The ELRA Newsletter

April - June 99

Vol.4 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.

ISSN: 1026-8200

ELRA/ELDA
CEO: Khalid Choukri
Assistant: Wahiba Boukern

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

Dear Members,

As we announced in the previous issue of the ELRA Newsletter, the call for Language Resource (LR) Packaging and Production was very successful and attracted 29 proposals which addressed most of the topics listed in the preference lists that we drew up and posted. A report on the review process, and its outcome, is provided in this issue.

During this quarter, ELRA has been actively involved in the preparation of several proposals that will be submitted to the Fifth Framework Programme (FP5) of the EU. One of these proposals was planned during a preparatory meeting hosted by ELRA on 14 April 1999 and focuses on Multi-Media/Multi-Modal LRs. This type of resource came out of our user requirements survey that was carried out prior to the 1999 ELRA Packaging and Production call, and will allow ELRA to extend its services to more Human Language Technology (HLT) players.

With support from ELRA, the DG13 HLT sector, and other organizations (e.g. ELSNET-ELAN), representatives from National Programmes across Europe gathered in Paris on 16 April 1999 for a kick-off meeting in order to discuss the topic of Language Resources for all European National Programmes and international activities in the field with issues including: complementarity and subsidiarity between national and international activities; synergies between national activities and funding agencies; towards a common general policy in the field of Language Resources; proposal of a set of short and medium term agreed objectives for our efforts: e.g. a minimal set of LR for as many languages as possible, types of LR urgently needed, research topics, the problem of low-density languages, etc.

In this issue of the ELRA Newsletter, we continue our efforts to give you various standpoints on system evaluation processes A paper from Eduard Hovy (ISI, University of Southern California) discusses current efforts to establish a set of certification standards for MT and related translation products.

Nicholas Ostler's (Foundation for Endangered Languages) article elaborates on some of the ways that new media in language technology can directly or indirectly favor low-density and sparse-data languages, thus highlighting ELRA's concern about minority languages for which market forces alone cannot guarantee HLT transfer.

This is followed by an article by Sharon O'Brien (ALPNET) on Translation Memory (TM) that is the first of several articles that are expected to appear in ELRA Newsletter issues on the topic of TMs. O'Brien focuses on the implementation and standardization of translation memory technologies from the perspective of a software localization provider.

We are glad to announce the availability of the PAROLE and EuroWordNet LRs which constitute re-usable and inter-operable LRs that you have been waiting for. The PAROLE Dutch corpus and lexicon are the first to be distributed. We expect a large number of PAROLE partners to sign similar license agreements during the next quarter. We hope that the PAROLE resources will bridge the gap that is regularly reported by developers of HLT in the area of written corpora and lexica development, similarly to what the SpeechDat family of resources has provided to the speech community.

We are sorry to inform you of the departure of our assistant Rébecca Jaffrain, who left our team at the end of May. We are very grateful to her significant contribution to ELRA and ELDA over the past three years and would like to wish her our best in her new ventures.

We are pleased to announce that with ELDA's new activities, it is offering a Research/Technical Engineer and a bilingual assistant position. Feel free to contact Jeff Allen (jeff@elda.fr) for more specific information about these positions.

ELRA is proud to welcome several new members since the beginning of 1999: The following is a list of those who have joined ELRA as members since the beginning of 1999: Speechworks International, Inc., USA; Institut d'Estudis Catalans, Spain; National Technical University of Athens, Greece; Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA), France; IBM, Spain ; IBM, Italy; Elan Informatique, France; Institut de Recherche en Informatique de Toulouse (IRIT), France ; DIBE University of Genova, Italy; Laboratoire Lorrain de Recherche en Informatique et ses Applications (LORIA), France; Aculab PLC, United Kingdom; Università Degli Studi di Pisa, Italy; Synapse Développement, France.

Antonio Zampolli, President Khalid Choukri, CEO

Does Size Matter? Language Technology and the Smaller Language

Nicholas Ostler, Foundation for Endangered Languages

The census figures for the use of Welsh in Wales, as between 1961 and 1991, paint a puzzling picture. On the one hand, as expected, the level of Welsh-competence in the country, away from the bigger towns and cities, has declined; but in the meantime, Welsh competence in the cities and suburbs, which had been declining throughout recorded history, has started to go up[1]. Why should this be?

