The ELRA Newsletter

January - March 2001

Vol.6 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.

Editor in Chief:
Khalid Choukri

Editor:
Khalid Choukri &
Valérie Mapelli

Layout:
Magali Duclaux
Emilie Marquois

Contributors:
D. Broeder
H. Brugman
Martine de Calmès
Jean-Pierre Chanod
Khalid Choukri
Guy Pérennou
P. Wittenburg

ISSN: 1026-8200

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

Dear Members,

For the first newsletter of the Millennium, we would like to start with the announcement of our Members' Annual General Assembly, a main event for ELRA during the first quarter of this year 2001, which will take place on Friday, 6th April at UNESCO premises in Paris. The necessary reports have been mailed in due time to all our members.

In this issue, we will also give you a brief overview of ELRA activities for the previous year, and the perspectives for the year to come, and we are glad to announce that our team now counts 3 more permanent members since the beginning of February. This will allow us undoubtedly to offer you better services.

Some major events took place during the year 2000, covering several areas. The LREC conference was organised in Athens, bringing together 500 to 600 attendees. We collected a large number speech data in partnership with some industrial partners, and prepared new proposals for the IST programme of the European Commission.

We also continued our work in many projects funded at a national or European level (GEMA, AVISE, Network - DC…) and have been actively involved in several evaluation initiatives such as AURORA, CLEF and AMARYLLIS.

To have a concrete overview, here are a few figures which illustrate the results of our activities in 2000: 173 items were sold in 2000 (compared to 110 in 1999). The number of language resources available in our catalogue has increased from 181 Spoken Language Resources (SLR) to 200, from 101 Written Language Resources (WLR) to 145, and our efforts in the terminology area have been reduced trusting the GEMA project to solve the structural problem we face in this area. 25 new members joined ELRA (12 for the spoken college, 12 for the written college, and 1 for the terminology college).

ELDA signed a collaboration agreement with and its US counterpart, the LDC (Linguistic Data Consortium), to launch a new project which aims at harmonising the activities and operations of national and international data outers. We started to re-design the ELRA and ELDA web sites, and to update the information available to make it more friendly and more efficient, i.e. by improving the presentation of our catalogue.

At a European level, there are several activities which are worth being noted here: the web site of the GEMA project, which aims at creating and developing a portal entirely dedicated to language resources and terminology, should now be open, @ www.lingoo.com, the kick-off meeting for C-oral-rom took place on 15th January. This project, which aims at building a large database of aligned corpora for 4 spoken romance languages, has now been officially launched.

A word about the European Commission action: a new programme, called eContent, has recently been launched in the area of Internet content products and services. For more information, please refer to page 5.

In this issue of the Newsletter, a report on ELRA 2000 activities is included, as well as a brief introduction to the new European eContent programme, a call for eContent experts' participation, and, finally, the announcement of a few job openings at the Commission.

Several articles from experts in the field of Language Resources are also part of this issue: the first one, written by Guy Pérennou and Martine de Calmès from the "Institut de Recherche en Informatique (IRIT)" deals with lexical resources designed for an automatic speech processing and the modelling of the pronunciation variability.

The second article, entitled "Multilingual Resources at XRCE" by Jean-Pierre Chanod, presents the multilingual components which have been developed over the years and which are used for terminology extraction, information retrieval, knowledge extraction or question answering at XRCE.

A report on the "Workshop on Annotation Architecture and Software Tools for Multi-Media Language Resources and Large Corpora", from P. Wittenburg, H. Brugman and D. Broeder, following a pre-conference workshop to LREC2000, is also available.

As usual, the final section is dedicated to the newly acquired resources, which are:

- ELRA-S0094 Czech SpeechDat(E)
- ELRA-S0095 Slovak SpeechDat(E)
- ELRA-S0099 Russian SpeechDat(E)
- ELRA-S0096 German SpeechDat(II) MDB-1000
- ELRA-S0098 British English SpeechDat(II) SDB-2400
- ELRA-S0097 British English SpeechDat(II) FDB-4000
- ELRA-S0100 MHATLex
- ELRA-W0026 Parole Irish Corpus
- ELRA-L0043 English Parole Lexicon

We would like to remind our members that they should have received an invoice to renew their membership to ELRA for 2001. Please do not forget either to proceed to your renewal, or to advise us in case you do not want to, by emailing Valérie Raymond, raymond@elda.fr.

Antonio Zampolli, President

Khalid Choukri, CEO

ELRA Annual Report 2000

Khalid Choukri

Year 2000 can be considered, in many respects, as a transition year. Despite the problems we faced to recruit new staff, our activities continued to expand in various areas. Our efforts to attract new members were successful (with an increase of 20% of paid up members), and probably illustrate the usefulness of the resources we include in our catalogue.

We also increased the number of resources in our catalogue: compared to the last year, from 181 Spoken Language Resources (SLR) to 200, from 101 Written Language Resources (WLR) to 145. Our efforts in the terminology area have been reduced trusting the GEMA project to solve the structural problem we face in this area.

The distribution effort has seen a substantial growth of our revenues (more that 1.25 M€ in 2000 compared to 780 K€ in 1999) and the number of items distributed (173 in 2000 compared to 110 in 1999). We continued to publish our newsletter on a quarterly basis, both in English and French, with some contributions from key people in our areas of involvement. The lack of personnel forced us to suspend the electronic bulletin we used to e-mail every month to our members. This action will be resumed in 2001 as we are recruiting new people. The LRs-P&P project, funded by the EC, which aimed at commissioning the production of resources is now over. This project helped us commission the production and/or the packaging of useful resources, now part of our catalogue, after a validation procedure. In addition to that, LRs-P&P funded a number of surveys which help us better understand the current situation of HLT market(s) and its evolution. Funding from the French government is also being used to prepare a written corpus of modern French. ELDA launched a data collection service and already collected four speech databases : UK English, French, US English, and German with very strong speaker distribution requirements e.g. demographic, dialectal, gender, etc. We expect this service to be very useful to our members.

ELRA/ELDA also put some efforts to address the validation issues. Important tasks are conducted to improve the quality of our resources particularly in the speech area via our validation unit (SPEX, The Netherlands).

Our participation to evaluation activities has been pushed one step ahead through data supply to projects such as Amaryllis and CLEF but also through the management of data delivery to the Aurora participants in real conditions of evaluation campaigns.

ELRA/ELDA has been very active in the MLIS project called GEMA, for which we conducted a user needs survey, various marketing actions and re-negotiation of our distribution licenses to obtain the rights for the use of terminological resources on the GEMA portal for consultation. We will also be contributing to the evaluation of the Portal in 2001.

The work of the other MLIS project (Network-DC) concerning a partnership between ELRA and our counterpart in the US LDC (Linguistic Data Collection) has not been conducted according to the initial plans because of the very late response of the US funding agency (NSF). A kick-off meeting took place in December 2000 and we had to ask for an extension of the project duration to ensure that we will fulfil our commitments.

ELRA/ELDA has also been part of several consortia that submitted proposals to the European Commission within the IST program. It is likely that three proposals will be accepted for funding (more information at the General assembly and in coming issues). A fourth one has already been accepted (C-ORAL-ROM) and will be launched in January 2001.

New projects have been signed with the French agencies in which we will focus on LR identification for some specific purposes.

The major event of this year was probably the very successful LREC conference organised in Athens. LREC is becoming the main event in the field and we hope that we will be able to continue this series in 2002.

ELRA/ELDA has been very active in coordinating the actions and initiatives of many national players and agencies, through the work of ENABLER (European National Activities for Basic Language Resources). This initiative has been packaged as an accompanying measure and submitted to the IST program for funding.

Our web site(s) will be re-designed in 2001 to account for the new services we offer and to incorporate some of the new features that internet can offer today. The catalogue will be reworked to consider the work being carried out on validation and to offer better search capabilities.

Last but not least, our financial situation is safer and may encourage us to invest in new Language Resources and related issues (Validation, Production, …).

Membership

Concerning the membership drive, we managed to attract several new members. We have now 108 members compared to 95 last year. What is also noticeable is the increase of paying members (about +25% more in 2000). We had 95 paid up members (out of 108 who registered), compared to 79 (out of 95 in 1999).

This shows a progress of +15% in the Speech college (from 44 to 52), +30 % in the Written (from 22 to 33) and +1 member in the terminology college.

Distribution of resources

During this fiscal year we substantially improved our sales from 110 resources in 1999 to 173 in 2000 and our revenues from 780 K€ in 199 to over 1.25 M€ (+60%). The ELRA/ELDA margin ratios were stable : 32% (1999) and 33,7% (2000).

Our sales to members still represent over 85% in 2000 compared to 92% in 1999; Speech resources represented 76%, written resources 21.5% and terminology databases about 2.5% (to compare to 86%, 13.9% and 0.1% in 1999 respectively). This clearly indicates that the new written resources in our catalogue are now appropriate for a large set of applications. We are still playing a balanced role within R&D and Commercial environments with 156 resources distributed for R&D and 110 for Commercial use (to compare to 104 and 74 respectively in 1999). Of course, the revenues are different: 6% in R&D and 94% from commercial users, to compare to 4.7% and 95.3% in 1999.

Identification of LRs

As usual, we have devoted a lot of efforts to enter into new agreements to secure distribution rights. Our speech resources increased from 101 to 145 items with some key resources from SpeechDat projects (SpeechDat-II, SpeechDat-E, SALA), from Babel and other resources developed by private industrial partners. Our written resources have grown from 181 to 200 items with a very important agreement with EDR (Japan) and several resources from the Parole Project.

