The ELRA Newsletter

April - June 2000

Vol.5 n.2

Contents

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

Editor in Chief:
Khalid Choukri

Editor:
Jeffrey Allen

Layout:
Audrey Mance

Contributors:
Dr Els den Os
Nelleke Oostdijk

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,

You are about to read the issue of the ELRA Newsletter printed just a few days before the LREC-2000 Conference. I hope that you can attend it this year and enjoy what will be a great event about Language Resources and Evaluation.

ELRA and ILSP are happy to welcome you to Athens from May 28 until June 3. Both May 28 and 29 will be devoted to the LREC satellite workshops. The main conference will take place from May 30 till June 2. A post-workshop is planned for Saturday June 3.

The large number of papers accepted for the conference (about 280, compared to 200 presented at LREC'98) illustrates the variety of issues that need to be addressed and discussed in forums like LREC. The LREC2000 programme is being updated regularly. A preliminary version is enclosed in this issue (page 6) and will be soon available at http://www.elda.fr/lrec2000.html, along with a detailed programme. The "Workshops" section is being updated according to the information provided by each workshop organiser. Many industrial and academic institutions involved in Natural Language and Speech Processing will be present at the LREC2000 Exhibition which will take place in parallel with the conference.

During this quarter, we devoted a lot of effort to the organisation of LREC but also to the organisation of the General Assembly, which took place in Paris on 27 March 2000 at La Grande Arche de la Défense.

In addition to the adoption of the management and financial reports, the General Assembly adopted the changes of the statutes suggested by the board. These changes will allow ELRA to acquire and sell shares in organisations actively involved in the field of Language Resources. Such acquisitions or sales have to be agreed upon by the General Assembly.

During the General Assembly, the board was renewed and we would like to express our warm thanks and gratitude to the members of the board who left: G. Carayannis (ILSP, Greece), H. van den Heuvel (SPEX, The Netherlands), B. Maegaard (CST, Denmark), T. Schneider (Consultant, Germany), H. Sonneveld (Topterm V.O.F., The Netherlands). We would also like to welcome those who joined the board: C. Fluhr (CEA, France), P. Heisterkamp (DaimlerChrysler, Germany), P. Isabelle (Xerox, France), T. McEnery (Lancaster University, U.K.), J. Odijk (L&H, Belgium). The new statutes and short biographies of the new board will be soon available on the web. The General Assembly minutes will be mailed to our members.

In this issue of the Newsletter, we include two articles. The first one, by Dr Els den Os (KPN Research, the Netherlands) deals with Babelweb, a EURESCOM project aiming at giving practise guidelines for the use of tools and architectures for the handling of multilingual web sites. The second article, by Nelleke Oostdijk (University of Nijmegen, The Netherlands), presents the Spoken Dutch Corpus Project, which aims at promoting Dutch through the production of Language Resources.

As usual, the final section includes a list of newly acquired Language Resources:

- ELRA S0034 Verbmobil;
- ELRA-S0058 RVG1 (Regional Variants of German 1);
- ELRA-S0081 Norwegian SpeechDat(II) FDB-1000;
- ELRA-S0082 Siemens Synthesis Corpus - SI1000P;
- ELRA S0083 ISLE Speech Corpus;
- ELRA-L0033 LusoLEX European Portuguese Lexicon;
- ELRA-L0034 BrasiLEX Brazilian Portuguese Lexicon;
- ELRA-L0035 PAROLE Portuguese Lexicon;
- ELRA-W0024 PAROLE Portuguese Corpus.

The text corpus ELRA-W0015 Le Monde for the year 1999 is now available.

Antonio Zampolli, President Khalid Choukri, CEO

BABELWEB, a Eurescom project on guidelines for multilingual web sites

Dr Els den Os, KPN Research, the Netherlands

In June 1999 the research institutes of five network operators, British Telecom, Portugal Telecom, France Telecom, Telecom Italia, and Royal Dutch Telecom (KPN), started a project concerning the design and management of multilingual web sites, called 'Babelweb'. The total duration of Babelweb was planned to be 18 months. EURESCOM is an institute for performing collaborative projects on research and strategic studies in all areas of telecommunications. Currently, there are 24 network operators and service providers from 23 European countries involved in EURESCOM.

The Telecommunications industry has developed a long way since its only goal was the provision of a basic telephony service. Today, telecommunications operators see the Internet as providing the key infrastructure for both new and traditional services. The result of Babelweb will be better understanding of types of web architectures that optimally support the development and maintenance of multilingual sites, thus reducing the costs of multilingual web services. In addition the project will provide detailed insight into the possibilities of (commercial) language technology (tools) that can be used for multilingual web site creation and use. Babelweb will give best practise guidelines for the use of tools and architectures for the handling of multilingual web sites for different applications and domains. Finally, the project will give some guidelines for multilingual human factors issues, especially related to language choice and navigation, and to the acceptability of machine translated pages.

From the very beginning, the project distinguished between making web sites multilingual and making multilingual web sites. In the first case one can translate (or localise) all (or part) of the information in an existing monolingual site to other languages. In the second case one can create web sites to be multilingual from their first inception. It is expected that in the second case it will be easier to incorporate language technology tools in the web architecture. We think that is better to use the term localisation instead of translation. The term localisation is commonly used in software development and it means in our case adapting a certain web design for a specific locale, which normally involves more than just translation of textual information.

