develop the idea (and the pre-prototype) further.

Subsequent to this a European consortium was set-up. This consortium is comprised of a number of high-profile European research organisations, commercial translation service providers and translation technology developers. Working together the consortium successfully proposed the Transrouter project under the 4th call for proposals of the Language Engineering sector of the EU's 4th Framework Programme. In addition to ETAT, findings from the EAGLES case study of work in the Commission Translation Services (SdT) also influenced the development of this proposal.

The project

The Transrouter project is lead by Berlitz, who is the world's largest provider of translation services. Other partners in the consortium are: the Localisation Research Centre (LRC) at the University of Limerick, ISSCO, the Centre for Language Technology (CST), the universities of Edinburgh and Regensburg, and Sail-Labs, L&H's strategic research organisation.

A user group that consists of a wide range of software publishers, translation technology developers, localisation vendors and translation service providers has been put in place to support the consortium (1).

The aim of Transrouter is to develop the prototype of a computer-based application that will help translation managers to decide whether a project should be translated using human translators, translation memory applications or machine translation. Transrouter will present a number of possible routes to translation managers, along with the consequences in terms of time, cost and quality of taking each route.

The suggestions by Transrouter's decision kernel is based on information stored in three types of profiles:

* *Project profiles* containing detailed data on specific projects, such as project size, source and target languages, file formats, time and budget available etc. This information will be acquired by user input and the automatic analysis of the project using the Transrouter *component tools.*
* *Agent profiles* containing detailed information on the features of translation technology applications, such as file formats and language combinations a system is able to handle, its integration options with other systems, the level of adaptability to specific user requirements etc. This information must be updated as necessary by translation technology experts or translation technology vendors.
* *Agent resources profiles* containing information on the resources available to agents such as terminology databases, translation memories, MT terminology databases etc. This information will be maintained by the in-house translation technology expert at the user site.

Users will eventually be able to 'connect' their own component tools to Transrouter. However, the consortium is also supplying some essential component tools with the prototype.

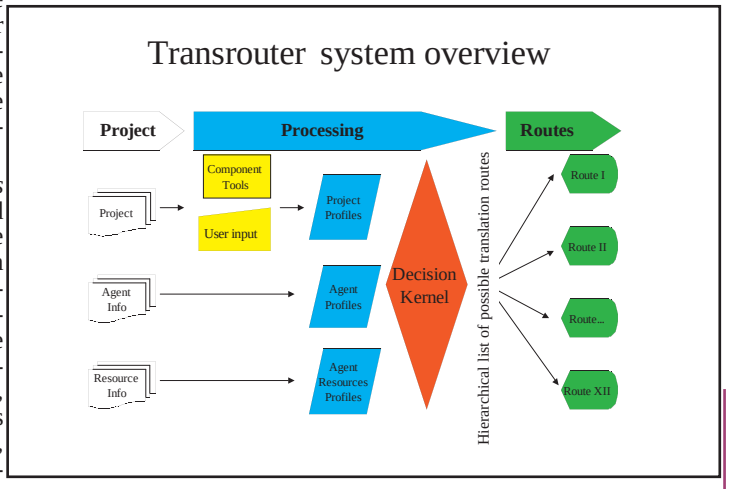
Among these are:

* Word counter and sentence length estimator
* Translation memory/previously translated texts coverage calculator
* Version comparison
* Repetitiveness detector
* Unknown term detector
* Sentence simplicity checker

The scope

While Transrouter has taken the initial *ETAT* prototype as one of its central starting points, its scope and target group have expanded substantially to include a wider range of translation technology applications in comparison to the translation memory centred approach of the *ETAT* prototype. Transrouter will also cover a broader range of subject matter in comparison to the localisation centred approach of *ETAT.*

Although localisation scenarios are still being used as a test bed for Transrouter, the benefits of the tool will outreach this industry alone. The Transrouter consortium carried out detailed studies of translation practices in different environments to ensure that the requirements of the broadest possible range of users will be catered for.



The prototypes

The project has already produced a first prototype that was successfully demonstrated at an expert review-session involving the European Commission. This prototype demonstrated the implementation of the Transrouter graphical user interface under the MS Windows operating system and a restricted number of possible routings to be considered by the Transrouter decision kernel.

The Transrouter consortium is now developing the second prototype, which will integrate all component tools in the application and offer a larger number of possible translation routes to the decision kernel.

Some of the major challenges for the development of the second prototype include:

* The development and integration of the component tools
* The definition of *quality* in the context of Transrouter and the impact of quality considerations on the recommendation and selection of specific routes
* The modeling of a finite set of routes
* The presentation to the user of the multidimensional data emerging from the decision kernel in a meaningful way
* The support for the extendibility and openness of the Transrouter framework

The user requirements

The second prototype will also take into account the suggestions and feedback from the Transrouter user advisory group and other potential Transrouter users by focussing on the main criteria of time, cost and quality in relation to each translation route considered. The reports produced by Transrouter will offer a detailed breakdown of the implications on time, cost and quality for each route and for each language and vendor considered. By offering accurate information to non-technical translation managers at this level, the implications on the project's schedule and budget will become evident. This should increase the level of confidence in the correctness of the 'routing' decisions taken and, in turn, make the use of translation technology products more likely.

Outlook

When the Transrouter project will finish in April 2000, the project will have achieved some major innovative results. The project will have:

* produced a template for and prototype implementation of the electronic profiling of essential translation technology applications and of different types of translation projects;
* created a model for the simple but effective approach to routing based on user requirements (cost, time, quality);
* implemented a prototype of this model making use of electronically available information on projects, translation technology applications and linguistic resources in an open architecture.

If you are a translation manager, Transrouter will put up-to-date information on translation technology applications and their respective linguistic resources at your fingertips. It will enable you to automatically analyse the implications on time, cost and quality of each of a number of possible translation routes involving different vendors - in-house and third party - using different approaches ('routes') to the translation of your project. It will provide you with an added level of confidence in your decision to choose a particular route and, ultimately, lead to a wider use of translation technology applications.

I would like to acknowledge the suggestions and comments provided by Roísín Cleary (LRC) and Maghi King (ISSCO) who reviewed the draft of this article.

(1) Among the members of the Transrouter user advisory group are: Alpnet Corporation, Corel Corporation Ltd., CTS Teoranta, EGT, Filenet Company Ireland Ltd., Gateway 2000, Lingtech A/S, LioNBRIDGE Technologies, Logos GmbH, Novell Ireland Software Ltd., Oracle WPTG, Oversaetterhuset A/S, Praetorius, STAR Deutschland GmbH, Symantec Ireland Ltd., TRADOS, Translation Experts Ltd., Translation Service of the European Commission, VistaTEC.