Validation Work

In late 1999, we established a first unit of our validation Network for Spoken Language Resources which carried out its work as planned in 2000. This unit (SPEX, The Netherlands) prepared a number of documents related to validation issues and run validation procedures for us on several resources. The quality will be an important part of our catalogue that is being re-designed. We plan to add a validation flag on the catalogue page (together with a validation report when available). We plan to start a bug reporting mechanism to ensure the flow of feedback from the users. This action will be hopefully extended to the written area in 2001.

Evaluation

ELRA has started contributing to evaluation programs through the supply of Language Resources, appropriate for evaluation and testing. We are actively involved in initiatives such as:

- CLEF (Cross-Language Evaluation Forum, Information Retrieval System Evaluation for European Languages), which consisted in three main evaluation tracks: Multilingual Information Retrieval (searching a multilingual document collection for relevant documents. The multilingual document collection contains English, German, French, and Italian documents); Bilingual Information Retrieval (A cross-language task has been provided in which the query language can be French, German or Italian and the target document collection is English), and Monolingual (non-English) Information Retrieval;

- AMARYLLIS which consists in an evaluation project to assess systems and tools for the access to textual information in French. The extension (AMARYLLIS-II) aims at addressing the needs to access multilingual textual databases in French;

- AURORA which was originally set up to establish a world wide standard for the feature extraction software which forms the core of the front-end of a DSR (Distributed Speech Recognition) system. ELRA/ELDA has been asked to be in charge of the distribution of the databases developed for this purpose. The Aurora group decided to carry out a blind-evaluation: the participant got training data in 2000 and will get an unseen database by the 1st of February 2001 and have a limited period of time to carry out the evaluation and to deliver the results. ELRA/ELDA has to ensure that the data is supplied to each participant in due time.

Commissioning the production of Language Resources

ELRA/ELDA had selected 8 proposals submitted within its 1999 Call for proposals to produce or package Language Resources. The Language Resources produced within these projects were delivered to ELRA by May 2000. As initially planned, most of them had to go through a validation procedure and are now part of our catalogue.

Promotion and awareness

We published 4 issues of the ELRA newsletter in 2000, in French and English. We attended a number of events to promote our activities. The LREC'2000 is considered as a very successful event. Our web site continues to be very attractive (more than 250,000 visitors in 2000, compared to 138,000 in 1999) and we do our best to update it very frequently.

Relationships with the European Commission

The first European project that helped ELRA establish its infrastructure (LE1-1019) is now officially over (final notification received from the commission). The LRsP&P ended in May 2000 and we are very glad to report that we fulfilled all our commitments.

The Language Resources accepted for funding under the project were selected by an expert group and the board of ELRA, following a call for applications issued in February 1999. All projects delivered the resources as planned, except the German-French Parallel Corpus of 30 Million words, for which the delivered data seems not to correspond in its nature to the one planned (sources of raw data).

The new MLIS project called GEMA is progressing as planned. The GEMA portal will be in operation by early 2001 with (only) 2 months deviation from the very initial planning.

Just to remind you, the GEMA project (Gates for an Enhanced Multilingual resource Access) aims at providing a central and organized access point for the linguistic sector. A friendly web site has already been set up (after brainstorming on the web name and main services). ELDA is actively searching and negotiating new resources for the portal that will be set up within GEMA. ELDA is also involved in the marketing and promotion of the outcomes of the project.

The Network-DC project has been delayed, waiting for notification from the US funding agency (received by our US partner LDC in October 2000). The work is now on the track.

Its objective is to start a transatlantic collaboration between the European Language Resources Distribution Agency (ELDA) and the US Linguistic Data Consortium (LDC) that includes networking and cross-agreements, for the production, acquisition, normalization, certification and distribution of spoken and written language resources for research and technology development.

A new project, C-ORAL-ROM, started in January 2001. It is about spoken data from conversational/colloquial speech from France, Italy, Spain and Portugal. ELRA/ELDA is involved in the distribution of the outcome of the project, in addition to addressing legal and information dissemination issues.

Future work

A major workpackage of 2001 would be the revision of our business plan and the preparation (and the implementation) of a strategy plan for the period 2001-2005. In particular specific and targeted marketing actions following the users analysis and market monitoring (as a follow up of LRs-P&P) should be deeply considered to update our business and investment plans.

In 2001, we will continue to carry out the regular activities related to the identification of new resources, the distribution, and the sales. We will continue to promote our activities through the quarterly newsletter (issued in French and English) and other information dissemination means.

We are re-designing our web sites to consider new services and new technical features to make it more friendly and more efficient. In particular, we would like to improve the presentation of our catalogue, taking into account the new technical possibilities and also the discussions we had on validation aspects and other services being offered by ELRA/ELDA.

Validation will be also a major keyword in our daily work: we need to implement the recommendation of the Spoken Language Resources validation committee and extend it to Written Language Resources later on.

We will also stress our role in the evaluation field, through a very active involvement in evaluation projects/campaigns. It is important to envisage that ELRA starts formally and officially a new branch of activities related to Evaluation. This should also apply to Multimodal/multimedia resources to fulfil the requirements we have learnt from our recent surveys.

It is also important that we carry out the tasks we are responsible for in several European and French projects and we will make sure that we capitalise on them.

This includes the "cooperation project" with DGLF on the "Corpus du Français Contemporain" (Modern French corpus), that will be carried out in cooperation with third parties (producers of LRs). Another part of the DGLF grant will be used to survey existing multimedia/multimodal resources, designed for HLT purposes. The other projects that benefit from the support of French Ministry of research and Ministry of industry have as a main task to identify resources suitable for search engines, MT systems, speech synthesis, etc. It is of paramount importance that we strongly capitalize on such projects as the work carried out within these projects should be fed into our regular activities enriching our catalogue with more attractive/useful resources.

The next LREC is scheduled for 2002. By the second quarter of 2001, we will produce a guideline booklet to help organize such a huge event with proven and professional procedures. We will launch the preparation (selection of location, dates, call for papers, etc.) by mid 2001.

We will have the opportunity and the pleasure to share some information with our members at the General assembly that will take place on April 6th 2001 in Paris.

eContent Programme

Overview

The European Commission is about to set in motion a new market-orientated programme in the area of Internet content products and services, which has become known as eContent. The programme is intended to stimulate the development and use of "European digital content on the global networks", and consists of three action lines:

- AL1: Improving access to and expanding use of public sector information.
- AL2: Enhancing content production in a multilingual and multicultural environment.
- AL3: Increasing the dynamic of the digital content market.

The overall goal of the projects and other actions established within Action Line 2 is to investigate and experiment with new partnerships, strategies and solutions for designing and producing e-content services which can be speedily and effectively tailored to the requirements of European and global markets.

Cost-shared projects established within this Action Line will focus in the 2001-2002 time frame on products and services designed to be used in connection with Internet access points, ranging from PCs through mobiles and communicating appliances, to television sets and game consoles.

Transnational projects and other collaborative actions are expected to address three broad communities:

- private and public e-content players planning to enhance their offerings (e.g. web portals, mobile services, broadband information and entertainment services) through cost effective internationalisation strategies and localisation processes;
- businesses and public-sector actors (e.g. utilities) which intend to establish or strengthen their presence on the e-commerce scene through e.g. web marketing, retailing and customer care offerings adapted to the linguistic and cultural requirements of a broad range of user groups;
- private/public partnerships geared towards a wider deployment and commercial exploitation of public sector information.

Further information on the eContent programme can be found at www.cordis.lu/econtent

For additional details on Action Line 2 and Language technologies and applications in general: www.hltcentral.org

Early information on upcoming calls for proposals will be published on the above web sites towards the end of February.

Inquiries re eContent: econtent@cec.eu.int

Inquiries re Action Line 2 and associated R&D developments: hlt@cec.eu.int

Roberto Cencioni
Head of unit INFSO/D4

Search for eContent experts

In the framework of a new market-orientated programme which has become known as "eContent", we would like to invite you to register as a candidate evaluator/reviewer in response to a recent call for experts. Further details can be found on www.cordis.lu/econtent/evaluators.htm

A number of call evaluations and project reviews will have to be performed by independent experts in the coming months. The evaluation of the proposals submitted in response to the first call, which is due for publication around mid March, is scheduled for the first half of July. A second call for proposals will be launched in early November.

Please note that all applicants, including those already on experts' lists drawn up for other European programmes (including the 5th framework), must submit a NEW application.

An online registration facility is available at: www.cordis.lu/econtent/expert_form.htm

For any further information and assistance please contact us at infso-experts.econtent@cec.eu.int

Job openings at the European Commission

Unit D4 - Linguistic Applications of the Information Society (INFSO) will have in the coming months a few openings for candidates willing to join a dynamic team managing sizeable research and non-research programmes in the field of content, language and speech technologies and applications.

Successful candidates will be offered a 12-month contract as an A-grade auxiliary agent. They will work on the INFSO premises in Luxembourg, and will be entrusted with one or several of the following tasks under the supervision of senior INFSO staff.

Interested candidates should submit a 2-page curriculum with a recent photo and relevant references to the following e-mail address: hlt@cec.eu.int no later than 31 March 2001.