At the end of March 2000, the first, exploratory phase of the project was in its final state. The following four tasks were carried out.

TASK 1: Identification of important issues related to multilingual web services

This task listed the most important factors related to multilinguality in the Internet, by reviewing existing multilingual sites and by interviewing web and service managers within the five network operators. It is very obvious that multilinguality on the Internet is quickly becoming an important topic for all service providers in the world. At the moment, the Internet contains millions of web sites and about 90% of them are in English. However, it is expected that the non-English speaking web users will outnumber the English-speaking users already during this year. This means that much effort has to be put into localisation of existing web sites and into the creation of new multilingual services, since it is certain that most web users prefer to be addressed in their native language, at least at the top-level pages of services. These top-level pages need to be perfect, since otherwise the risk is high that the customer will gain a poor impression of the company. In the multilingual Internet the motto "The competition is just a mouse click away" is very true indeed. Multilingual web sites may cost a lot of money and effort, since for most services it is inevitable that human translators are involved. Our review and interviews showed that localisation often is an ad hoc process.

TASK 2: Inventory of tools for building and maintaining multilingual web sites

There is a number of web architecture tools available that support the creation of multilingual sites. One of the aims of Babelweb is to investigate whether language technology tools can enrich these tools.

This task in the project gave an overview of web site management tools, of language technology tools, and of what we have called 'voice input and output' on the Internet. The architectural tools are intended to provide procedures and support for managing web sites in a multilingual environment and also address issues like multilingual fonts and characters. The language technologies of interest are: Multilingual HTML editors, Machine Translation, Translation Memories, Text Summarisation, Thematic Analysis, Language Identification, Information Extraction, Language Generation, and Cross-linguistic Information Retrieval. Most of these technologies can be used in an on-line and off-line manner. This means that they can both support the creation of multilingual sites, and that users can use them interactively while surfing the web. The last activity within this task provided an overview of the possibilities related to voice browsers that may use (a combination of keyboard and) speech recognition for input, and pre-recorded speech files or speech synthesis for output. In this way the users can have access to the Web anytime and anywhere (at home, on the move or at work). By using only a fixed or cellular phone, they can choose to interact by a keystroke or a spoken command. The overview comprises an inventory of the tools that allow voice access to Web pages, a view on mark-up programming languages used to develop Web page browsing by voice, and a description of the scientific community's activities for developing Voice Browsers.

TASK 3: Architectures for Multilingual Web sites

The third task defined possible web architectures for a number of services. In order to be able to do this we have broken down a typical multilingual web service into five 'functional elements':

1. information provision; display of information in the form of natural language phrases, sentences, paragraphs, and larger documents
2. user input; getting information from/about the user
3. searching; mono and cross language information retrieval
4. e-commerce; multilingual issues involved in the use of currencies, measures, legal issues, (pragmatic distribution)
5. multimedia information; graphics, sounds, etc.

For each of these elements the impact on the web architecture was first defined at a rather high level of abstraction. Detailed implementation will be taken into account when building demonstrators in the next phase of the project.

TASK 4: Human factors issues in multilingual web services

In order to get an idea about the human factors issues related to multilingual sites, we built and tested a mock-up multilingual (English-French) service. This service, called 'EURESCOM meeting support', helps researchers to plan their trip abroad. The advantage of this service is that it combines internal sites (with for example information about the lab to be visited) and external sites (e.g. hotel sites, weather forecast sites). Such a combination reflects the reality of Internet surfing. 'Internal' sites are sites that are managed by the provider of a multilingual service, and that are therefore accessible at an arbitrarily low level. 'External' sites are all the web sites that are simply linked, and therefore lie outside the direct influence of the provider of a multilingual web service. The use of those sites is often limited to what can be accessed through the public interface.

We conducted a user study with French and English subjects who had to perform a number of tasks that addressed the issues of language choice/navigation and acceptability of machine translated pages. The results indicate that these users prefer to have language choice on each page. However, the choice should not appear very dominantly on each page, since it invites users to click on the language button even when they already are in the appropriate language version. Machine Translation was accepted by both groups of subjects. The acceptability was less for a site that contained only text, compared to a site that also showed pictures. Interestingly, it took one-third longer to read a machine-translated text than the same text translated manually.

Dr Els den Os
Babelweb project leader
KPN Research
System Integration and Application for Multimedia
Postbus 421
2260XZ Leidschendam
The Netherlands
Tel.: + 31 70 332 62 82
Fax: +31 70 332 64 77
E-mail: e.a.denos@kpn.com
Web site:
http://www.eurescom.de/Public/Projects/P900-series/P923/p923.htm

The Spoken Dutch Corpus Project

Nelleke Oostdijk, University of Nijmegen, The Netherlands

Introduction

The Spoken Dutch Corpus project is a five-year project that started in June 1998. It is funded jointly by the Flemish and Dutch governments and Science Foundations with a budget of some 4.6 MEuro. The project aims at the compilation and annotation of a 10-million-word corpus of contemporary standard Dutch as spoken in the Netherlands and Flanders. The entire corpus will be orthographically transcribed, lemmatized and annotated with part-of-speech information. For a selection of one million words further, more detailed annotations are envisaged, including a broad phonetic transcription and a syntactic annotation. To enable effective access to the speech recordings, the transcriptions will be enriched with pointers into the speech files. The automatic time alignment will be manually checked on the word level for that part of the corpus for which a verified phonetic transcription is available.