Reinhard Schaler
Localisation Research Centre (LRC)
Department of Computer Science and Information Systems
University of Limerick, Plassey
Limerick, IRELAND
Tel. +353-61-20213176
Fax +353-61-330876
E-mail: Reinhard.Schaler@ul.ie

Machine Translation Summit VII, A Summary of the Language Resources Thematic Session and the Language Resources Demo Session

Hitoshi ISAHARA, Communications Research Laboratory, Ministry of Posts and Telecommunications

MT Summit VII (September 13-17, 1999) turned out to be a great success with more than 250 participants from around the world. There were two sessions related to the use of corpora and lexicons on the 15th. One was the Language Resources Thematic Session, proposed by ELRA, and the other was the Language Resources Demo Session.

In the Thematic Session, there were four presentations. (1) Parallel Text Collections at the Linguistic Data Consortium, by Xiaoyi Ma, (2) The ELAN Slovene-English Aligned Corpus, by Tomaz Erjavec, (3) Harmonised Large-Scale Syntactic/Semantic Lexicons: a European Multilingual Infrastructure, by Nicoletta Calzolari and Antonio Zampolli, and (4) Developing Knowledge Bases for MT with Linguistically Motivated Quality-Based Learning, by Evelyne Viegas.

In (1), past and current work on the creation of parallel text corpora by the LDC was presented, e.g. the Canadian Hansard Corpus, the United Nation Parallel Text and the European Corpus Initiative Multilingual Corpus. In (2), a parallel resource was also presented. A sentence-aligned, tokenised, Slovene-English corpus developed in the scope of the EU ELAN project was discussed. (3) provided an overview of the situation of Language Resources in Europe. It included the PAROLE (corpora and morphological and syntactic lexicons), SIMPLE (semantic lexicons) and SPARKLE projects, as well as the distribution activities of ELRA. In (4), Viegas discussed the creation of new lexicon entries using lexico-semantic rules and the creation of new concepts for unknown words, using a new linguistically-motivated model to trigger concepts in context.

In the Demo Session, four demonstrations were also presented. (1) A Japanese-English aligned corpus from the Japan Electronic Industry Development Association, given by Hirofumi Sakurai, Ichiko Sata and Hitoshi Isahara, (2) The ORCHID Corpus Toolkit, given by Virach Somlertlamvanich, (3) KIBS: Korean Language Information Base System, given by Key-Sun Choi, and again (4) The ELAN Slovene-English Aligned Corpus, given by Tomaz Erjavec. In contrast to the Thematic Session which had four presentations from Europe and the United States, three demos in the Demo Session were from Asian countries.

(1) was from Japan and showed their Japanese-English parallel corpus which contains SGML tagged text, sentence-level alignment data, phrase-level alignment tags and correspondences between proper nouns and compound nouns in Japanese and English, using governmental white papers as a source of texts. (2) was from Thailand and is a linguistic toolkit that helps in viewing both plain text corpora and POS tagged corpora in Thai. Both types of corpora are indexed to support text retrieval with any combination of surface words and part-of-speech. The keywords can be multiple with a defined distance between them. With this toolkit, the linguists can browse the patterns of word cooccurrence, the patterns of POS occurrence and the occurrence of word with its POS. This toolkit is scheduled to be distributed under the ORCHID project. (3) is from Korea and demonstrated their Korean Language Information Base System (KIBS). KIBS has a mission to expedite the progress of Korean information processing through the construction, management, integration, distribution, and the practical use of large-scale Korean language information base including various Korean corpora (raw, agged, treebank, etc.), multilingual corpora, electronic dictionary, terminology, speech database, off-line handwritten character database planned for a period of 10 years from 1995 through 2004. (4) the Slovene-English parallel corpus discussed in the Thematic Session was also demonstrated.

Hitoshi Isahara
Communications Research Laboratory
Ministry of Posts and Telecommunications
isahara@cri.go.jp

Automated Orthography Conversion and Lexical Standardization for Vernacular Languages

Marilyn P. Mason, Mason Integrated Technologies

Introduction

Over the past 24 years, the tools for typesetting and publishing of literature in the Haitian Creole language have changed radically from electric typewriters to IBM Selectric composing machines to electronic word processors to today's second and third generation Pentium computers. Yet, none of this technological change has prevented the processing of Creole texts; it has rather enhanced the possibilities. This paper presents a few areas of research and product development for Haitian Creole language engineering.

"Major" Languages vs. Vernacular Languages

There are those who would have us believe that computer technology is only for the "major" languages. The large corporations tend to support the "major" languages first because that is where there is financial return on investment. Although the major market players have established priorities for producing language engineering applications for the major languages, the vernacular languages have been in definite need of publishing tools and applications for many years and this need will continue to grow in the near and far future.

Case Study: Haitian Creole

Take the case of written Haitian Creole (HC). Over a period of the last 45 or so years, it has gone through 3 major orthographic system changes, as well as many hybrids [1-3]. The literature base is in a state of confusion, depending upon who wrote what and when [4-8].

Allen & Hogan [7] have provided detailed frequency counts on variation found for 27 HC lexical items within electronic textual data collected from 13 independent sources. Their initial study on variation in HC spelling is limited only to the context of nasalization. An example, provided below, is taken from that study.

The word for "government"

Frequency	*Written form*
10	gouvèman
8	gouvèmnan
7	gouvènmam
924	gouvènman
5	gouvènnman
20	gouvenman

In taking into account the above words from the perspective of a phonemically based orthography for HC, it is obvious that the pronounced forms of the lexical items would be quite varied. This is just one example among hundreds of examples [5, 7-8] that demonstrate the variation of the written lexicon of this vernacular language. It has been clearly shown [5, 7] that variation in HC spelling for the same lexical items has been found not only to be 'inter-textual' (i.e., between the many different editorial teams writing in Creole), which is something probably to be expected, but also that variation is very frequent at the 'intra-textual' level (i.e., within the same texts produced by the same editorial team). This tendency toward a high level of variation in real-life publishing and authoring contexts leads one to consider how to develop automated authoring tools for processing text written in vernacular languages, in light of the great need that has been shown in various articles on the topic [9-10].