Multilingual Resources at XRCE

Jean-Pierre Chanod, Xerox Research Centre Europe, France

Over the years, XRCE has been engaged in a systematic effort to build a suite of consistent and reusable multilingual components ranging from morphology and part-of-speech (POS) tagging for most languages, to parsing and semantic disambiguation for a more limited number of languages. Descriptions of more than 15 languages are now available, at various levels of complexity. Those components are based on core, language-independent techniques, such as finite-state calculus, parsing engines or statistics. They are integrated into the same unified architecture for all languages, the Xerox Linguistic Development Architecture (XeLDA). They are used into a variety of commercial and research applications, such as terminology extraction, information retrieval, knowledge extraction, or question answering.

Finite-state calculus

Finite-state technology is one of the fundamental technologies for developing language resources, esp. tokenisers, morphological analysers, noun phrase extractors and other language-specific components [Kar et al. 97].

- The basic calculus is built on a central library that implements the fundamental operations on finite-state networks. It is the result of a long-standing research effort [Kap & Kay 94, Kar 95, Moh 97, Roc & Sch 97]. An interactive tutorial on finite-state calculus is also available at http://www.xrce.xerox.com/research/mltt/fst/home.html. Besides the basic operations (concatenation, union, intersection, composition, replace operator) the library provides various algorithms to improve further the compaction, speed and ease of use of the networks. The calculus also includes specific functions to describe two-level rules and to build lexical transducers.

Morphology

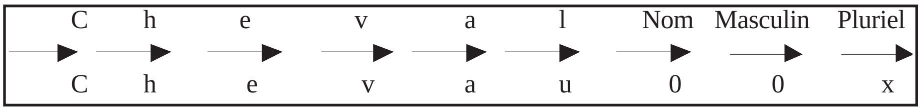
Morphological variations can be conveniently represented by finite-state transducers, which encode on the one side surface forms and on the other side normalised representations of such surface forms [Kar 94]. More specifically:

1. the allowed combinations of morphemes can be encoded as a finite-state network;

2. the rules that determine the context-dependent form of each morpheme can be implemented as finite-state transducers (cf. two-level morphology [Kos 83]);

3. the lexicon network and the rule transducers can be composed into a single automaton, a lexical transducer, that contains all the morphological information about the language including derivation, inflection, and compounding.

For example, the following diagram shows how the plural masculine form of the French noun *cheval* produces the surface form *chevaux* (where the 0 symbol represents the empty symbol):



Lexical transducers have many advantages. They are bi-directional (the same network for both analysis and generation), fast (thousands of words per second), and compact. They also provide an adequate formalism for a multilingual approach to language processing, as major European languages and non-Indo-European languages (e.g. Finnish, Hungarian, Arabic, Basque) can be described in this framework.

Part-of-speech tagging

The general purpose of a part-of-speech tagger is to associate each word in a text with its morphosyntactic category (represented by a tag), as in the following example:

This+PRON is+VAUX_3SG a+DET sentence+NOUN_SG .+SENT

The process of tagging consists in three steps:

1. tokenisation: break a text into tokens

2. lexical lookup: provide all potential tags for each token

3. disambiguation: assign to each token a single tag

Each step is performed by an application program that uses language specific data:

- The tokenisation step uses a finite-state transducer to insert token boundaries around simple words (or multiword expressions), punctuation, numbers, etc.

- Lexical lookup requires a morphological analyser to associate each token with one or more readings. Unknown words are handled by a guesser that provides potential part-of-speech categories based on affix patterns.

- In XRCE language suite, disambiguation is based on probabilistic methods (Hidden Markov Model), [Cut & al. 92], which offer various advantages such as ease of training and speed. However, some experiments [Cha & Tap 95] showed that a limited number of disambiguation rules could reach the same level of accuracy. This may become the source of interesting developments in POS tagging, as one deals with highly inflective, agglutinative and/or free-word order languages for which simple contextual analysis and restricted tagsets are not adequate [Haj & Hla 98].

Noun Phrase extraction

Finite-state Noun Phrase extraction [Bou 93, Lau & Dra 94, Str 95, Sch 96] consists in extracting patterns associated with candidates NPs. Such patterns can be defined by regular expressions based on sequences of tags such as:

ADJ* NOUN+ (PREP NOUN).

The example above specifies that an NP can be represented by a sequence of one or more nouns [NOUN+] preceded by any number of adjectives [ADJ*] and optionally followed by a preposition and a noun [(PREP NOUN)], the optionality being indicated in the regular expression by the parentheses.

Such a pattern would cover phrases like "digital libraries" "relational morphological analyser" "information retrieval system" or "network of networks". Due to overgeneration, the same pattern would also cover undesirable sequences such as "art museum on Tuesday" in "John visited the art museum on Tuesday".

This highlights that simple noun phrase extraction based on pattern matching requires further processing, be it automatic (e.g. by using fine-grain syntactic or semantic subcategorisation in addition to part-of-speech information or by using corpus-based filtering methods) or manual (e.g. validation by terminologists or indexers).

Incremental finite-state parsing

Incremental Finite-state Parsing (IFSP) is an extension of finite-state technology to the level of phrases and sentences, in the more general framework of robust, i.e. parsing of unrestricted texts such as newspaper or web pages [Jen & al. 93, Abn 91]. IFSP computes syntactic structures, without fully analysing linguistic phenomena that require deep semantic or pragmatic knowledge. For instance, PP-attachment, coordinated or elliptic structures are not always fully analysed. The annotation scheme remains underspecified with respect to yet unresolved issues, especially if finer-grained linguistic information is necessary. This underspecification prevents parse failures, even on complex sentences. It also prevents some early linguistic interpretation based on too general parameters.

Syntactic information is added at the sentence level in an incremental way [AM & Cha 97a, AM & Cha 97b], depending on the contextual information available at a given stage. The implementation relies on a sequence of networks built with the replace operator.

The parsing process is incremental in the sense that the linguistic description attached to a given transducer in the sequence relies on the preceding sequence of transducers and can be revised at a later stage.

The parser output can be used for further processing such as extraction of dependency relations over unrestricted corpora. In tests on French corpora (technical manuals, newspaper), precision is around 90-97% for subjects (84-88% for objects) and recall around 86-92% for subjects (80-90% for objects). The system being highly modular the strategy for dependency extraction may be adjusted to different domains of application, while the first phase of syntactic annotation is general enough to remain the same across domains.

Here is a sample sentence extracted from this current section:

Annotation:

[SC [NP _The parsing process NP]/SUBJ :v is SC] [AP incremental AP] [PP in the sense PP] [SC that [NP the linguistic description NP]/SUBJ attached [PP to a given transducer PP] [PP in the sequence PP] :v relies SC] [PP on the preceding sequence PP] [PP of transducers PP] and [SC :v can be revised SC] [PP at a later stage PP].

Dependency extraction:

- SUBJ(description,rely)
- SUBJ(process,be)
- SUBJPASS(description,revise)
- SUBJPASS(process,revise)
- VMODOBJ(revise,at,stage)
- VMODOBJ (rely,at,stage)
- VMODOBJ(rely,on,sequence)
- VMODOBJ(be,in,sense)
- ADJ(late,stage)
- ADJ(given,transducer)
- ADJ(linguistic,description)
- NPPASDOBJ(description,attach)
- ATTR(process,incremental)
- NNPREP(sequence,at,stage)
- NNPREP(sequence,of,transducer)
- NNPREP(transducer,in,sequence)
- NNPREP(description,to,transducer)

Sense disambiguation

The word sense disambiguation (WSD) system developed at XRCE is based on two existing WSD systems. The first system [Seg & al. 98], the Semantic Dictionary Lookup, is built on top of Locolex (cf. infra). It uses information about collocates and subcategorization frames derived from the Oxford-Hachette French Dictionary [Oxf 94]. The disambiguation process relies on dependency relations computed by the incremental finite-state parser.

The second system [Din & al. 99], GINGER II, is an unsupervised transformation-based semantic tagger first built for English. Semantic disambiguation rules are automatically extracted from dictionary examples and their sense numberings. Because senses and examples have been defined by lexicographers, they provide a reliable linguistic source for constructing a database of semantic disambiguation rules. In that respect, dictionaries appear as valuable semantically tagged corpus.

An overview of language resources

The following diagram gives an overview of currently available language resources.

language	**en**	**fr**	**es**	**de**	**it**	**pt**	**nl**	**sw**	**no**	**da**	**fi**	**hu**	**cz**	**pl**	**ru**	**zh**	**ar**	**tu**	**ro**
tokenizer	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X			
morphology	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
POS tagger	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X			
NP mark-up	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X			
Parser (IFSP)	X	X	X																
Sense dis.	X	X																	

The development of morphology or part-of-speech taggers largely benefited from pre-existing resources, such as lexicons and annotated corpora. This multilingual development effort was realised through a number of external collaborations, esp. in Central and Eastern Europe. European projects like Elsnet-goes-East were great facilitators in that respect. Our most recent development is in Chinese, again in collaboration with a Chinese university. Even in the case of Chinese, we could reuse pre-existing resources to create new resources compatible with our own underlying technologies and formats (e.g. finite-state lexical transducers). We then expanded the Chinese suite (tagger, NP-extractor) as we did for other languages.

The most advanced resources (parsing and sense disambiguation) cover a more limited number of languages. This is mostly due to the fact that they rely on more recent research. Additionally, they are based on detailed language-specific descriptions for which pre-existing resources are not easily accessible, if at all.