Background, motivation

Standard Dutch is the official language in the Netherlands (some 15 million people speak northern standard Dutch) and in Flanders (the northern part of Belgium, 5.6 million people speak southern standard Dutch)[1]. While variants of the same language, there are considerable differences between northern standard Dutch and southern standard Dutch. These differences occur with regard to syntax, morphology, lexis and phonetics/phonology (cf. Donaldson, 1983; Van de Velde et al., 1998).

As one of the smaller languages in Europe, Dutch is under serious threat of gradually disappearing as it is losing ground to English. The availability of the necessary resources has placed the English-language-based language and speech technology in the leading position it holds today and has thus further strengthened the position of English for business communication. The fact that to date for Dutch few relevant language resources are available forms a serious complication for the advancement of Dutch language and speech technology (cf. Bouma and Schuurman, 1998). The present project seeks to ameliorate this situation.

Apart from the interests held by language and speech technologists, the corpus is intended to serve several other research interests. The corpus addresses the needs of linguists from various backgrounds. So far for Dutch the only more or less substantial data collections derive from written sources. As a consequence, studies of Dutch linguistics in the past have focused on the written language, leaving the spoken language rather poorly documented. Another field in which the corpus will be of significant use is that of education. The insights that can be gained into everyday actual language use are indispensable for developing Dutch language courses and course materials.

Project organization

The Spoken Dutch Corpus project is directed by a board whose members include representatives of the two governments, the Dutch Language Union[2], Dutch and Flemish research foundations and one of the Dutch national research schools (LOT). The Chairman of the board is Professor W. Levelt of the Max Planck Institute for Psycholinguistics in Nijmegen, The Netherlands.

Appointed by the board there is a steering committee consisting of experts from various linguistics (sub)disciplines and expert language and speech technologists, that is responsible for the project progress and finances.

Project activities are coordinated from two sites: Ghent for Flanders and Nijmegen for the Netherlands. Each site is managed by a project leader. The project leaders in collaboration with three specialist working groups (one for corpus design and compilation, one for signal processing and one for corpus annotation) are responsible for the design and implementation of the various project activities.

Project outline and timetable

The project aims to compile a 10 million word corpus that will constitute a plausible

sample of contemporary standard Dutch as spoken in Flanders and the Netherlands. One third of the data will be collected in Flanders, two thirds will originate from the Netherlands. The entire corpus will be transcribed orthographically, lemmatized and tagged with part-of-speech information. Users will be able to access the speech recordings through pointers in the transcriptions. For a selection of one million words it is envisaged that an auditorily verified, broad phonetic transcription will be available, while for this part of the corpus the automatic time alignment will be manually checked on the level of the word. For most recordings which are not checked by hand the pointers are expected to be accurate within less than 100 ms. Also for one million words a syntactic annotation will be available and 250,000 words will receive a prosodic annotation.

The first year of the project has been devoted to corpus design, the development of various protocols and annotation schemes, and the selection and adaptation of tools and supporting resources. During this year also a 50,000-word pilot corpus was compiled which was used for testing purposes.

Over the remaining four years the corpus will be compiled, transcribed and annotated incrementally in eight six-month periods. At the end of each period part of the material will be released. Thus the data will be available to users from an early stage onward, while the project may benefit from the feedback given by these users.

Corpus design

The design of the corpus was guided by a number of considerations. First of all, there is the fact that the corpus must serve many and rather diverse interests. Different user groups have different requirements when it comes to the quality and quantity of the data, the number and type of speakers, and so on. Second, the total budget available for the entire project is fixed at 4.6 MEuro, i.e. this should cover all costs involved in recording and collecting data, transcribing and annotating these data, etc. And finally, the issue of copyright complicates matters. Since the corpus will be distributed including the speech files, the consent of all speakers is required as well of any other parties that have any rights to the recorded material.

The design of the corpus takes into account the various dimensions underlying the variation that can be observed in language use. In the overall design of the corpus the principal parameter is taken to be the socio-situational setting in which language is used. This leads us to distinguish a number of components, each of which can be characterized in terms of its situational characteristics such as communicative goal, medium, number of speakers participating, and the relationship between speaker(s) and hearer(s). The specification of each of the components is given in terms of sample sizes, total number of speakers, range of topics, etc. Where this is considered to be particularly interesting speaker characteristics such as gender, age, geographical region, and socio-economic class are used as sampling criteria, otherwise they are merely recorded as part of the meta-data. The overall design of the corpus is given in Table 1 (see page 8).