Systematically dealing with Variation in Haitian Creole

Despite the high level of variation in HC, this language is quite systematic and the data is computationally usable when compiled into machine-readable form [5, 7, 11-14]. Once decisions are made with regard to the standardized forms, the algorithms can be appropriately configured. Past R&D on HC systems (e.g., OCR software, machine translation systems, spell checkers, text-to-speech synthesis) has been conducted by both the DIPLOMAT project of Carnegie Mellon University (www.lti.cs.cmu.edu/Research/Diplomat) and Mason Integrated Technologies, Ltd (http://hometown.aol.com/mit2usa/Index2.html). This research and product development has shown that automated processes can be developed to assist speakers of vernacular languages to create and/or process textual data in their native tongue.

Vernacular Language Lexical Standardization via Orthography Conversion

Research has been conducted by several institutes [5, 7, 14-16] on how to process vernacular language written text, although few projects have thus far produced a usable, functional, tool for end-user native speakers of these types of languages. In 1991, however, the prototype of a flexible, semi-automated process was completed for the conversion of texts written in earlier HC orthographies to conform to the Institut Pédagogique National (IPN) orthography that had been officially established in Haiti by the Orthography Law of 1979 [11]. The algorithm is based on a core of fixed phonemic-to-graphemic rules along with a set of other rules for the use of apostrophes, hyphens, contractions, punctuation, capitalization, proper names, and nasalization that were established within the framework of the IPN orthography.

Development of CreoleConvert™

In order to conduct research on orthography conversion, it was necessary to have a large corpus of electronic HC texts. Because such a "digital treasure" did not yet exist in the late 1980's, a total of 7 1/2 person years of time were spent retyping the entire HC Bible into a computer. Why was that book chosen? In 1989, what was to become one of the most widely-read books in the Haitian Creole language [17] was published, not in the official IPN orthography that had been established 10 years before, but rather in an earlier, outdated orthography. From the beginning, this book was a prime candidate for orthographic updating but, because the original text had been manually typeset, it still needed to be digitized in order to be manipulated and edited [18].

Why choose to type (hand enter) the text instead of scan it? Since the text was printed on both sides of "see-through" India paper, the scanner captured both sides of each page during scanning attempts. Even taking the additional step of photocopying the pages did not altogether eliminate the background interference. Also, the tiny superscripted Bible verse numbers created havoc ten years ago for OCR software, and this has not improved since then. Data entry of the HC Bible began in 1989 [18].

By using copies of the newly digitized HC texts, it was possible to begin experimenting with a basic, intuitive "character matching" approach of speed editing away from older orthographies toward IPN (i.e., texts in the Pressoir-Faublas orthography converted to IPN, texts in the McConnell orthography converted to IPN, etc.). This was conducted within the standard editing tools of "off-the-shelf" word processing software [11].

Over a number of years of development, along with benchmark and validation tests of the program, the process matured from a semi-automated process taking 2 hours to orthographically process a 250-page book, to a fully-automated process requiring less than 2 1/2 minutes to convert that same 250-page book from one orthography into another. This technique, known as the Mason Method of Haitian Creole Orthography Conversion (MMHCOC™), led to a software product called CreoleConvert™, that was then used to automatically and successfully convert the outdated orthographies of samples from well-known books such as Boukan, Jé Nou Louvri and Chanmòt la to the IPN orthography [19].

Since May 1996, this methodology/software has been demonstrated [14, 20-22] and test-marketed in Haiti, in Florida (USA), in France, and in the Seychelles by Mason Integrated Technologies Ltd (MIT2). This is a Boston-based (USA) start-up company that was formed to enable publishers, writers, educators, and governmental and non-governmental agencies in developing nations to quickly and efficiently standardize printed materials. This company fosters further research and development for the broad-based delivery of such tools in Haiti, the Haitian Diaspora, and other nations and languages for which this methodology has shown to be applicable. Negotiations are currently underway with the Seychelles government in order to proceed with implementation of this technology in that Indian Ocean nation. This is just a first step at a worldwide level toward the standardization of written texts of vernacular languages.

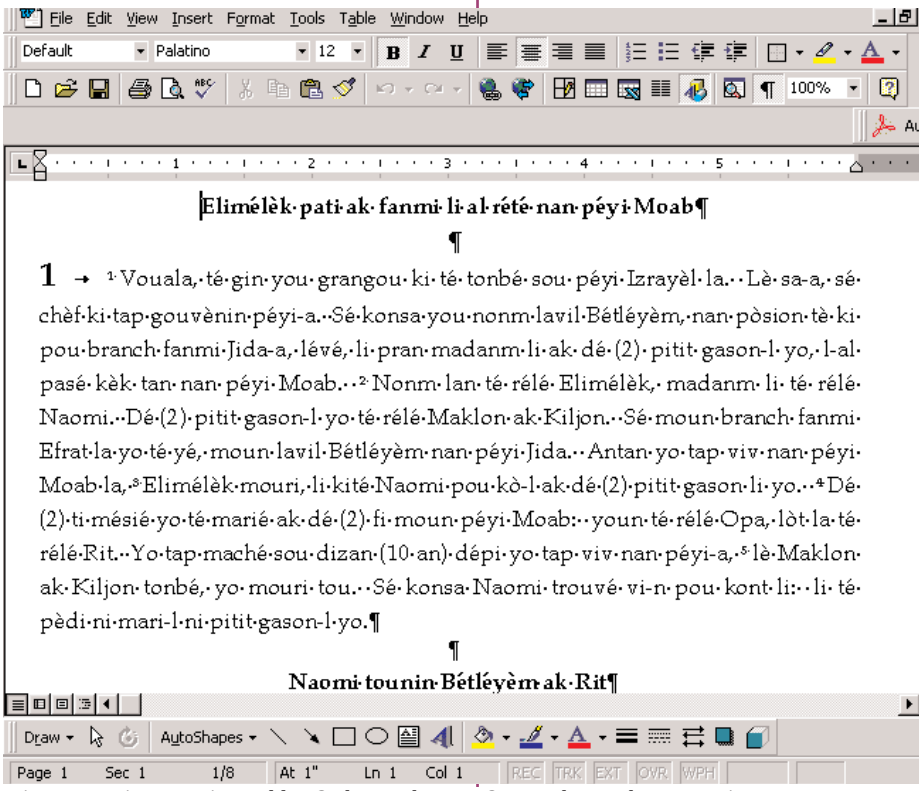


Fig. 1: Text in Pressoir-Faublas Orthography BEFORE orthography conversion

Conclusion

The development of usable authoring and translation systems and strategies is based on standardization of the lexicon of the languages to be processed. For some of the international languages, such standardization has been achieved over time and with the recent help of integrated spelling checkers in Microsoft Word and other applications. The majority of the world's languages, being minority and vernacular languages, have not been able to benefit from such advantages of the modern technological world. Due to recent efforts, it is now possible to provide services to a majority of languages of the world by allowing them to achieve lexical standardization by using existing and upcoming corpora. In applying these technologies to the standardization of corpora, the next step would be to see how additional multilingual documentation technologies could be developed for such languages. However, if techniques are not developed and implemented to provide for something as simple as lexical standardization and spell-checking, these minority languages of today and tomorrow will suffer greatly with respect to the authoring and translation technologies that have been developed to meet the globalization needs of the modern world.