References

S. Abney, [Abn 91] Parsing by chunks, in *Principled-Based Parsing*, R. Berwick, S. Abney, and C. Tenny, (eds.), Kluwer Academic Publishers, Dordrecht, 1991.

Salah Aït-Mokhtar, Jean-Pierre Chanod [AM & Cha 97a] Incremental finite-state parsing, in *Proceedings of Applied Natural Language Processing* 1997, Washington, DC. April 97.

Salah Aït-Mokhtar, Jean-Pierre Chanod [AM & Cha 97b] Subject and Object Dependency Extraction Using Finite-State Transducers, *ACL workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. 1997, Madrid.

D. Bourigault [Bou 93] An endogenous corpus-based method for structural noun phrase disambiguation, *6th Conf. of EACL*, Utrecht, 1993.

Jean-Pierre Chanod, Pasi Tapanainen [Cha & Tap 95] Tagging French-comparing a statistical and a constraint-based method in *Seventh Conference of the European Chapter of the ACL*. Dublin, 1995.

Doug Cutting, Julian Kupiec, Jan Pedersen and Penelope Sibun [Cut & al. 92] A Practical Part-of-Speech Tagger. In *Proceedings of ANLP-92*, pages 133--140. Trento, 1992.

Dini, V. Di Tomaso, F. Segond [Din & al. 99] GINGER II: an example-driven word sense disambiguator. In *Computer and the Humanities*, to appear.

Jan Hajic and Barbora Hladka [Haj & Hla 98] Czech Language Processing / POS Tagging, *First International Conference on Language Resources and Evaluation*, Antonio Rubio, Natividad Gallardo, Rosa Castro and Antonio Tejada (eds.) Granada,1998.

Karen Jensen, George E. Heidorn, and Stephen D. Richardson, eds., [Jen & al. 93] *Natural language processing: the PLNLP approach*, Kluwer Academic Publishers, Boston, 1993.

Ronald M. Kaplan, Martin Kay. [Kap & Kay 94] Regular Models of Phonological Rule Systems. *Computational Linguistics*, 20:3 331-378, 1994.

Lauri Karttunen [Kar 94] Constructing Lexical Transducers. In *Proceedings of the 15th International Conference on Computational Linguistics*, Coling, Kyoto, Japan, 1994.

Lauri Karttunen [Kar 95] The Replace Operator. In *Proceedings of the 33rd*

Annual Meeting of the Association for Computational Linguistics, ACL-95} 16-23, Boston, 1995.

L. Karttunen, JP Chanod, G. Grefenstette, A Schiller [Kar et al. 97] Regular Expressions for language Engineering, *Journal of Natural Language Engineering*, vol 2 no 4 (1997) pp 307-330, 1997 Cambridge University Press.

Kimmo Koskenniemi. [Kos 83] *A General Computational Model for Word-Form Recognition and Production.* Department of General Linguistics. University of Helsinki. 1983

M. Lauer & M. Dras [Lau & Dra 94] A probabilistic model of compound nouns. *7th Joint Australian Conference on Artificial Intelligence*. 1994.

Mehryar Mohri [Moh 97] Finite-State Transducers in Language and Speech Processing. *Computational Linguistics* 23:2, 269-312, 1997.

Oxford-Hachette [Oxf 94] *The Oxford Hachette French Dictonary*, Edited by M-H Corréard and V. Grundy, Oxford University Press-Hachette.

E. Roche and Y. Schabes [Roc & Sch 97] E. Roche and Y. Schabes, eds. *Finite-State Language Processing*. MIT Press, Cambridge, Massachusetts. 1997.

Anne Schiller [Sch 96] Multilingual Finite-State Noun Phrase Extraction *ECAI '96 Workshop on Extended Finite-state Models of Language*. *Aug.* 11-12, 1996 Budapest.

F. Segond, E. Aimelet, L. Griot. [Seg & al. 98] “All you can use!" or how to perform Word Sense Disambiguation with available resources *Second Workshop on Lexical Semantic System*, Pisa, Italy, 1998.

T. Strzalkowski [Str 95] "Natural Language Information Retrieval." In Information Processing and Management, Vol. 31, No. 3. Pergamon/Elsevier. 1995.

Jean-Pierre Chanod
Xerox Research Centre Europe
6, chemin de Maupertuis,
38240 Meylan, France
Email: Jean-Pierre.Chanod@xrce.xerox.com
Web site: http://www.xrce.xerox.com

Lexical resources for spoken and written French at IRIT

Guy Pérennou and Martine de Calmès, Université Paul Sabatier, France

Introduction

Lexical resources for spoken and written languages play an important role in a variety of *Human Language Technology* applications like speech recognition and comprehension, dialogue systems, text or speech corpora annotation…

Such resources capture a part of the relationship between text and speech which underlies applications such as automatic dictation and certain approaches of the spelling correction or the text-to-speech synthesis.

Which information is needed in such resources? If we cannot give a categorical answer to this question we can however observe that a number of language engineering applications need fast surface processing where spelling, pronunciation and morpho-syntactic features are involved; word frequency may also be useful. Moreover, automatic speech or text processing may simultaneously involve several aspects of this lexical information.

It should be noted that two kinds of resources must be distinguished: on the one hand, the lexicons designed and optimised for a particular application; on the other hand, reference resources involving generality, portability and a large linguistic coverage.

With this perspective, BDLex resources for spoken and written French have been developed during the two last decades -the first versions within the GDR-PRC CHM (Man-Machine Communication of the French National Research Co-ordination Program). Recently, we have introduced new lexical resources, called MHATLex, specially adapted to the automatic speech processing and the modelling of the pronunciation variability.

BDLex resources

The present version contains about 450.000 inflected words derived from 50.000 canonical words...

A BDLex lexical entry is illustrated in table 1. The present materials also include a version BDLex-syll where the syllabic divisions are given in the field PHONO -where for example *samedi (Saturday)*, including two feet and three syllables, is represented by |sa,m@|di instead of sam@di.

Table 1: BDLex entry structure and a few examples

Spelling	Pronunciation		Morpho syntax			
ORTHO	PHONO	FPH	CS	VS	M	LIEN
prendre	pRa~dR	@	V		inf	=
prennent	pREn	@t”	V	3P		prendre
petites	p@tit	@z”	J	FP		petit
un	9~	n”	d	MS	di	=
avion	avjo~		N	MS		=

Modelling pronunciation via BDLex

A pronunciation model must allow the prediction and the generation of the possible pronunciation(s) of a word in each given sentence -or, on the contrary, the recognition of the possible words underlying a phonetic transcription.

BDLex allows such a modelling thanks to its phonological representations which make easier the use of phonological rules. The rules annexed to the material may help the user for designing his proper phonological engine in view of given applications. This engine may be very simple, as for example in the case of a text-to-speech pilot.

Phonological representations[1]. The phonetic code used is SAMPA (http://www.phon.ucl.ac.uk/home/sampa/home.htm) completed with a few particular conventions concerning

- the schwa /@)/ (or elidable "e"), for example in the words *prendre*, *prennent*, *petites* of table 1 -it may be elided or pronounced as the central and neutral unit, then transcribed by the symbol 6 at the phonetic (pronunciation) level;

- the liaison consonant -a liaison consonant C is represented by C" in FPH field, for example in the words *prennent*, *petites*, *un*.

Pronunciation free-variants (variants that can be predicted without using the context sentence) are easily generated thanks to the MPGs (Multiple Pronunciation Groups). These can be optional or frequently elided units, for example (l) in *fusil* (gun) /fyzi(l)/, (k) in *extraction* /E(k)stRaksjo~/, schwa in *samedi* /sam@di/.

More complex GPMs exist, like those occurring in borrowed word, for example : *starter* /staRt{6R}/ where {6R} may be pronounced [ER] or [9R] (that is as "èr" or "œr"), *adagio* /ada{dZj}o/ with a MPG {dZj} that may be pronounced [dZj], [Zj] or [dZ] ([Z] pronounced as "j" in French *je* or as "s" in English *pleasure*).

[1] Following the usual conventions, the transcriptions are represented between slashes at phonological level, for example /sam@di/, and between square brackets at phonetic (pronunciation) level, for example [samdi].

Table 2 - Examples of sentences (column 2) transcribed by mean of BDLex PHONO and FPH field (column 3) and after the use of phonological rules (column 4).

n°	Examples + (CS)	Transcription PHONO FPH	Pronunciation
(1)	*le(d) héros(N) (the hero)*	/l @ *****E/Ro/	[l6 E/Ro]
(2)	*ils(P) prennent(V) un(d) avion(N)*	/il z" pREn @**t"** 9~ **n**" avjo~/	[il pREn **(6)t** 9~**n** avjo~]
(3)	*(they take a plane)* *si(c) l'on(P) a(V) envie(N)* *(if we feel like it)*	/si lo~ n" a a~vi/	[il pREn 9~**n** avjo~] [si lo~n a a~vi]
(4)	*où(A) va(V) -t-il(P)?* *(where does he go?)*	/u va til/	[u va til]
(5)	*sers(V) -en(P) un(P) (serve one)*	/sER za~ 9~ n"/	[sER za~ 9~]
(6)	*sers(V) un(d) verre(N)* *(serve a glass)*	/sER 9~n" vER @/	[sER 9~ vER(6)]

In French an important category of pronunciation variants depends on the sentence context. The field FPH takes that into account; among other things, it can contain a schwa @ and/or a liaison consonant -this is illustrated in table 2.