Recording and collecting data; digitization

Ten million words of data amount to roughly 1,000 hours of speech. The recordings are obtained in a variety of ways. Where, as in the case of broadcast data, recordings (sometimes accompanied by rough transcripts) can be obtained through other parties, contracts are negotiated that allow us to use the data. For components such as the direct face-to-face conversations, volunteers are recruited and asked to participate in the recording of conversations in their home environment, while a relatively small group of people is instructed to go out and record in a variety of settings (in shops, at work, in a restaurant, etc.). For yet other components, such as the lectures, research assistants working for the project contact the schools or colleges, ask permission and make the necessary arrangements for them to come and do the recording on site. On occasion there are collaborative actions where the Spoken Dutch Corpus project obtains data through other projects, as in the case of the private interviews that have been recorded within the project The pronunciation of Standard Dutch. Varieties and variants in Flanders and the Netherlands (Van de Velde et al. 1998). All recordings are digitized. Information about the recording conditions, the equipment that was used, etc. is recorded as part of the meta-data.

Orthographic transcription

For all recordings a verbatim transcript will be available[3]. To facilitate the transcription process, use is made of the interactive signal processing tool PRAAT[4]. During the transcription process, transcribers segment the audio files in relatively short chunks by inserting time markers in unfilled pauses between words. At a later stage these markers are used as anchor points for the automatic alignment of the transcript and the speech file.

Lemmatization and part-of-speech (POS) tagging

After an evaluation of taggers and tagsets available for Dutch, it was decided to define a tagset for Dutch that would conform to the EAGLES guidelines and would be compatible with the authoritative Dutch reference grammar, viz. the ANS (Haeseryn et al., 1997). The tagset distinguishes ten major word classes, while with each of these word classes additional morpho-syntactic features are recorded. In all, the tagset consists of some 300 tags. For the tagging process a tagger has been developed which assigns the most likely tag for a word in a given context. All output is manually checked and - where necessary - corrected. Apart from the POS tag, for each word also the associated lemma is given.

Phonetic transcription

For the broad phonetic transcription of the data, use is made of the SAMPA set. In order to speed up the transcription process and also to maximize consistency, transcribers are provided with an automatically generated transcript which they are asked to verify and/or correct.

Syntactic annotation

An annotation scheme is currently being developed. The syntactic analyses will contain information about the constituent structure of major linguistic units as well as the syntactic relations that hold between constituents. Syntactic annotation will be carried out semi-automatically, using the ANNOTATE software[5].

Dissemination of the results

During the project, prospective users are kept informed about its progress by means of a newsletter and a website[6]. Intermediate results of the project will be made available at regular (roughly) six-month intervals. The first release of the first part of the corpus was on March 1st, 2000. The date for the second release is set for September 1st, 2000. On a regular basis workshops and seminars are organized at which progress reports are presented and results discussed

(continued on page 8)

Language Resources and

31 May - 2
Athens,

Preliminary

Wednesday May 31, 2000

09:00	**2 oral sessions in parallel**	
	Opening Session *Room 4*	
09:40	*5 min break*	
09:45	**4 oral sessions in parallel**	
	Panel 5: Speech Database Processing Tools: the state of the art in automatic labeling of speech, *Campbell* *Nick*	*Room A*
	Session WO13: Multilingual Resources and Applications	*Room B*
	Lunch	
11:05	*coffee break*	
11:20	**3 oral sessions in parallel**	
	Session WO15: Language Resources Projects	*Room B*
	Session WO16: Corpus Annotation and Information	*Room C*
	Session EO4: Grammars and Systems Evaluation	*Room D*
12:20	**6 poster sessions in parallel**	
	Session SP4: Tools for Evaluation and Processing of Spoken Language Resources	*Peristylion*
	Session SP5: Multimodal - Multimedia Resources and Tools	*Peristylion*
	Session WP7: Corpus Projects	*Peristylion*
	Session WP8: Corpus Tools	*Peristylion*
	Session WP9: Applications using Written Language	*Peristylion*
	Session EP1: Evaluation and Written Area	*Peristylion*
13:40		
15:00	**4 oral sessions in parallel**	
	Session SO6: Recognition	*Room A*
	Session WO17: Semantic Lexicons	*Room B*
	Session WO18: Morphology in Lexical and Textual Resources	*Room C*
	Session EO5: Information Retrieval and Question Answering Evaluation	*Room D*
17:20	*coffee break*	
17:40	Closing Session	

Thursday June 1, 2000 (a.m.)

09:00	**2 oral sessions in parallel**	
	Invited speaker	*Room A*
	Invited speaker	*Room B*
09:40	*5 min break*	
09:45	**4 oral sessions in parallel**	
	Panel 5: Speech Database Processing Tools: the state of the art in automatic labeling of speech, *Campbell* *Nick*	*Room A*
	Session WO13: Multilingual Resources and Applications	*Room B*
	Session WO14: Named Entity Recognition	*Room C*
	Session EO3: Evaluation and Semantics	*Room D*
11:05	*coffee break*	
11:20	**3 oral sessions in parallel**	
	Session WO15: Language Resources Projects	*Room B*
	Session WO16: Corpus Annotation and Information	*Room C*
	Session EO4: Grammars and Systems Evaluation	*Room D*

d Evaluation Conference

June, 2000
Greece

Programme

Thursday June 1, 2000 (p.m.)