REFERENCES

1. DEJEAN, Yves. 1977. *Comment Ecrire le Créole d'Haïti.* [Abridged and revised Ph.D. Thesis, Indiana University, 1977], Outremont, Québec: Collectif Paroles, 1980.

2. VALDMAN, Albert. 1978. *Le Créole: Structure, Statut et Origine.* Paris. Editions Klincksieck, pp. 349-350.

4. VALDMAN, Albert. 1988. *Diglossia and Language Conflict in Haiti.* In International Journal for the Sociology of Language, 71, p. 76.

5. ALLEN, Jeffrey. 1998. *Lexical variation in Haitian Creole and orthographic issues for Machine Translation (MT) and Optical Character Recognition (OCR) applications.* Paper presented at the Embedded MT Systems Workshop at AMTA98, Philadelphia, 28 October 1998.

6. SCHIEFFLIN, Bambi and Rachelle Charlier DOUCET. 1992. *The 'Real' Haitian Creole: Metalinguistics and Orthographic Choice.* In Pragmatics 2:3, pp. 427-443.

7. ALLEN, Jeffrey and Christopher HOGAN. 1998. Evaluating Haitian Creole orthographies from a non-literacy-based perspective. Paper presented at the annual meeting of the Society for Pidgin and Creole Linguistics, New York City, 9-10 January 1998.

8. ESKENAZI, Maxine, HOGAN, Christopher, ALLEN, Jeffrey, and Robert FREDERKING. 1998. *Issues in database design: recording and processing speech from new populations (poster session).* In LREC Proceedings, 28-30 May 1998, Granada, Spain. Vol. 2, pp. 1289-1293.

9. OSTLER, Nicholas. 1999. *Does Size Matter? Language Technology and the Smaller Language.* ELRA Newsletter, Vol 4 N1. Jan-Mar 1999.

10. BAKER, Paul, MCENERY, Tony, SEBBA, Mark, and Lou BURNARD. 1998. *Minority Language Engineering.* In ELRA Newsletter, Vol 3 N4, Nov 1998, p. 10.

11. MASON, Marilyn. 1991. Unpublished manuscript *"Novel Method for Orthography Conversion in Haitian Creole"* (June 18, 1991).

12. MASON, Marilyn. 1994. Unpublished manuscript *"Story behind Color Coded Mason Method of Haitian Creole Orthography Conversion (CCMMHCOC)"* (May 3, 1994).

13. HOGAN, Christopher. 1999. *"OCR for Minority Languages".* In Proceedings of the 1999 Symposium on Document Image Understanding Technology, Annapolis, Maryland, April 1999, pp. 235--244.

14. MASON, Marilyn. 1999. *Orthography Standardisation Tools: Preparing Creole Languages for the New Millenium.* Paper and demo presented at the Seychelles '99 Creole Symposium, Mahé, Seychelles, October 26-28, 1999.

15. NIRENBURG, Sergei. 1998. *Project Boas: "A Linguist in the Box" as a Multi-Purpose Language Resource.* In LREC Proceedings, 28-30 May 1998, Granada, Spain. Vol. 2, p. 740.

16. CAMARA, Émile, CÉLESTIN NSTADI, Véronique Rey, and Jean VÉRONIS. December 1995. *Traitement Informatique des Langues Africaines: Problèmes et Perspectives.* Action de Recherche Partagée, Section 3.2. ALAF (AUPELF-UREF) Document ALAF ALA1. Version 1.0. http://www.lpl.univ-aix.fr/projects/alaf/ALA1.html.

17. BIB LA an Ayisyin, Société Biblique Haïtienne, Port-au-Prince, 1989. Although the author typed the text initially for research purposes, written permission was granted to her in 1996 by Société Biblique Haïtienne to distribute versions of the texts which have been orthographically converted by CCMMHCOC™.

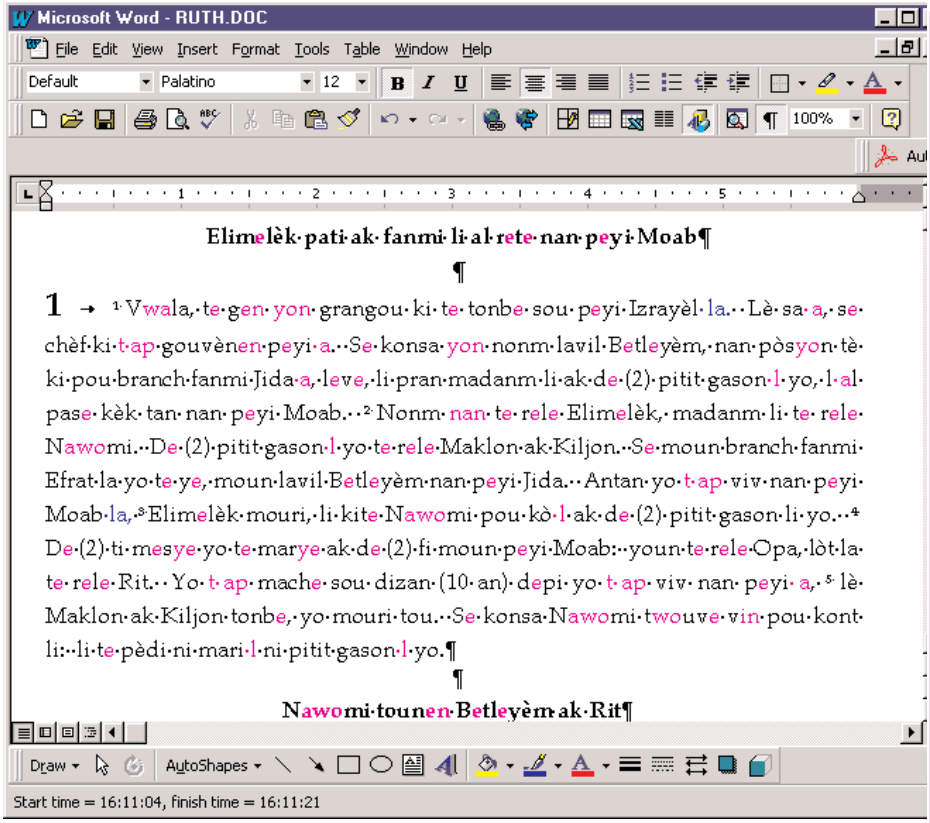


Fig. 2: Text in IPN Orthography AFTER orthography conversion

18. MASON, Marilyn. 2000, *forthcoming. Automated creole orthography conversion.* To appear in Journal of Pidgin and Creole Languages, 15:1, April 2000.