An adequate use of this field supposes a morpho-syntactic control (more often a local control). For example, if a liaison is possible it is required after an article as *un*, optional between a verb and its complement-see the example (2) of table 2 where the n-liaison is required and the t-liaison optional. Moreover BDLex has specific representations and entries adapted to the phonological problems raised by euphonic consonants and enclitic pronouns - see examples (3) to (6) in table 2.

Variability of the pronunciation and speech recognition

During the last decade, automatic speech recognition has made important progresses in applications close to human language communication as automatic text dictation and conversational speech access to services. Now recognisers must be speaker-independent and accept fluent speech.

So the variability of the pronunciation has become a salient question, which has motivated various experiments; they have shown that modelling pronunciation variants can improve recognition accuracy.

MHAT (Markovian Harmonic Adaptation and Transduction). We introduced this model to take into account the pronunciation of markovian multi-level sources. Thus we can develop lexical resources easily usable in speech recognisers.

Three levels of representation related to the lexicon are considered:

- the syntactic surface level *S*, where a representation consists of a string of syntactic boundaries and references to inflected words (for the sake of legibility we refer to a word through its spelling);
- the phonological word level *W*, where a representation consists of phonological-units strings and phonological boundaries;
- and the phonetic level *P*, where a representation consists of phonetic-units strings.

At each level, a representation has two stages: the input level where the unit representations are put together, and the output level where contextual adaptations, required for a well formed representation, have been performed.

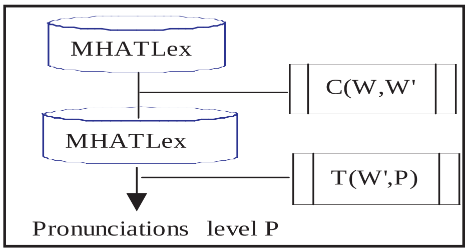
Phonological level W. At the input stage *W*, word phonological representations are inserted. The lexicon MATLexW includes the words provided with these phonological representations.

At the output stage *W'*, the representations have adapted to their context. A word *m* may have several variants W'(CtxG,m,CtxD) where CtxG and CtxD are the contextual conditions making possible the insertion of these variants, called *phonotypical words*, in a sentence. This is similar to the insertion of an inflected word under the control of syntactic features such as gender, number…

Thus, a phonotypical word would be a phonological inflected word. We can observe that generally such phonological inflexion is not taken into account by the spelling (in opposition to the morpho-syntactic inflexions). However exceptions exist: for example the words *l', cet, nouvel…* requiring CtxD=-C, that is a subsequent word starting with a non-consonant. They have a spelling variant: *le, ce, nouveau…* that occurs in other contexts. Words adapted to the context (that is phonotypical words), constitute the lexicon MHALexW'. A component C(W,W') makes the adaptation from the stage W to the stage W'.

MHATlex resources. They are identical to BDLex resources for the vocabulary, the spelling and the morpho-syntactic features. It is from the pronunciation point of view that the two resources are different: MHATLex allows pronunciation modelling including more explicitly free and contextual variability. They are better adapted to speech recognition. The basic resources (Fig.1) include the lexicon MHATLexW and the components C(W,W') and T(W',P) -defined below-designed for a pronunciation model. By simple table modifications the user can obtain a new model, in particular his own lexicon MHATLexW'.

Figure 1 Resources MHATLex



BDLex phonological and phonetic conventions are used in MHATLex; in addition, the latter uses specific units, the CPGs (Phonological Contextual Groups).

The CPGs model the effect of the context on the pronunciation. They may occur at the beginning or at the end of the words. Certain monosyllabic words consist in a single CPG. (for example the article or pronoun le).

To take into account the contextual effects, MHATLex distinguishes a double phonological point of view. On the one hand, a word generates contextual effects, regressive influence (InflD) on the antecedent word and/or progressive influence (InflG) on the subsequent word. On the other hand, a word pronunciation is possible in an antecedent context (CtxG) and/or a subsequent context (CtxD). This is illustrated in tables 3, 4 and 5.

Adaptation component C(W,W'). It transcribes CPGs in each given context by means of non-recursive rules such as those given in table 3.

Table 3: Excerpt of tables used by C(W, W’)

CtxG	W:GPC	CtxD	W’
	<n”#>	-C	N
	<n”#>	C	
	<~dR@>	-C	DR
	<~dR@>	C	(~dR@)
SO	<#p@>		p@
SF	<#p@>		p6

Tables 4 and 5 - Excerpts from MHATLexW and MHATLexW'. The phonotypical variants that cannot be inserted, given the sentence context of prendre un avion, are shaded.

ORTHO	InflD	PHONO-W	InflG
prendre	Q	pRa~<~dR@>	SF
un	-C	9~<n"#>	SO
avion	-C	avjo~	So

ORTHO	InlD	CtxG	PHONO-W'	CtxD	InflG
prendre	Q		pRa~<~dR@>	C	SF
			pRa~dR	-C	
un	-C		9~	C	SO
			9~n	-C	
avion	-C		avjo~		SO

Phonotypical word pronunciation. Once the phonotypical word selected, the phonological context of the sentence does not play any role. The transducer T(W',P) generates the pronunciations at stage *P*. For example, it will generate the pronunciations [dR6], [n]… for the (~dR@) of a variant of *prendre*.

The phonetic units are under the coarticulation effects. This is the motivation for the diphones or triphones used in speech recognisers. Such units are assigned to the stage *P'*. They do not involve lexical resources.

Perspectives

Two types of lexical resources of spoken and written French have been presented with a particular attention for the pronunciation variability. The first one, BDLex, supposes the recourse to phonological rules, the last one, MHATLex, can be integrated into the HMM framework of speech recognition.

About the content of such resources, two points at least still need research efforts.

On the one hand, the question of linguistic coverage, from the pronunciation point of view, remains. Indeed, seeing that conversational servers will be open to the public as for example in ARISE project, it becomes necessary to enlarge the coverage to dialectal and sociolectal speaking, including pronunciations with foreign accent.

On the other hand, the language models generally used in speech recognition should be improved to take into account better free and contextual pronunciation variants. At present, these models are essentially based on the inflected words. In French, as shown in this paper, the phonotypical word, inflected according to the phonological context, must be considered. Experiments to clarify these questions still remain to be done.

Bibliography

Pérennou, G. (1995). Phonological Component of an Automatic Speech Recognition, The Case of Liaison Processing. In *Levels in Speech Communication, Relations and Communications*, pp. 211-24. Elsevier.

Pérennou, G. (1996). Les règles et les niveaux en phonologie: du générativisme aux modèles markoviens. In *Fondements et perspectives en traitement automatique de la parole* (pp. 185-204). AUPELF-UREF, HACHETTE ou ELLIPSES.

De Calmès, M. Pérennou, G. (1998). BDLEX : a Lexicon for Spoken and Written French. In *1st International Conference on Language Resources and Evaluation*, pp. 1129-36, Granada.

Pérennou, G. de Calmès, M. (2000) MHATLex: Lexical Resources for Modelling the French Pronunciation. In 2d International Conference on Language Resources and Evaluation. pp. 257-64. Athens.

A complete panorama on pronunciation modelling can be found in :

Strik, H. Kessens, J.M. Webster M. (Eds.), (1998) Proceedings of ESCA Tutorial and Research Workshop on Modelling Pronunciation Variation for Automatic Speech Recognition, Rolduc, Kerkrade, The Nederlands.

Sites related to the material presented in the paper:

http://www.irit.fr/ACTIVITES/EQ_IHMPT /ress_ling/

http://www.icp.grenet.fr/ELRA/fr/cata/tab speech.html

Guy Pérennou
Institut de Recherche en Informatique (IRIT)
118, route de Narbonne
31062 Toulouse cedex, France
Email: perennou@irit.fr
Web site: http://www.irit.fr

Martine de Calmès
Institut de Recherche en Informatique (IRIT)
118, route de Narbonne
31062 Toulouse cedex, France
Email: decalmes@irit.fr
Web site: http://www.irit.fr

Report on the Workshop on Annotation Architecture and Software Tools for Multi-Media Language Resources and Large Corpora

Pre-conference workshop to LREC2000

P. Wittenburg, H. Brugman and D. Broeder

Also these two workshops ran under the flag of the new EAGLES/ISLE project, i.e. they were organized to define the actual needs of the community to be tackled in the project. At the end of this note we will draw some conclusions.

Contributions

It is not possible to be comprehensive and mention all contributions of the two workshop parts. We will limit ourselves to contributions and comments which are related to our tool-oriented work at the MPI. Some contributions focussed on the encoding of multi-modal behavior. It is fully clear that we don't have good insights about what people are doing in this area and that the EAGLES/ISLE project has to work on this. In this summary we will not comment on these contributions although they are very important for many of us.

ATLAS

The Atlas concept/architecture was introduced. It is based on an API which offers all functionality to deal with relational database structures implementing LDC's formal model (acyclic directed graphs). On top of this API various applications and APIs are planned which make use of this API. The architecture mentions AIF (ATLAS Interchange Format) files on the same level as the relational database, but operations are not symmetrical: AIF is only an import/export format which can be either generated or consumed. LDC's annotation graph model is well-known, it was generalised to be able to cope with higher-dimensional cases. The term *region* was introduced to denote a stretch in some n-dimensional space, so a time interval is a stretch in a 1-dimensional space, but a gesture occurs in a spatial as well as in a time stretch. Based on this an ATLAS Object Model was developed which served as a basis for developing the API. Currently, LDC people try to get the API stable, design the AIF, and start adapting/creating tools which work with the API. In another talk from LDC it was reported that a query language is being developed which seems to make use for the described API. Also the well-known Transcriber tool was told to support the API. If these tools were ready they were the first making LDC's ideas available in an operational form. ATLAS is one part within the TalkBank initiative which aims at understanding the needs of a large variety of disciplines and creating a universal format and a set of tools operating on it.