These three decades from 1961 to 1991 have witnessed the spread of mass media and information technology into every home and school. This has meant that English-language entertainment can now penetrate every Welsh person's world, every day. But the same period has also seen the set-up of a Welsh-language TV channel, and growing public insistence on the use of Welsh in schools and government offices. The Welsh Language Board, in its Strategy published in 1996, explicitly expects support of Welsh through information technology. Somehow, Wales is showing that new media can provide some benefits for traditional smaller languages, as well as competition to them.

This article examines some of the ways that new media in language technology can favour smaller languages, directly or indirectly. The benefits are real, but they do not come without costs.

What Can Language Technology Offer, at the State of the Art?

There is a wide profusion of applications possible for that confluence of computing and linguistics that we call Human Language Technology. So much so, that I have found it useful to organize them into a table, with the various aims of the applications on the horizontal axis, and the various technologies that can be deployed down the vertical axis.

The column listed under "Develop Foundations" is not in itself a list of applications, but rather of the kinds of studies, most carried out at research institutions, which may support progress in the other applications further down the row.

When the various applications are displayed like this , one immediately sees that applications which require high-level analysis of grammar and meaning are in a small minority, perhaps only Interpreting (not yet available) and Machine Translation; while Summarization, Grammar-checkers, Text Retrieval and Computer-Aided Language Learning might be expected to make much more use of it in the future. This only underlines the fact that smaller languages can begin to apply the technology even though very little work has been done as yet on formal analysis of their structures.

	Develop Foundations	*Production and Publishing*	*Improve Access for Insiders*	*Improve Access for Outsiders*
Speech Processing	Speech databases; recognition; generation	Dictation, vocalization	Voice control, alarms	[Interpreting]
Text Processing	Coding standards; localization	Word processing	Text retrieval, summarization	Multilingual document search
Compiling Reference Materials	Morph analysers; parsers; corpora	Spell-checkers; gram-checkers	Multimedia, document libraries	Machine(-aided) translation
Networking	Interchange standards; protocols	World Wide Web	E-mail, discussion lists	Electronic networks, WWW
Computer-aided Instruction	Dictionaries (computer tractable)	Literacy	Classroom materials	Computer-aided language-learning

Who Can Use What Language Technology Can Offer?

The European Commission has claimed that languages which do not take a full part in the electronic media are doomed to stagnate, if not atrophy:

> … many of the minority languages are experiencing difficulties, often under the influence of changing patterns of communication. Penetration of the new technologies could substantially accelerate this process, threatening to diminish the linguistic and cultural diversity of European society.
>
> … The rapid rise in use of information and communication technologies will naturally favour languages which can be successfully processed. Languages supported by key software products offering powerful facilities for manipulating text also provide almost unlimited access to information services in those languages… The long-term viability of languages not specifically supported is therefore put at risk.
>
> *EC proposal for a Council Decision, Multilingual Information Society, 1995*

Language, any language, serves many purposes, from acting as a kind of community flag, through business interactions and literary creations, to private conversations and inward musings. Since electronic media are novel, whereas all these purposes of language go back for æons, the question arises how these "changing patterns of communication" actually fit into the overall web of language use.

There is no model to hand of how all the purposes fit together in a community's life, let alone their relative importance in it. It is intuitively clear, however, that the immediate user community foreseen for language technology applications is the business community, especially larger concerns who are able to invest to bring down costs in the longer term. It is they who would have an immediate use for the various applications. A second major community would be those engaged in research and education (except perhaps for the applications on the top line, requiring speech processing). Without radical changes in urban living, however, there seems little reason to expect massive growth of the use of language technology in everyday home, social, leisure and indeed religious life, which are inevitably the centres of language use for smaller languages.

If the EC are insistent, therefore, in their claim that abstinence from language technology may endanger a language's long-term future, they seem to be saying that a language which cannot function in modern business, research and education is likely to drop out of use in domestic life as well.