19. MASON, Marilyn. 1991. Unpublished manuscript *"Optical Character Recognition (OCR) Technology Widens Impact of Mason Method of Haitian Creole Orthography Conversion (MMHCOC)"* (June 28, 1991).

20. MASON, Marilyn. 1998. *Automated Approach to Haitian Creole Orthography Conversion.* Paper presented at the Fourth International Creole Language Workshop: "Standardizing the Orthography, Vocabulary and Structure", Florida International University, Miami, FL, March 19-21, 1998.

21. MASON, Marilyn. 1999. *Automated Approach to Haitian Creole Orthography Conversion: Can This Methodology Be Adapted to Other Creoles?* Paper presented at the 9è Colloque des Études Créoles. Held at the Université de Provence, Aix-en-Provence, France, 24 - 29 June 1999.

22. MASON, Marilyn. 1999. *Kreol + Computers + Internet = A Bright Future for Kreol!* Paper and demo presented at the 14th Annual Creole Festival, Mahé, Seychelles, October 23-31, 1999.

Marilyn Mason
Mason Integrated Technologies Ltd
P.O. Box 181015
Boston, MA 02118 USA
Tel.: +1 617 247-8885
Fax: +1 617 262-8923
E-mail: MariLinc@aol.com

The Evaluation Paradigm as a Resource Producer

Patrick Paroubek, LIMSI

Origins

MULTITAG (of the joint research program in Language Engineering of CNRS departments SHS and SPI) had the goal of producing and making available a 1 Million words corpus annotated with Part-Of-Speech (POS) tags out of the corpus tagged by the participants of the GRACE evaluation campaign (Adda et al., 99). The tags in the standard format proposed by EAGLES/MULTEXT/GRACE and the corpus documentation will represent a very useful material for linguistics studies, an essential resource for POS tagger training but also to study how machine learning by system combination.

To cut down the cost of proofreading the corpus, it has been semi-automatically corrected by verifying only the forms for which the annotations proposed by the different systems did not converge. This idea was inspired from what is done in machine learning with system combination. The level of convergence in the annotations provided a confidence measure to identify which forms needed to be checked.

The laboratory participating to MULTITAG are the INaLF, Institut National de la Langue Française (USR-705 Nancy) and the LIMSI, Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (UPR 3251). MULTITAG has collaborated with the CLIF project, in particular with the University of Provence (Jean Véronis) and TALANA (A. Abeillé) teams.

The linguistic resource

The lack of interactive tool readily available and customizable for correcting the corpus has been a hindrance to the project. For the linguistics aspects, defining the annotation procedure that the correctors had to follow has been much harder than was anticipated. In addition to the validation of the corpus annotations and of the method of system combination itself, this work yielded, from what the documents already been produced in GRACE about this problem, a refined annotation manual for the correctors, which will be very handy not only for further annotation work (for what concern decision making and consistency checking), but also for generic linguistic studies and the POS tagging problem itself. This annotation guide is part of the corpus documentation. Preliminary results of a study of the text typology of the different material composing the corpus can be found in [Illouz, 99]. The GRACE corpus tagged by the participants during the campaign is made of two parts: first the dry run corpus of roughly 450,000 forms (100,000 forms from the Le Monde newspaper and 350,000 forms from the FRANTEXT database of INaLF) and second the test corpus of approximately 830,000 forms (460,000 forms from the Le Monde newspaper and 370,000 forms from the FRANTEXT database). The dry run corpus has been normalized and the result of the system combination has been built but no manual validation of the resulting material has been performed. Because the test corpus was bigger and more interesting because it was annotated with the latest version of the GRACE morphosyntactic formalism we decided to work on this one. First it was normalized, then

some test where run to determine the best annotation combination procedure. The results show that it is possible to obtain an important increase in decision with a quasi-null loss in precision when combining the annotations of the five best systems (out of 15). These results were measured using the reference data of the GRACE campaign. In the first phase, manual validation (conducted by the University of Provence) was done on 38,643 forms of the test corpus (out of 830000 forms, which represents roughly 4%) for which the system combination procedure had produced an ambiguous annotation for the main morphosyntactic category or the subcategory (independently of other morphosyntactic information like the gender or the number). In a second phase of validation, all the forms whose annotations contained number, gender or person information (64,061 forms of the test corpus, roughly 8%) were manually checked. The test corpus is now undergoing final validation and will be added to the ELRA catalog in a near future as the first version of the MULTITAG resource.

References.

[Adda et al., 99] Gilles Adda, Joseph Mariani, Patrick Paroubek, Martin Rajman, Josette Lecomte, " L'action GRACE d'évaluation de l'assignation des parties du discours pour le français ", Langues vol.2 No 2, Juin 1999.

[Illouz 99] Gabriel Illouz, "Méta-Étiqueteur Adaptatif, Vers une utilisation pragmatique des ressources linguistiques", Conférence TALN, Cargèse 12-17 Juillet 1999, pp 185-194

URLs

GRACE documentation is available at : http://www.limsi.fr/TLP/grace

GRACE evaluation toolkit, which has been rewritten in the scope of the European project ELSE, is available (in beta version) at URL: http://www.limsi.fr/TLP/ELSE

Patrick Paroubek
LIMSI-CNRS
Human-Machine Communication Department
Bat 508 University Paris XI
BP 133
91 403 Orsay Cedex, France
Tel.: +33 1 69 85 80 54
Fax: +33 1 69 85 80 88
E-mail: paroubek@limsi.fr

Report on ELDA's Survey of Language Resource User Needs

Jeffrey Allen, ELDA

Introduction

This is a summary report describing the current state of an on-going survey that aims at determining the needs of users with respect to available and potentially available Language Resources (LRs). Within the framework of market monitoring activities outlined in the Language Resources - Production & Packaging (LRsP&P) LE4-8335 project, the main objective of this survey is to provide concrete figures for developing a more reliable and workable business plan for the European Language Resources Association (ELRA) and its Distribution Agency (ELDA), and to determine investment plans for sponsoring the production of new resources.