Comment

LDC has done a great job with analysing the various formats and describing a formal model. It helped all of us to clarify concepts and can serve as a reference. The results are similar but more comprehensive compared to a study which was made at the MPI as a basis for the EUDICO abstract corpus model years ago. And, of course, the community is highly interested in the results of the TalkBank project. Until now LDC has the universal formalism and ideas (some code, however, not yet stable) of how to implement this with the help of a relational database structure covered by an API. More has to be available to make better judgements about the impact of this work. Until now it is not clear for us what exactly will be accepted by the community. The AIF seems to be much more important than the API, since it would allow other developers to independently build tools and in doing so support the AIF. The AIF also is the documentation format. However, the AIF is not yet specified. Using the API might only be interesting for a few developers, if the underlying machinery (database engine) provides more efficient access as other methods. Relational databases, however, are fairly common. Much excellent analysis work has been done by the LDC people until now, but the hard programming results have to come. A format unification could be achieved when the TalkBank project would be able to describe a generic AIF in not too far time.

MATE

The MATE spoken corpus annotation program is demonstratable, although it has still some bugs[1]. SDU presented a tool which has as one of its core concepts the so-called coding modules. A coding module is a realization of an encoding scheme and it can be easily (in normal cases) specified by the user. MATE is delivered with a set of ready-made coding modules. These coding modules are used in two ways: (1) They are used to constrain the annotation and (2) they are used to generate DTDs which describe the structure of the XML files which MATE can handle. MATE also uses XML as an interchange format, i.e. internally MATE operates with a relational database. MATE is delivered with a powerful search tool which allows the user to do IR by using structural information and some statistics. MATE comes with a number of well-designed user interface components.

Comment

The MATE people have demonstrated a tool with a nice and to a large extent convincing user interface. Surprising for us was the decision that MATE cannot be used as a transcription tool. This is supported by the fact that the speech viewer is comparatively simple and attached. You need a first transcript and then can carry out further annotations. MATE is the first annotation program (as far as we know) which implemented an XML-import/export module. However, MATE does not apply the stand-off format, this decision is coherent with its goal to function as annotation tool based on a ready transcription. MATE might therefore have problems with multiple independent streams (channels) as they occur in multi-media annotations. Nevertheless, MATE is (almost) ready and may be used by many as a tool for manual annotations. A problem might be the limited number of input filters currently available (Xlabel, BAS). Some design decisions might make it difficult to extend MATE to a full-fledged multi-media annotation and exploitation tool, operating in distributed environments as is required for the work in our institute. Nevertheless, we can learn a lot from the MATE project.

Ghorbel

The most complex annotation situation seems to be given in TV studios where complex workflow processes influence the way annotations emerge from multiple interacting annotators. Complex relations between the different annotations are given such that the EPFL colleagues decided to use a knowledge base on top of the annotation system.

Comment

To us it is not clear whether this application introduces new types of structural phenomena in the annotation scheme which were not yet been described by others. If this complexity is covered by what has been described already, then the knowledge base can be seen as complementary, but some of the tools currently under developent and presented at the workshop should be able to cope with the annotation task. MPI investigation indicate that EUDICO's internal abstract corpus model is rich enough to handle such situations. But we are not yet sure about this.

CELLAR

With CELLAR a spin-off of the challenging but not finished Lingua-Links project was presented. The user can specify his/her data model and both a DTD and an SQL schema are created. The DTD could be used by an editor which is used to create XML-structured data. Such XML files can be imported to the CELLAR system which is based on a relational database engine. Applications can operate with the database. For Cellar it is claimed that the model can cope with data objects having many simultaneous properties and highly interrelated data requiring to encode associative links between related pieces of data.

Comment

In principle similar to MATE, Cellar offers a possibility to specify annotation schemas. It does so by creating a DTD both for defining the structure of an XML document and of a relational database. The idea is excellent. It seems that the designers had typical text-based annotations in mind and did not think of multi-media environments. It is not clear to us whether CELLAR can be used for complex structured annotations as we know them for for example gesture databases. It would make sense, if CELLAR would be available as a specification tool which is independent from concrete relational DBMS (since people are using different systems) and if it would be easily integratable into annotation tools. For us it is also unclear whether CELLAR can cope with dynamic environments, i.e. environments where people frequently change the annotation structure.

Romary/Lopez

LORIA people presented a layered framework to create annotation structures and to transform them into efficient internal representations. In the focus of their work is the term "free of redundance" which is similar to the term "normalized structure" in the field of database design. The first step is to create a "Relational Resource Organization Model" which describes the set of resource entities and the set of relations between entities. Resource entities are thought to be independent, i.e. basical-

[1] This is not to blame the developers, since we know that bug-free programming is a very hard job.

ly every annotation tier has to be represented in a separate file. A tier such as an orthographic transcription or an original text is the basis, i.e. all annotations refer to words or group of words of this basic tier. Based on this model an XML structure is derived where each independent resource element is stored in one XML document. This comes close to what is known as a stand-off model. Also the relations between the resource elements are stored in a separate XML document. Since this set of XML documents does not lend itself for efficient processing, a Finite State Representation mechanism is derived which is free of redundance. This is used to implement an efficient access machinery.

We speak about channels or in MPI's terminology about independent streams.

Comments

The approach to first build a good model of the data and from that derive an orthogonal XML structure seems to be helpful. However, it presupposes that the person exactly knows which kind of linguistic units will occur. In a dynamic environment which is often the case it is not known beforehand what users will encode, i.e. it is not possible to generate a model which goes down to the linguistic units. It is claimed that the redundancy-free FSR mechanism can be used for efficient access. However, this can only be true for certain type of access patterns. Increasing redundancy in general makes access faster. FSR are theoretical concepts which have to be mapped to physical database structures. Since the paper does not tell how this is done, it cannot be seen which type of access might be efficient which not. So, although the conceptual procedure is convincing it is not clear to us whether this framework is generally applicable.

Ide

Nancy gave two papers: one mentioning requirements for the work we all are doing and one explaining the possible gain in applying XML. The first was very useful as a general reference and will not be commented further. The second reported about extended functionality in XML to create links between annotations such as XLink, XPath, and XPointer. These mechanisms may have to be applied when complex annotation structures have to be represented within the XML formalism. XML transformation possibilities such as XSL and XSLT are more on the tool side where we don't know yet where these can be applied and whether they are appropriate in multi-media environments. XML schemas will be of large importance to better describe (and constrain) the contents of XML documents. However, XML Schemas are not yet accepted as an international standard and they are still subject of changes.

EUDICO

EUDICO is MPI's baby and will not be commented by us. It is ready as a player version to demonstrate its basic concepts. Still it has some functional gaps before it can be described as a full-fledged annotation and exploitation tool for multi-media language resources. Since it is still under development, it is not yet debugged. Nevertheless, it is one of the few operational true multi-media tools.

Discussion

The discussion after the talks and at the end of the session resulted in a number of interesting points:

- One major question focussed on the value of XML. It was generally agreed that XML will be very important as an open exchange format. The structure of a document will be well-described such that everyone can read XML-documents and use the data in some form. Therefore, it is also good for long-term documentation. However, much data will remain as it is and will not be converted into XML files. Also some of the non XML formats (TIPSTER, …) are much more suitable to the specific work people are doing, so there is no reason to step over to another format. However, tool developers should provide XML import/export modules. The main argument for using XML often is the availability of tools. However, in case of multi-media environments there is nothing. Further, there is the clear statement from LORIA people that XML is not a good modeling framework.

- There is still a debate whether XML structures can directly be used for processing. All major tool builders currently tend to provide XML import/export modules, but they internally often use relational databases or in case of LORIA a FS representation. One question which adressed the speed of retrieval was not answered although it is an important one.

- Extensibility of annotations is an important issue. Often people don't know beforehand how they will encode linguistic phenomena, i.e. there must be ways for individuals to enter just what they want and define arbitrary references and add arbitrary comments.

- The stand-off model seems to be widely accepted for XML documents. It implies that independent annotation layers are stored in different files and that links are set between these files by using structure pointers.

- Often the term "object" is used when people speak about structure elements in XML documents. This could lead to irritations, since one of the problems some toolbuilders have is exactly how to map rich object models to linear document structures. This mapping is not trivial.

- It is a general agreement that the tools or formats should not impose biases towards a certain linguistic unit. This implies that the annotation structure has to allow the user to define new tiers where he/she can choose new stretches (spatial or temporal) and label them. This was already well-described in the paper from SB&ML.

- There is a debate in how far tool developers have to provide "stereotypic" views on the data or whether formalisms such as XSL can be given to the user to have him/her create their own view on the data. In a multi-media environment only stereotypic viewers will work, i.e. viewers which were defined by the system developer. Most people see XSL as a way for specialists to easily create other type of layouts for textual documents. So XSL could form a medium layer for the specialist to create new views in the case of textual data.

- There was a short discussion about the usage of SMIL. As far as could be seen from the documentation so far SMIL is a tool for making synchronised representations via the web, but it can't be seen as a multi-media analysis and exploitation tool which would serve our needs.