09:00	**2 oral sessions in parallel**	
	Invited speaker	*Room A*
	Invited speaker	*Room B*
09:40	*5 min break*	
09:45	**4 oral sessions in parallel**	
	Panel 5: Speech Database Processing Tools: the state of the art in automatic labeling of speech *Nick*	
	lunch	
	Session WO13: Multilingual Resources and Applications	*Room B*
	Session WO14: Named Entity Recognition	*Room C*
	Session EO3: Evaluation and Semantics	*Room D*
11:05	*coffee break*	
11:20	**3 oral sessions in parallel**	
	Session WO15: Language Resources Projects	*Room B*
	Session WO16: Corpus Annotation and Information	*Room C*
	Session EO4: Grammars and Systems Evaluation	*Room D*

Friday June 2, 2000

09:00	**2 oral sessions in parallel**	
	Invited speaker	*Room A*
	Invited speaker	*Room B*
09:40	*5 min break*	
09:45	**4 oral sessions in parallel**	
	Panel 5: Speech Database Processing Tools: the state of the art in automatic labeling of speech, *Campbell* *Nick*	*Room A*
	Session WO13: Multilingual Resources and Applications	*Room B*
	Session WO14: Named Entity Recognition	*Room C*
	Session EO3: Evaluation and Semantics	*Room D*
11:05	*coffee break*	
11:20	**3 oral sessions in parallel**	
	Session WO15: Language Resources Projects	*Room B*
	Session WO16: Corpus Annotation and Information	*Room C*
	Session EO4: Grammars and Systems Evaluation	*Room D*
12:20	**6 poster sessions in parallel**	
	Session SP4: Tools for Evaluation and Processing of Spoken Language Resources	*Peristylion*
	Session SP5: Multimodal - Multimedia Resources and Tools	*Peristylion*
	Session WP7: Corpus Projects	*Peristylion*
	Session WP8: Corpus Tools	*Peristylion*
	Session WP9: Applications using Written Language	*Peristylion*
	Session EP1: Evaluation and Written Area	*Peristylion*
13:40	*lunch*	
15:00	**4 oral sessions in parallel**	
	Session SO6: Recognition	*Room A*
	Session WO17: Semantic Lexicons	*Room B*
	Session WO18: Morphology in Lexical and Textual Resources	*Room C*
	Session EO5: Information Retrieval and Question Answering Evaluation	*Room D*
17:20	*coffee break*	
17:40	Closing Session	
21:00	GALA	*Hilton Hotel*

		Synsets	No. of senses	Sens./ syns.	Entries	Sens./ entry	LIRels.	LIRels/ syns	EQRels-ILI	EQRels /syn	Synsets without ILI
Dutch Wordnet	Nouns	34455	54428	1,58	45972	1,18	84869	2,46	26724	0,78	6070
	Verbs	9040	14151	1,57	8826	1,60	25973	2,87	26724	2,96	1133
	Other	520	1622	3,12	1485	1,09	797	1,53	n.a.	n.a.	n.a.
	Total	44015	70201	1,59	56283	1,25	111639	2,54	53448	1,21	7203
Spanish Wordnet	Nouns	18577	41292	2,22	23216	1,78	40559	2,18	18634	1,00	0
	Verbs	2602	6795	2,61	2278	2,98	3749	1,44	2602	1,00	0
	Other	2191	2439	1,11	2439	1,00	10855	4,95	n.a.	n.a.	n.a.
	Total	23370	50526	2,16	27933	1,81	55163	2,36	21236	0,91	0
Italian Wordnet	Nouns	30169	34552	1,15	24903	1,39	83021	2,75	43848	1,45	98
	Verbs	8796	12473	1,42	6607	1,89	30757	3,50	27941	3,18	0
	Other	1463	1474	1,01	1468	1,00	3290	2,25	n.a.	n.a.	n.a.
	Total	40428	48499	1,20	32978	1,47	117068	2,90	71789	1,78	1561
French Wordnet	Nouns	17826	24499	1.37	14879	1.65	39172	2.20	17815	1.00	16
	Verbs	4919	8310	1.69	3898	2.13	10322	2.10	4915	1.00	4
	Other	0	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
	Total	22745	32809	1.44	18777	1.75	49494	2.18	22730	1.00	20
German Wordnet	Nouns	9951	13656	1.37	12746	1.07	23856	2.40	10570	1.06	0
	Verbs	5166	6778	1.31	4333	1.56	10960	2.12	5762	1.12	0
	Other	15	19	1.27	19	1,00	2	0.13	15	1.00	0
	Total	15132	20453	1.35	17098	1.20	34818	2.30	16347	1.08	0
Czech Wordnet	Nouns	9727	13829	1.42	9277	1.49	19856	2.04	9729	1.00	0
	Verbs	3097	6120	1.98	3006	2.04	6403	2.07	3097	1.00	0
	Other	0	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
	Total	12824	19949	1.56	12283	1.62	26259	2.05	12824	1.00	0
Estonian Wordnet	Nouns	5028	8226	1.64	7209	1.14	10873	2.16	5683	1.13	0
	Verbs	2650	5613	2.12	3752	1.50	5445	2.05	3321	1.25	0
	Other	0	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
	Total	7678	13839	1.80	10961	1.26	16318	2.13	9004	1.17	0
English	Nouns	4751	14188	2,99	2524	5,62	20707	4,36	n.a.	n.a.	n.a.
	Verbs	11363	25761	2.27	14726	1,75	21070	1,85	n.a.	n.a.	n.a.
						,13	363	1,47	n.a.	n.a.	n.a.