The results presented herein indicate only the information obtained from the summer 1999 questionnaire on User Needs which was conducted primarily among non-members of ELRA. Other questionnaires that were sent to members have also been analyzed, but they will not be treated here. This questionnaire was conducted by direct contact with personalized messages sent to 667 individual addresses. Provided below are some general overall statistics obtained from the questionnaire. The full range of questions in the study include: speech systems; speech evaluation and assessment; text processing; text processing systems; authoring and translation environments; information processing systems; multi-media and multimodal LRs; languages needed, LR domains/fields; regional areas of respondents; formatting of LRs; and the medium of delivery of LRs.

It is very important to note that the Summer 1999 questionnaire was not sent to regular ELRA members or clients, as was the case with earlier questionnaires. This does not mean that past LR clients members or clients were not contacted, but rather that the intention was not to use the current customer list as a basis for obtaining information about LR user needs. Addresses were extracted and compiled from a single database of contact addresses, and the general objective was to contact as many different players as possible, acknowledging of course that a single database is not exhaustive. In addition, those individuals contacted for this survey were known to possibly be more interested in Written LRs since ELDA has been focussing on improving the network of contacts in the Written and Terminology LR fields in 1999.

Survey statistics

Overall, of the nearly 670 questionnaires sent out individually to language engineering specialists, 17.5% returned as bad addresses. A clean-up procedure has been undertaken since that time in order to correct and/or remove the invalid addresses in further survey efforts. After discounting the invalid addresses, a total of 16.4% (90 respondents) of the total valid addresses returned a completed questionnaire to us. For this first round of sending out the new version of the questionnaire, this was a significant improvement over past questionnaires in this survey series. Additional follow-up strategies are underway to recontact those people who did not respond, and for contacting other individuals who did not receive the questionnaire in the summer 1999 batch of recipients.

Each LR type was divided into basic non-annotated data vs. annotated data. It can be noted that 30% of respondents are interested in basic speech data and 29% of respondents are interested in annotated speech data. This provides a round figure of 30% of participants that seek speech LRs. Those who conduct work on written LRs include 28% of respondents who seek basic data and 42% seek annotated data for syntactic bases. Those interested in lexical databases are

very numerous, amounting to 54% of respondents for basic data and 63% for annotated data. The fourth major category of types of LRs is that of text databases of which 63% of respondents seek basic data and 58% seek annotated data. In general, this survey demonstrates from a group of 90 respondents, approximately 1/3 are interested in speech LRs, and approximately 2/3 are interested in written LRs. This is a different audience than targeted in past survey efforts in 1997, 1998, and early 1999. These figures show that our survey work is reaching a high number of potential users of written LRs, as compared with past survey results. These figures demonstrate that ELDA's efforts to target the area of written LRs in 1999 has been successful.

Speech processing:

One section of the questionnaire aims at gathering information about the type of work being conducted in the Speech domain. In this section and all subsequent sections, the users are divided into those who conduct research and those who develop products. According to the results, the highest figures are between Speech Recognition and Speech Synthesis. 30% of respondents are involved in Speech recognition Research and 14% in Speech Recognition product development. On the other hand, 24% of respondents are involved in Speech synthesis Research and 9% in Speech synthesis product development. This is followed by those conducting work on the development of Speech databases (22% for Research and 19% for Product Development) and Speech Analysis (22% for Research and 6% for Product Development). The lower end of the spectrum include Speech Coding (9% for Research and 4% for Product Development), followed by Speech Workstation software (8% for Research and 4% for Product Development). From these results, we see that the major two types of users of Speech LRs are involved in Speech Recognition and Speech Synthesis (between 1/4 to 1/3 of respondents). Additional statistics on each speech processing subtopic are provided in the full report that is available to ELRA members.

Text processing

A section on general types of text processing systems was also included in the questionnaire. The results indicate that the development of text corpora is attested by 69% of respondents for Research and 22% Product Development. Syntactic Parsers (56% of respondents for Research and 23% for Product Development); Grammar Development (54% of respondents for Research and 23% for Product Development); Automatic Lexicon Recognition (41% of respondents for Research and 14% for Product Development); Text/Message Understanding (44% of respondents for Research and 11% for Product Development); Dialogue Management (22% of respondents for Research and 06% for Product Development); and Discourse Understanding (19% of respondents for Research and 1% for Product Development).

A very high amount of total participants (including those working on speech and text processing) are involved in Automatic Machine Translation activities (41% for Research purposes and 23% for Product development). This is followed by Terminology Management tools (32% for Research purposes and 23% for Product development) and Translation Memory applications (19% for Research purposes and 16% for Product development). These were followed by Grammar checkers (22% for Research purposes and 18% for Product development), Style checkers (20% for Research purposes and 11% for Product development), and Spell checkers (19% for Research purposes and 17% for Product development).

Multi-media and Multi-modal LRs

One of the most recent demands for LRs falls in the area of Multi-media and Multi-modal data. As for Multi-modal Processing, the recent survey shows that 50% of all respondents are interested in Multi-media data and 35% are interested in Multi-modal data. Approximately 10% of all respondents state that they want one of several types of Multi-modal LRs for Research. Product development is still low, but this is expected for a new area of research. There is an overwhelming increase from the information obtained in the 1997 Autumn/Fall Survey in which only 1/18th of the surveyed participants were interested in Multi-modal LRs.

Languages needed

Another one of the questionnaire sections asked for the languages desired with regard to LR data. These statistics clearly help us understand the needs, correlated with what is currently offered, and to see where there is a potential lack in what is offered. Taking into account that each respondent could tick more than one language box in the questionnaire, the following percentages refer to the total number of individual boxes ticked on language, not to the total number of respondents: European Languages - 67%; Eastern European Languages - 15%; Asian Languages - 12%; Mid-East Languages - 5%.

Many more details on each of the above-mentioned points are found in the full report. The results obtained from the 1999 summer User Needs Questionnaire aimed at complementing information already received from ELRA members and customers, and to determine if there are similar trends among the non-ELRA member institutions. We have taken the results of this questionnaire, along with previous questionnaires, and are redesigning the strategy for further work. This includes extending the survey to cover a larger base of recipients, and by targeting other domains that are specific to the Human Language Technology field. The current questionnaire results are therefore setting a benchmark for future survey work. Also, these results are helping ELDA rework its overall marketing strategy for promoting Language Resources.