Summary Statement

- Together with Nancy Ide we organized two workshops about annotation structures, encoding schemes, and the architecture of tools. While part of the talks were dedicated to rich textual structures other were focussing on the special requirements when working in a multi-media environment. The requirements are partly different.

- A great problem is seen in the fact that although we speak about very similar and largely overlapping things, still the terminology is very different. This refers to the statement of HT about the non-existing ontology of our field. The area in which we are active is very dynamic.

- This dynamic situation is the reason that makes us sure that we need the competition of different approaches. This is true for the representation formats as well as for the analysis and exploitation tools. LDC did do a great job with describing the various phenomena in complex annotation structures and deriving a common logical framework for annotations. But there is no doubt that we will have to try various formats in the area of multi-media corpora and that we need a variety of tools to create them and to exploit their content. New APIs such as that one from LDC are emerging, but we don't know yet whether they will be sufficient and whether it will do what we need. The availability of open exchange formats will help us a lot on the way to re-use language resources, but there is still a long way until suitable XML-structures for multi-modal content will have stabilized.

- As already mentioned at the beginning some projects started with annotating multi-modal behavior. But there are still many open questions in for example encoding gestures. What we need therefore is an overview about what people are doing in this area, how they are encoding multi-modal behavior, and what kind of analysis they intend to carry out. This may end up in suggestions for new projects to achieve greater coherence and thereby improve re-usability. On the other hand we need flexibility in this area, since we just started encoding multi-modal behavior.

- Only briefly during the workshop we spoke about how to integrate media and how to do streaming. This area is suffering from high dynamics on various levels. On the signal encoding level we have the trend form MJPEG (->Cinepak) to MPEG1, MPEG2, and MPEG4 which will keep those busy who have to build multi-media tools. On the higher level we have container APIs such as Quicktime and player APIs such as Java-Media-Framework, and much incompatibilities with respect to file formats. Driven by the media community we also have media annotation initiatives such as MPEG7 and Dublin-Core which will influence what we are doing to a certain extent.

- We have seen a number of architectures of software tools (MATE, GATE, ATLAS, EUDICO, CALIN, CELLAR, …). It seems that a multi-level structure is widely accepted: (1) At the physical level systems mostly operate with a relational database as internal format for efficiency reasons. Most tend to support an XML-based format for import/export. Few also support native formats such as CHAT. (2) Although terminology differs between the teams the essential point is that after methods of abstraction a universal layer was introduced. ATLAS for example speaks about an API which is based on a generalised object model. EUDICO speaks about an abstract corpus model. The difference in these two cases is that ATLAS makes the logical level available as an API, while in EUDICO the abstract level is part of the kernel. (3) Consequently, the next level, the application level, is different as well. In ATLAS applications are separate programs on top of the APIs, i.e. due to a lack of API descriptions it is not yet clear what the shared machinery is. In EUDICO there is a kernel based on the abstract model and applications are realized as class hierarchies on top of this machinery. GATE is designed for a somewhat different purpose. Its main objective is to allow language engineers easily add NLP modules to an existing framework which provides common functions such as data access and visualization. It makes use of the TIPSTER format which is widely accepted in the LE community and has proven its usefullness as a component framework at many sites. MATE's architecture is not yet fully clear to us. It seems that the search module was built separately from the annotation environment although all functionality is available via a unifying user interface. It is not clear to us whether there is a common API designed for such components or whether the logical description of the database is the common interface. CELLAR's major intention is the data modeling interface which generates structure descriptions for relational database as well as for XML documents. With respect to the architecture of the CALIN we cannot make statements yet, since the talk was not about such aspects.

- A short discussion was about the question to what extent we have to re-invent the wheel. It is good to have a limited number of data models which is the gasoline in our field. Therefore the analysing work about common formats is very important. Still due to the dynamics we will be far away from a situation where we have narrowed down the number of formats. In the area of multi-media annotations we see a number of activities such as TalkBank, MPEG7, Dublin-Core, EAGLES/ISLE etc. all dealing with partly similar type of questions, but raised from the perspectives of different communities. Additionally, we see the many different projects which still use their own formats from various reasons which are sometimes mission critical. Of course, we need to come to unification, but it will take a while. With respect to the machinery which makes use of the gasoline we believe that we need competition of different concepts. The interests are differing and we are far away from being able to design a framework which will handle all of them.

- Some "users" argued that it would be very helpful for the field to have unbiased descriptions of what the tools can and especially what they can't do. It was also required that it would be very useful to have demo examples (possibly in the web) to make it easy for the user to understand the main concepts.

EAGLES/ISLE Project

From the workshop we can extract a few major tasks for the EAGLES/ISLE project:

- We should start making an overview about the encoding schemes used in annotations of multi-modal behavior.

- The project should make an analysis of the architectural basics of the major tools and describe the available functions. It would also be useful to select a number of corpora such that the tool builders can show how the tools can deal with such corpora. The goals must be that the users can easily understand what the tool can do for them and that the professionals get a deeper insight about structural phenomena and requirements fo the community.

Comments and questions should be addressed to ISLE@mpi.nl

Peter Wittenburg
Max-Planck-Institute for
Psycholinguistics
Wundtlaan 1
6525 XD
Nijmegen
The Netherlands
Email: pewi@mpi.nl
Web site: http://www.mpi.nl

H. Brugman
Instituto Tecnológico de Informática
Universidad Politécnica de Valencia
Spain
Email: e.vidal@iti.upv.es
Web site: http://www.upv.es/

D. Broeder
Instituto Tecnológico de Informática
Universidad Politécnica de Valencia
Spain
Email: e.vidal@iti.upv.es
Web site: http://www.upv.es/

New Resources

ELRA-W0026 PAROLE Irish Corpus

The PAROLE Irish Distributable Corpus consists of over 8 million words (a subset of the 15+ million words Irish Reference corpus).

The text is marked-up in accordance with the PAROLE encoding standard which incorporates the Corpus Encoding Standard (CES) and Text Encoding Initiative (TEI) Guidelines. All the files are in SGML format with a detailed header and the body of the text tagged to paragraph level. The header includes information such as title, author(s), number of words, ownership, publication details and also a standard coding for Medium, Topic and Genre categories.

A subset of the Distributable Corpus is morpho-syntactically tagged. Included in this distribution is approximately 3,000 manually checked words. Below is a breakdown of the sources of texts.

Medium	No. of texts	Source	No. of Words
Book	196	- An Gúm (state publishing) - Peanntrónaic (typesetting and design company)	5,900,000
Newspaper	109	- Anois (weekly) - Lá (weekly)	2,580,000
Miscellaneous	9	- ITÉ - Aontas Eorpach	278,000
Total			8,758,000

Price: 250 Euro

ELRA-L0043 English PAROLE Lexicon

The English PAROLE Lexicon has been compiled by two partners, Sheffield University and the Corpus Linguistic Group (CLG) at Birmingham University.

The Lexicon was compiled from existing resources: CRL-LKB and the COBUILD dictionary database. Both have restricted availability and contain extensive syntactic, semantic and morphological information.

The lexicon contains 22,000 morphological units, of which 12998 are common nouns, 40 proper nouns 4195 verbs, 3208 adjectives, 606 adverbs, 71 adpositions, 2 articles, 21 conjunctions, 25 determiners, 53 pronouns.

The English PAROLE lexicon comprises the following information: morphological encoding for all nouns, verbs, adverbs, adjectives and function words; syntactic encoding of all verbs, nouns, adjectives and adverbs.

The organizational procedure was as follows: I. Selection: Lemmata were mostly selected on the basis of frequency from the COBUILD corpus. Most proper nouns were deselected and some verbs were added because of the decision to encode deverbal nominalisations and compound information. II. Coverage: the headword list was checked against the resources to make sure there was adequate coverage of syntactic and morphological information. III. Composition: the nominal lemmata were checked for derivations and compounds. These were extracted and analyzed into their constituent parts and compounds were checked for lexicalisation. Components were flagged with their base forms and grammatical class. IV. Conversion: Morphosyntactic information was either directly transferred from existing resources or, in the case of inflectional information and subcategorisation patterns, programs were written to extract information and convert it into the PAROLE format. V. Cross-reference: all components contained in nominal derivations and compounds were cross-referenced with their base PoS. VI. Integrity checks were made and the lexicon was parsed using nsgmls.

	ELRA Members	Non Members
Price for research use	3,400 Euro	5,100 Euro
Price for commercial use	9,000 Euro	13,500 Euro

ELRA-S0096 German SpeechDat(II) MDB-1000

The German SpeechDat(II) MDB-1000 comprises 1295 German speakers (663 males, 610 females, 22 speakers with gender not specified) recorded over the German mobile telephone network. The database was produced by the Department of Phonetics and Speech Communication of the University of Munich under a subcontract with Vocalis Ltd., Cambridge, UK. The MDB-1000 database is partitioned into 8 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: 3 isolated digits; 1 sequence of 10 isolated digits; 4 connected digits: 1 prompt sheet number (5+ digits), 1 telephone number (9-11 digits), 1 credit card number (15-16 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); 3 application words; 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 isolated words.

The following age distribution has been obtained: 34 speakers are below 16 years old, 587 speakers are between 16 and 30, 376 speakers are between 31 and 45, 199 speakers are between 46 and 60, 48 speakers are over 60, and 51 speakers of unknown age.