Table 1. *Overall design of the corpus*

(continued from page 5)

and evaluated. Upon completion of the project, the corpus including the recordings will be distributed on CD-ROM through ELRA.

For more information, please contact the Spoken Dutch Corpus secretariat at the following address:
Bureau Corpus Gesproken Nederlands
NWO, Geesteswetenschappen
Ms. A. Dijkstra
P.O. Box 93120
2509 AC The Hague
The Netherlands
Email: dijkstra@nwo.nl

Notes

This publication was supported by the Netherlands Organization for Scientific Research (NWO) under grant number 014-17-510.

1. In addition, Dutch is the official language in Surinam and the Dutch Antilles. However, since it concerns very small populations (some 360,000 en 240,000 speakers respectively), who use Dutch predominantly in formal settings, these have not been included.

2. The Dutch Language Union was founded in 1980 and is the result of a treaty between Flanders and the Netherlands concerning their language policy. In the case of the Spoken Dutch Corpus it is the Dutch Language Union which holds all rights.

3. The orthographic transcription conforms to a large extent to standard spelling conventions. A protocol has been developed which describes what to transcribe and how to deal with new words, dialect, mispronunciations, and so on. See Goedertier and Goddijn (2000). At present, the protocol is in Dutch. An English motivation will be available shortly.

4. For more information on PRAAT see http://www.fon.hum.uva.nl/praat/

5. More information on ANNOTATE can be found at http://www.coli.uni-sb.de/sfb378/negra-corpus/annotate.html

6. http://lands.let.kun.nl/cgn/

References

Bouma, G. and Schuurman, I., (1998). De positie van het Nederlands in Taal- en Spraaktechnologie. Rapport in opdracht van de Nederlandse Taalunie.

Donaldson, B.C., (1983). Dutch. A Linguistic History of Holland and Belgium. Leiden: Martinus Nijhoff.

Goedertier, W. and Goddijn, S., (2000). Protocol voor orthografische transcriptie. CGN Internal publication. See also http://lands.let.kun.nl/cgn/release1/info/paginas/gg_2000.htm

Haeseryn, W., Romijn, K., Geerts, G., de Rooij, J., and van den Toorn, M.C., (1997). Algemene Nederlandse Spraakkunst. Groningen: Martinus Nijhoff.

Van de Velde, H., De Schutter, G., van Hout, R., Adank, P., Huinck, W. and Op 't Eynde , L., (1998). 'The Pronunciation of Standard Dutch in Flanders and the Netherlands'.

Nelleke Oostdijk
Dutch project manager
Spoken Dutch Corpus
University of Nijmegen, The Netherlands
E-mail: n.oostdijk@let.kun.nl

New Resources

ELRA-S0058 RVG1 (Regional Variants of German 1) - Extension

The corpus consists of single digits, connected digits, phone numbers, phonetically balanced sentences, computer command phrases and spontaneous speech.

Each of the 498 speakers has read a subcorpus of 85 items:

· 11 single digits (0-9, with the two pronunciations of 2 ('zwei', 'zwo'),

· 19 connected digits (10-19, 20-100 in steps of ten),

· 12 computer command phrases,

· 30 phonetically balanced sentences,

· 5 6-digit phone numbers,

· 5 7-digit phone numbers,

· 2 phone numbers with area code,

· 1 minute spontaneous speech (monologue).

The speaker was placed in front of a standard IBM-compatible PC. The backround noise was limited to the usual noise in office environment, eg. door slam, backround crosstalk, phone ringing, paper rustle, PC noise, etc. The head of the speaker is in a range between 2-4 feet to the screen, 1-2 feet from the desktop microphones. The speaker is not forced into a special position. The speaker is wearing a Sennheiser HD 410 and is free to use the keyboard or the mouse in front of him. The three desktop microphones are: Sennheiser MD 441 U, Telex (Soundblaster) and Talk Back (AT&T). Speakers were selected to achieve the demoscopic density of the German spoken areas in Europe (including Austria and Switzerland).

The recorded sound samples are stored in NIST SPHERE format. The resolution is 16 Bits. The sampling frequency is 22.050 Hz except for speakers 001 to 036 which were recorded with 11.025 Hz. Each microphone channel is stored into a separate file. A transliteration of spontaneous speech according to Verbmobil Format is also provided.

RVG1, Part 1 contains 498 speakers recorded through high quality microphones.

RVG1, Part 2, contains 419 speakers recorded through low quality microphones.

	ELRA Members	Non Members
For research use	7,925.02 Euro	11,759.71 Euro
For commercial use	11,759.71 Euro	16,872.63 Euro

ELRA-S0083 ISLE Speech Corpus

This corpus contains approximately 20 minutes of speech (per speaker) from 23 German and 23 Italian intermediate learners of English. Each speaker recorded sentences from several blocks of various types (reading simple sentences, using minimal pairs, giving answers to multiple choice questions). The prompts were of varying perplexities.

About 2/3 of the data for each speaker was annotated by a team of linguists. The files were corrected first at the word level, and an automatic recogniser was then used to produce phone-level annotations. The annotator then re-annotated each sentence to mark phone and stress errors (e.g. substitutions, insertions, or deletions).

Corpus details:

· a total of 46 speakers (23 German and 23 Italian)

· 11484 utterances

· 1.92 gigabytes of WAV files (4 CDs)

· 17 hours, 54 minutes, and 44 seconds of speech data.