Jeff Allen
ELRA/ELDA
55-57, rue Brillat Savarin
75013 Paris, France
Tel.: +33 1 43 13 33 33
Fax: +33 1 43 13 33 30
E-mail: jeff@elda.fr

New Resources

ELRA-S0072 Danish SpeechDat(II) FDB-1000 / ELRA-S0073 Danish SpeechDat(II) FDB-4000

The Danish SpeechDat(II) database, recorded over the Danish fixed telephone network, comprises two sets, FDB-1000 (1000 Danish speakers) and FDB-4000 (4000 Danish speakers). The SpeechDat databases have been collected and annotated by the Center for PersonKommunikation (CPK). The speech databases made within the SpeechDat(II) project were validated by SPEX, the Netherlands, to assess their compliance with the SpeechDat format and content specifications.

Speech samples are stored as sequences of 8-bit 8 kHz A-law. Each prompted utterance is stored in a separate file. Each signal file is accompanied by an ASCII SAM label file which contains the relevant descriptive information.
Each speaker uttered the following items: 3 application words; 1 sequence of 10 isolated digits; 4 numbers : 1 sheet number (5-10 digits), 1 telephone number (9-11 digits), 1 credit card number (16 digits), 1 PIN code (6 digits); 3 dates : 1 spontaneous (year of birth), 1 prompted date (word style), 1 relative and general date exp.; 1 word spotting phrase using an application word (embedded); 1 isolated digit; 3 spelled word : 1 spontaneous (own forename), 1 spelling of directory city name, 1 real word for coverage; 1 currency money amount; 1 natural number; 5 directory assistance : 1 spontaneous, own forename, 1 city of school at 7 years (spontaneous), 1 most frequent cities (set of 500), 1 most frequent company/agency (set of 500 names), 1 "forename surname" (set of 500 names); 2 yes/no questions : 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.

The following age distribution for the complete set of 4000 speakers has been obtained: 372 speakers are below 16 years old, 1004 speakers are between 16 and 30, 1109 speakers are between 31 and 45, 901 speakers are between 46 and 60, and 614 speakers are over 60.

A pronunciation lexicon with a phonemic transcription in SAMPA is also included.

ELRA-S0072 Danish SpeechDat(II) FDB-1000		ELRA-S0073 Danish SpeechDat(II) FDB-4000	
Price for ELRA members:	Price for non members:	Price for ELRA members:	Price for non members:
For research use: 9,000 EURO	For research use: 22,000 EURO	For research use: 28,000 EURO	For research use: 48,000 EURO
For commercial use: 18,000 EURO	For commercial use: 25,000 EURO	For commercial use: 40,000 EURO	For commercial use: 56,000 EURO

ELRA-S0074 British English SpeechDat(II) MDB-1000

The British English SpeechDat(II) MDB-1000 comprises 1000 British speakers recorded over the GSM digital mobile network. The database was produced by BT Labs in Suffolk, England. The MDB-1000 database is partitioned into 5 CDs in ISO 9660 format. The speech databases made within the SpeechDat(II) project were validated by SPEX, the Netherlands, to assess their compliance with the SpeechDat format and content specifications.

Speech samples are stored as sequences of 8-bit 8 kHz A-law. Each prompted utterance is stored in a separate file. Each signal file is accompanied by an ASCII SAM label file which contains the relevant descriptive information.

Each speaker uttered the following items: 1 sequence of 10 isolated digits; 3 connected 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); 2 isolated digits, 3 spelled words: 1 spontaneous name (own forename), 1 city name, 1 real / artificial word for coverage; 1 currency money amount; 1 natural number; 5 directory assistance names: 1 spontaneous name (own forename), 1 city of birth / growing up (spontaneous), 1 most frequent cities (set of 500), 1 most frequent company / agency (set of 500), 1 'forename surname' (set of 150 'full' names); 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.

The following age distribution has been obtained: 329 speakers between 16 and 300, 340 speakers between 31 and 45, 331 speakers between 46 and 60.

A pronunciation lexicon with a phonemic transcription in SAMPA is alsoincluded.

Price for ELRA members:	For research use: 20,000 EURO	For commercial use: 28,000 EURO
Price for non members:	For research use: 25,000 EURO	For commercial use: 35,000 EURO

ELRA-S0069 Swedish SpeechDat(II) FDB-5000 / ELRA-S0070 Swedish SpeechDat(II) FDB-1000

The Swedish SpeechDat(II) database, recorded over the Swedish fixed telephone network, comprises 2 sets, FDB-1000 (1000 Swedish speakers, 4 CDs, each of which comprises 250 speakers sessions) and FDB-5000 (5000 Swedish speakers; 25 CDs, each of which comprises 200 speakers sessions). The SpeechDat databases have been collected and annotated by the Department of Speech, Music and Hearing, KTH. The speech databases made within the SpeechDat(II) project were validated by SPEX, the Netherlands, to assess their compliance with the SpeechDat format and content specifications.

Speech samples are stored as sequences of 8-bit 8 kHz A-law. Each prompted utterance is stored in a separate file. Each signal file is accompanied by an ASCII SAM label file which contains the relevant descriptive information.

Each speaker uttered the following items: 1 isolated single digit; 1 sequence of 10 isolated digits; 4 numbers : 1 sheet number (5-10 digits), 1 telephone number (9-11 digits), 1 credit card number (16 digits), 1 PIN code (6 digits); 1 currency money amount; 1 natural number; 3 dates : 1 spontaneous (date or year of birth), 1 prompted date, 1 relative or general date expression; 2 time phrases : 1 time of day (spontaneous), 1 time phrase (word style); 3 spelled words : 1 spontaneous (own forename), 1 city name, 1 real word for coverage; 5 directory assistance utterances : 1 spontaneous, own forename, 1 city of school at 7 years (spontaneous), 1 frequent city name, 1 frequent company name, 1 common forename and surname; 2 yes/no questions : 1 predominantly "yes" question, 1 predominantly "no" question; 3 application words; 1 keyword phrase using an embedded application word; 4 phonetically rich words; 9 phonetically rich sentences

The database also contains additional Swedish specific material for speaker verification purposes and dialectal studies: 2 sentences for speaker verification purposes, same for all speakers, 4 connected digits strings (3-6 digits) for speaker verification purposes, 2 sentences for dialectal studies, same for all speakers

Age distribution

ELRA-S0069 Swedish SpeechDat(II) FDB-5000: 315 speakers below 16 years old, 2095 speakers between 16 and 30, 1080 speakers between 31 and 45, 1078 speakers between 46 and 60, and 432 speakers are over 60.
ELRA-S0070 Swedish SpeechDat(II) FDB-1000: 43 speakers below 16 years old, 429 speakers between 16 and 30, 208 speakers between 31 and 45, 241 speakers between 46 and 60, and 79 speakers over 60.