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

	ELRA Members	Non Members
Price for research use	20,000 Euro	25,000 Euro
Price for commercial use	28,000 Euro	35,000 Euro

ELRA-S0097 British English SpeechDat(II) FDB-4000

The British English SpeechDat(II) FDB-4000 comprises 4000 British English speakers (1968 males, 2032 females) recorded over the British fixed telephone network. The SpeechDat database has been collected by the Signal Processing, Control and Networks Division of the GEC-Marconi Research Centre. This database is partitioned into 20 CDs. 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 connected digits : 1 sheet number (6 digits), 1 telephone number (9-11 digits), 1 credit card number (16 digits), 1 PIN code (6 digits); 1 spontaneous phone number; 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 birth / growing up (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; keyword phrase using an embedded application word; 4 phonetically rich words; 9 phonetically rich sentences.

The following age distribution has been obtained: 1242 speakers are between 16 and 30, 1321 speakers are between 31 and 45, 1298 speakers are between 46 and 60, and 139 speakers of unknown age.

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

	ELRA Members	Non Members
Price for research use	35,000 Euro	45,000 Euro
Price for commercial use	45,000 Euro	55,000 Euro

ELRA-S0098 British English SpeechDat(II) SDB-2400

The British English SpeechDat(II) SDB-2400 is designed for development and assessment of speaker verification and identification systems. It contains 22 utterances for 120 different speakers who called 20 times, collected over the fixed and mobile telephone networks in quiet and noisy environments. The SpeechDat database has been collected by the Signal Processing, Control and Networks Division of the GEC-Marconi Research Centre. This database is partitioned into 8 CDs. 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; 2 connected digits: 1 credit card number (16 digits), 1 PIN code (6 digits); 2 spelled words : 1 fixed spelled "forename surname", 2 spelled "names/words"; 1 fixed "forename surname"; 2 "forename surname" (out of 10); 2 application words; 10 phonetically rich sentences.

The following age distribution has been obtained: 7 speakers are under 16 years old, 41 speakers are between 16 and 30, 33 speakers are between 31 and 45, 32 speakers are between 46 and 60, and 7 speakers of unknown age.

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

	ELRA Members	Non Members
Price for research use	32,000 Euro	39,000 Euro
Price for commercial use	39,000 Euro	47,000 Euro

ELRA-S0095 Slovak SpeechDat(E) Database

The Slovak SpeechDat(E) Database (Eastern European Speech Databases for Creation of Voice Driven Teleservices) comprises 1000 Slovak speakers (498 males, 502 females) recorded over the Slovak fixed telephone network. The database was collected by the Slovak Academy of Sciences in Bratislava, in co-operation with Lernout&Hauspie France. This database is partitioned into 5 CDs. The speech databases made within the SpeechDat(E) project were validated by SPEX, the Netherlands, to assess their compliance with the SpeechDat(E) format and content specifications.

The speech files are stored as sequences of 8-bit, 8kHz A-law speech files and are not compressed, according to the specifications of SpeechDat(E). Each prompt utterance is stored within a separate file and has an accompanying ASCII SAM label file.

Corpus contents: 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 (16 digits), 1 PIN code (6 digits); 3 dates: 1 spontaneous date (birthday), 1 prompted date (word style), 1 relative and general date expression; 1 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 assistance city name; 1 real/artificial name for coverage; 2 currency money amounts: 1 Slovak money amount, 1 International money amount (USD, EURO); 1 natural number; 6 directory assistance names: 1 spontaneous, e.g. own forename; 1 city of birth / growing up (spontaneous); 1 most frequent city (out of 500); 1 most frequent company/agency (out of 500); 1 "forename surname" (set of 150), 1 "surname" (set of 150); 2 questions, including "fuzzy" yes/no: 1 predominantly "yes" question, 1 predominantly "no" question; 12 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: 39 speakers are below 16 years old, 446 speakers are between 16 and 30, 253 speakers are between 31 and 45, 214 speakers are between 46 and 60, and 48 speakers are over 60.

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

Price for research use by a Slovak organisation	7,500 Euro
Price for research use	10,000 Euro
Price for commercial use	16,000 Euro

ELRA-S0094 Czech SpeechDat(E) Database

The Czech SpeechDat(E) Database (Eastern European Speech Databases for Creation of Voice Driven Teleservices) comprises 1052 Czech speakers (526 males, 526 females) recorded over the Czech fixed telephone network. The database was collected by the Institute of Radioelectronics of Brno University of Technology (VUT) and by the Department of Signal Theory of Czech Technical University (CVUT) Prague, in co-operation with Lernout&Hauspie France. This database is partitioned into 6 CDs. The speech databases made within the SpeechDat(E) project were validated by SPEX, the Netherlands, to assess their compliance with the SpeechDat(E) format and content specifications.

The speech files are stored as sequences of 8-bit, 8kHz A-law speech files and are not compressed, according to the specifications of SpeechDat(E). Each prompt utterance is stored within a separate file and has an accompanying ASCII SAM label file.

Corpus contents: 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 (birthday), 1 prompted date (word style), 1 relative and general date expression; 1 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 assistance city name; 1 real/artificial name for coverage; 2 currency money amounts: 1 Czech money amount, 1 International money amount (USD, EURO); 1 natural number; 6 directory assistance names: 1 spontaneous, e.g. own forename; 1 city of birth / growing up (spontaneous); 1 most frequent city (out of 500); 1 most frequent company/agency (out of 500); 1 "forename surname" (set of 150), 1 "surname" (set of 150); 2 questions, including "fuzzy" yes/no: 1 predominantly "yes" question, 1 predominantly "no" question; 12 phonetically rich sentences; 2 time phrases: 1 time of day (spontaneous), 1 time phrase (word style); 4 phonetically rich words; 4 additional questions (spontaneous).

The following age distribution has been obtained: 20 speakers are below 16 years old, 490 speakers are between 16 and 30, 238 speakers are between 31 and 45, 230 speakers are between 46 and 60, 71 speakers are over 60, and 3 speakers of unknown age.

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

Price for research use by a Czech organisation	7,500 Euro
Price for research use	10,000 Euro
Price for commercial use	16,000 Euro

ELRA-S0099 Russian SpeechDat(E) Database

The Russian SpeechDat(E) Database (Eastern European Speech Databases for Creation of Voice Driven Teleservices) comprises 2500 Russian speakers (1242 males, 1258 females) recorded over the Russian fixed telephone network. The database was collected by AudiTech Ltd. (Russia). This database is partitioned into 13 CDs. The speech databases made within the SpeechDat(E) project were validated by SPEX, the Netherlands, to assess their compliance with the SpeechDat(E) format and content specifications.

The speech files are stored as sequences of 8-bit, 8kHz A-law speech files and are not compressed, according to the specifications of SpeechDat(E). Each prompt utterance is stored within a separate file and has an accompanying ASCII SAM label file.

Corpus contents: 6 application words; 1 sequence of 10 isolated digits; 4 connected digits: 1 sheet number (5 digits), 1 telephone number (9-10 digits), 1 credit card number (14-16 digits), 1 PIN code (6 digits); 3 dates: 1 spontaneous date (birthday), 1 prompted date (word style), 1 relative and general date expression; 1 spotting phrase using an application word (embedded); 1 isolated digit; 3 spelled-out words (letter sequences): 1 spelling of surname, 1 spelling of directory assistance city name, 1 real/artificial name for coverage; 2 currency money amounts: 1 Russian money amount, 1 International money amount (USD, EURO); 1 natural number; 6 directory assistance names: 1 spontaneous (own forename), 1 city of birth / growing up (spontaneous), 1 most frequent city (out of 500), 1 most frequent company/agency (out of 500), 1 "forename surname" (set of 150), 1 "surname" (set of 150); 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: 10 speakers are below 16 years old, 854 speakers are between 16 and 30, 858 speakers are between 31 and 45, 679 speakers are between 46 and 60, 34 speakers are over 60, and 65 speakers are of unknown age.

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

ELRA Members	20,000 Euro
Non Members	25,000 Euro

ELRA-S0100 MHATLex

MHATLex is a new enhanced lexical resource for written and speech automatic processing for French (see article p. 8). It is derived from BDLex (see ELRA-S0003 and S0004). It contains three levels of representation: Syntactic level: S; Phonological word level: W; Phonetic level: P.

At the W level, a word has two representations: input representation (W representation) where words are simply imported from the lexicon; output representation (W' or phonotypical) where words have the phonotypical representation imposed by their context in the sentence. The lexicons contain inflected words (among which canonical words).

Type of entry	Number of entries	
	MHATLe	MHATLexSt (& BDLex) MHATLexW
Canonical	W81,456	49,962
Inflected	854,452	437,998

Words are represented with their orthography, pronunciation, morpho-syntactic features, and frequency indicator (L23 if the word is derived from the most frequent 23,000 canonical words, which corresponds to BDLex 23000). Only the pronunciation related part changes according to the lexicon (except if the user want to generate his own lexicon by skipping some features). Four lexicons can be generated from MHATLex: MHATLexW : this is the central lexical resource which enables to generate the other lexicons; MHATLexW' (or MHATLexPht) : gives the word representations for each pertinent context; MHATLexSt : with standard and simplified format of the pronunciation; BDLex (or BDLex50): already distributed by ELDA (ELRA-S0003 and S0004). The current BDLex, derived from MHATLexW, contains some updates.

	ELRA Members	Non Members
Price for research use	1,500 Euro	2,500 Euro
Price for commercial use	5,000 Euro	7,500 Euro