	ELRA Members	Non Members
For research use	50.00 Euro	100.00 Euro
For research use by a commercial organisation.	500.00 Euro	1,500.00 Euro

A much more detailed explanation of the ISLE corpus will be available in the proceedings of LREC 2000. An electronic copy of this paper may be obtained at ELRA (Reference: W. Menzel, E. Atwell, P. Bonaventura, D. Herron, P. Howarth, R. Morton, and C. Souter (in preparation). "The ISLE corpus of non-native spoken English", Proc. Second International Conference on Language Resources and Evaluation).

ELRA-S0081 Norwegian SpeechDat(II) FDB-1000

The Norwegian SpeechDat(II) FDB-1000 comprises 1016 Norwegian speakers (517 males, 499 females) recorded over the Norwegian fixed telephone network. The SpeechDat database has been collected and annotated by Telenor Research and Development. The FDB-1000 database is partitioned into 4 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 numbers : 1 sheet number (8 digits), 1 telephone number (8 digits), 1 credit card number (16 digits), 1 PIN code (6 digits)
· 1 currency money amount
· 2 natural numbers
· 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 artificial letter sequence for coverage
· 5 directory assistance utterances : 1 spontaneous own forename, 1 city of calling (spontaneous), 2 city names, 1 common forename and surname
· 2 yes/no questions : 1 predominantly "yes" question, 1 predominantly "no" question
· 6 application words
· 1 word spotting phrase using an embedded application word
· 4 phonetically rich words
· 9 phonetically rich sentences
· 1 additional sentence

The following age distribution has been obtained: 3 speakers are below 16 years old, 301 speakers are between 16 and 30, 363 speakers are between 31 and 45, 195 speakers are between 46 and 60, 137 speakers are over 60, and 17 speakers whose age is unknown.

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

	ELRA Members	Non Members
For research use	15,000.00 Euro	25,000.00 Euro
For commercial use	18,000.00 Euro	25,000.00 Euro

ELRA-S0082 Siemens Synthesis Corpus - SI1000P

The SI1000P recordings were done to provide material for high quality concatenate speech synthesis. It contains 1000 newspaper sentences read by two German professional broadcasting announcers in studio quality together with the laryngographic signal and the glottal pulse stream. Parts of the corpus were labelled and segmented phonemically (SAM-PA) and prosodically (borders + accents).

Both speakers are trained and experienced broadcast announcers at the local state broadcasting unit. They were asked to read the texts in a speaking style like broadcast announcing, very correct, but fluently and without pausing between words.

The recordings were done in a total echo-cancelling studio at the Institute of Phonetics at the University of Munich. Recording channels were:
- speech signal recorded by Sennheiser MKH20 omnidirectional, 30 cm from mouth.
- laryngograph signal, LxProc of Laryngograph Ltd. London.
- glottis pulse stream by laryngograph
- start/stop pulse at beginning and end of utterance

Recording machine was a high quality 4 channel DAT (48 kHz, 16 bit). The data were copied to hard disk and cut according the pulse information in the forth channel into separate utterances (one utterance per file).

Speech signals were filtered and down-sampled from 48 kHz to 16 kHz. Laryngograph signals were filtered and downsampled to 16 kHz. The format of the signal files is PhonDat 2.

The resulting segmentation and all information accompanying the signal is summed up in the corresponding Partitur File. The Partitur File format is an open structure that allows the easy description and processing of information aligned to a speech signal.

The database also provides an ordered list of all occurring words together with the standard pronunciation in SAM-PA and the orthography of all spoken utterances in the corpus.

	ELRA Members	Non Members
For research use	5,521.95 Euro	6,033.24 Euro
For commercial use	6,774.62 Euro	8,538.57 Euro

ELRA-L0033 LusoLEX European Portuguese Lexicon

LusoLEX is a multifunctional monolingual lexicon of the European variety of Portuguese, developed by the Natural Language Group of INESC. It has about 61,000 entries (lemmas) and 1,600 correspondent inflexion paradigms. The set of entries includes compound words and the inflexion paradigms include information regarding enclitics, augmentatives and diminutives. Morphological information is encoded with maximum granularity and is conformant with the EAGLES recommendations. Most entries were selected from the Palavroso lexicon. The morphological information was converted from the Palavroso rule based morphological analyser and completely revised.

Note: LusoLEX (ELRA-L0033) and BrasilLEX (ELRA-L0034) were designed with the same architecture and format. Each of these lexicons is complete and independent. Therefore, there is a set of entries that are identical in both of them, because they correspond to lexical items used in both national varieties, some of which have different morphological classification. Others, even when used in both National Varieties, are not included in one of them because its frequency of use does not justify it.

	ELRA Members	Non Members
For research use	3,000.00 Euro	5,000.00 Euro
For commercial use	25,00.00 Euro	30,000.00 Euro

ELRA-L0034 BrasiLEX Brazilian Portuguese lexicon

BrasiLEX is a multifunctional monolingual lexicon of the Brazilian variety of Portuguese, developed by the Natural Language Group of INESC. It has about 65,000 entries (lemmas) and 1,600 correspondent inflexion paradigms. The set of entries includes compound words and the inflexion paradigms include information regarding enclitics and augmentative/diminutive degree. Morphological information is encoded with maximum granularity and is conformant with the EAGLES recommendations.