A pronunciation lexicon with a phonemic transcription in SAMPA is also included.

ELRA-S0069 Swedish SpeechDat(II) FDB-5000		ELRA-S0070 Swedish SpeechDat(II) FDB-1000	
Price for ELRA members:	Price for non members:	Price for ELRA members:	Price for non members:
For research use: 35,000 EURO	For research use: 60,000 EURO	For research use: 9,000 EURO	For research use: 22,000 EURO
For commercial use: 50,000 EURO	For commercial use: 70,000 EURO	For commercial use: 18,000 EURO	For commercial use: 25,000 EURO

ELRA-S0071 Swedish SpeechDat(II) MDB-1000

The Swedish SpeechDat(II) MDB-1000 comprises 1000 Swedish speakers recorded over the Swedish mobile telephone network. The SpeechDat database has been collected and annotated by the Department of Speech, Music and Hearing, KTH. The MDB-1000 database is partitioned into 5 CDs, each of which comprises 200 speakers sessions. The speech databases made within the SpeechDat(II) project were validated by SPEX, the Netherlands, to assess their compliance with the SpeechDat format and content specifications.

Speech samples are stored as sequences of 8-bit 8 kHz A-law. Each prompted utterance is stored in a separate file. Each signal file is accompanied by an ASCII SAM label file which contains the relevant descriptive information.

Each speaker uttered the following items: 2 isolated single digits; 1 sequence of 10 isolated digits; 4 numbers : 1 sheet number (5 digits), 1 telephone number (9-11 digits), 1 credit card number (16 digits), 1 PIN code (6 digits); 1 currency money amount; 1 natural number; 3 dates : 1 spontaneous (date or year of birth), 1 prompted date, 1 relative or general date expression; 2 time phrases: 1 time of recording, 1 time phrase; 3 spelled words : 1 spontaneous (own forename), 1 city name, 1 real word for coverage; 5 directory assistance utterances : 1 spontaneous, own forename, 1 city of school at 7 years (spontaneous), 1 frequent city name, 1 frequent company name, 1 common forename and surname; 2 yes/no questions : 1 predominantly "yes" question, 1 predominantly "no" question; 6 application words; 1 keyword phrase using an embedded application word; 4 phonetically rich words; 9 phonetically rich sentences.

The database also contains additional Swedish specific material for speaker verification purposes and dialectal studies: 2 sentences for speaker verification purposes, same for all speakers, 4 connected digits strings (3-6 digits) for speaker verification purposes, 2 sentences for dialectal studies, same for all speakers.

The following age distribution has been obtained: 32 speakers are below 16 years old, 348 speakers are between 16 and 30, 253 speakers are between 31 and 45, 292 speakers are between 46 and 60, and 75 speakers are over 60.

A pronunciation lexicon with a phonemic transcription in SAMPA is also included.

Price for ELRA members:	For research use: 22,000 EURO	For commercial use: 25,000 EURO
Price for non members:	For research use: 25,000 EURO	For commercial use: 35,000 EURO

ELRA-S0075 Welsh SpeechDat(II) FDB-2000

The Welsh SpeechDat(II) FDB-2000 comprises 2000 Welsh speakers (918 male speakers et 1082 speakers) recorded over the British fixed telephone network. The database was produced by BT Labs in Suffolk, England and collected by the Speech Research Group at the University of Wales Swansea, Wales. The FDB-2000 database is partitioned into 10 CDs, each of which comprises 200 speakers sessions. The speech databases made within the SpeechDat(II) project were validated by SPEX, the Netherlands, to assess their compliance with the SpeechDat format and content specifications.

Speech samples are stored as sequences of 8-bit 8 kHz A-law. Each prompted utterance is stored in a separate file. Each signal file is accompanied by an ASCII SAM label file which contains the relevant descriptive information.

Each speaker uttered the following items: 6 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 or year of birth), 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 word (letter sequences): 1 spontaneous (e.g. own forename), 1 city name, 1 real/artificial for coverage; 1 currency money amount; 1 natural number; 5 directory assistance names: 1 spontaneous (own forename), 1 city of birth / growing up (spontaneous), 1 most frequent city name (set of 500), 1 common forename and surname; 2 yes/no questions : 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.

The following age distribution has been obtained: 509 speakers between 16 and 30, 645 speakers between 31 and 45, 565 speakers between 46 and 60 and 281 speakers over 60.

A pronunciation lexicon with a phonemic transcription in SAMPA is also included.

Price for ELRA members:	For research use: 18,000 EURO	For commercial use: 25,000 EURO
Price for non members:	For research use: 25,000 EURO	For commercial use: 35,000 EURO

Up date on Language Resources from the ELRA Catalogue

ELRA-S0034 Verbmobil

This resource consists of spontaneous speech recorded in a dialog task (appointment scheduling). The BAS edition of the German part is fully labelled and segmented into phonemic/phonetic SAMPA by the MAUS system (see ELRA Newsletter Vol.2n4) and partly segmented manually.

New corpora available via ELRA (for the complete list, please contact ELRA or visit ELRA or BAS Web sites):

VM CD 15.1 - VM15.1 (new edition)

Verbmobil II - 19 spontaneous dialogues (19 close mic, 19 room mic, 19 phone line (GSM)), 3117 turns, transliteration (VM II format), NIST headers, partitur files*.

VM CD 20.1 - VM20.1 (new edition)

Verbmobil II - 30 spontaneous dialogues (10 close mic, 27 room mic, 10 phone line (GSM)), 1957 turns, transliteration (VM II format), NIST headers, partitur files*

VM CD 21.1 - VM21.1 (new edition)

Verbmobil II - 38 spontaneous dialogues (38 close mic, 2 room mic, 22 phone line (GSM)), 2331 turns, transliteration (VM II format), NIST headers, partitur files*

* partitur files : files describing the different parts which constitute the corpus - word order, phrase order, etc.

Price for ELRA members:	127,82 EURO/CD-Rom
Price for non members:	255,65 EURO/CD-Rom

ELRA-W0015 Le Monde

The text corpus ELRA-W0015 Le Monde for the year 1998 is now available.

Price for ELRA members:				Price for non members:			
1yr	238.91 EURO	4yrs	955.65 EURO	1yr	310.59 EURO	4yrs	1,242.35 EURO
2yrs	477.83 EURO	5yrs	1,194.56 EURO	2yrs	621.17 EURO	5yrs	1,552.93 EURO
3yrs	716.74 EURO			3yrs	931.76 EURO		