Note: LusoLEX (ELRA-L0033) and BrasilLEX (ELRA-L0034) were designed with the same architecture and format. Each of these lexicons is complete and independent. Therefore, there is a set of entries that are identical in both of them, because they correspond to lexical items used in both national varieties, some of which have different morphological classification. Others, even when used in both National Varieties, are not included in one of them because its frequency of use does not justify it.

	ELRA Members	Non Members
For research use	3,000.00 Euro	5,000.00 Euro
For commercial use	25,00.00 Euro	30,000.00 Euro

ELRA-L0033 and ELRA L0034 can be purchased together at the following prices:

	ELRA Members	Non Members
For research use	3,000.00 Euro	5,000.00 Euro
For commercial use	25,00.00 Euro	30,000.00 Euro

ELRA-W0024/01 PAROLE Portuguese Corpus

The PAROLE Portuguese corpus contains approximately 3 million running words of European Portuguese distributed by Medium, as follows:

Newspaper:
about 65%, covering the period 1996-1997 of 3 titles;

Book:
about 20%, concerning 12 titles from 3 editing houses;

Periodical:
about 5%, concerning 7 weekly issues of 1 title, 1996;

Miscellaneous:
about 10%, concerning several files distributed by 8 titles.

The corpus was classified and encoded according to the common core PAROLE encoding standard. The file format of this corpus is SGML. It includes a subset of 250 thousand words (please see next page, resource ELRA-W0024/02).

	For research use by an academic organisation	For research use by a commercial organisation	For commercial use
ELRA Members	875.00 Euro	1,575.00 Euro	2,450.00 Euro
Non Members	1,250.00 Euro	2,250.00 Euro	3,500.00 Euro

ELRA-W0024/02 PAROLE Portuguese Sub-Corpus

This subcorpus of the PAROLE Portuguese Corpus reproduces approximately the whole corpus distribution by Medium (Newspaper: about 65%, Book: ab. 20%, Periodical: ab. 5%, Miscellaneous: ab. 10%)

It has about 250,000 words morpho-syntactically tagged accordingly to the PAROLE common tagset and morpho-syntactic annotation standards. Disambiguation was manually checked. The files are in SGML format.

	For research use by an academic organisation	For research use by a commercial organisation	For commercial use
ELRA Members	525.00 Euro	875.00 Euro	1,750.00 Euro
Non Members	750.00 Euro	1,250.00 Euro	2,500.00 Euro

ELRA-L0035 PAROLE Portuguese Lexicon

The PAROLE Portuguese Lexicon is constituted by 20 thousand entries morpho-syntactically and syntactically encoded, accordingly to the PAROLE common encoding standards. The data is in SGML format.

	For research use by an academic organisation	For research use by a commercial organisation	For commercial use
ELRA Members	1,400.00 Euro	3,500.00 Euro	10,500.00 Euro
Non Members	2,000.00 Euro	5,000.00 Euro	15,000.00 Euro

Up-date on Language Resources from the ELRA Catalogue

ELRA-S0034 Verbmobil

This resource consists of spontaneous speech recorded in a dialog task (appointment scheduling). The BAS edition of the German part is fully labelled and segmented into phonemic/phonetic SAM-PA by the MAUS system and partly segmented manually. New corpora available via ELRA (for the complete list, please contact ELRA or visit ELRA or BAS Web sites):

VM CD 22.1 - VM22.1 (BAS edition)
Verbmobil II - German, 60 spontaneous dialogues (28 close mic, 5 room mic, 27 phone line (GSM) recordings), 2004 turns, transliteration (VM II Format)

VM CD 23.1 - VM23.1 (BAS edition)
Verbmobil II - American English, 28 spontaneous dialogues (28 close mic, 0 room mic, 0 phone line (GSM) recordings), 2459 turns, transliteration (VM II Format)

VM CD 24.1 - VM24.1 (BAS edition)
Verbmobil II - German, 58 spontaneous dialogues (36 close mic, 0 room mic, 22 phone line (GSM) recordings), 2231 turns, transliteration (VM II Format)

VM CD 25.1 - VM25.1 (BAS edition)
Verbmobil II - Japanese, 10 spontaneous dialogues (10 close mic, 0 room mic, 0 phone line (GSM) recordings), 1654 turns, transliteration (VM II Format)

VM CD 26.1 - VM26.1 (BAS edition)
Verbmobil II - Japanese, 16 spontaneous dialogues (16 close mic, 0 room mic, 0 phone line (GSM) recordings), 1319 turns, transliteration (VM II Format)

VM CD 27.1 - VM27.1 (BAS edition)
Verbmobil II - Japanese, 24 spontaneous dialogues (24 close mic, 0 room mic, 0 phone line (GSM) recordings), 1149 turns, transliteration (VM II Format)

VM CD 28.1 - VM28.1 (BAS edition)
Verbmobil II - American English, 28 spontaneous dialogues (28 close mic, 0 room mic, 0 phone line (GSM) recordings), 2409 turns, transliteration (VM II Format)

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

ELRA-W0015 Le Monde

The text corpus ELRA-W0015 Le Monde for the year 1999 is now available.

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