In fact, all over the world, language technology is helping smaller communities either to bridge the gap from domestic to other uses of their language or to overcome particular difficulties in their domestic lives. This is happening in different ways for different purposes, and there is only space for a very few examples here.

A simple use of **word processing** and desktop publishing technology is to jump-start literacy in a community by producing books and other literature cheaply but attractively with small print-runs. Two examples of this happening are in Oaxaca, Mexico, through the efforts of ***CELIAC***, the Centro Editorial de Literatura Indígena (asociación civil)[2], and in Maesai, Chiangrai in northern Thailand, through the ***Akha Heritage Foundation***[3].

In the former of these, a teaching centre has been set up to show how literacy in one lan-

[1] J. Aitchison and H. Carter, A Geography of the Welsh Language 1961-1991, University of Wales Press, 1994.

guage (usually Spanish) can provide the basic know-how to start writing in one's own vernacular, even when it has never been written before. In this way, native speakers have produced publishable texts in the Mexican languages Totonac, Zapotec, Mixtec, Chinantec and Otomí-Ñahñu, and instruction has also been given to speakers of Quichua and Aymara in South America.

In both these cases, the attempt is being made to finesse the contentious step of agreeing an orthography for the language of publication, simply going ahead without an agreed standard, and trusting that (as happened in the case of Middle English a few hundred years ago) a workable standard or set of standards will emerge.

Not all accept this *laissez-faire* approach, especially in the case of the Akha language where two competing well-defined standards already exist. But there are considerable advantages to it. On the one hand, it encourages direct involvement of speakers in the literature of their own language on their own terms, so that the element of "linguist's patronage" is reduced. From a strictly lexicographical point of view too, it can be highly productive: rather than starting with the laborious work of compiling a dictionary through elicitation by a linguist from a native speaker, and checking conformity of words used in texts with the inevitably incomplete lexicon that emerges, the texts are primary. The task of compiling a dictionary, when it is approached, is then lightened by having a substantial corpus of spontaneous literature to work from, and to discuss between speakers and linguists if necessary.

The networking potential of the **World Wide Web, E-mail and discussion lists** can be applied to the benefit of small-language communities which are in diaspora, with local groups small and scattered. A good example of what can be done can be found at ***Nineveh on Line***, <http://www.nineveh.com> which provides an on-line home for the Assyrians, mostly those now resident in the USA, but also in Australia, Sweden, Lebanon, Iraq, and Canada. It provides an on-line newspaper, and a chat line in English. But it is also a ready source of information about, and support for learning, the Assyrian language in its distinctive written form. More specifically linguistic is the coverage of Kurdish on the ***Kurd_lal Archive***, <http://www.cogsci.ed.ac.uk/~siamakr/Kurdish> providing almost an encyclopaedia of language resources on line, with pages on standardization of Kurdish, its linguistics, a corpus of literary and newspaper texts, fonts, discussion of the right to use Kurdish, access to language engineering applications such as an online dictionary, a bibliography and a chronology of recent events.

Computer aided language learning is most simply applied at the moment to distribute language learning materials, not all of them involving computing, or even machinery at all. This is another aspect of the World Wide Web's value to smaller languages, and the SSILA Learning Aids site (Society for the Study of the Indigenous Languages of the Americas <http://trc2.ucdavis.edu/ssila/learning/stm>) gives access to such materials for over eighty distinct languages, all from North America. Particular sites may offer useful background of a psycho- or sociolinguistic nature (e.g. the Cheyenne site <http://www.mcn.net/~wleman/cheyenne.htm> gives useful references on the educational value of bilingualism). And of course it is not necessary that a language still be alive to be the focus of this kind of coverage: the Jiwarli website <http://adhocalypse.arts.unimelb.edu.au/Dept/LALX/research/jiwarli> is now the only place where this western Australian language can now be heard, along with a linguistic analysis of some ethnographic tales from the mouth of its last speaker.

At the other extreme, with adequate technical back-up, such network technology can provide proxy access to broadcasting in the language, and so gives a close equivalent of full current participation in the discourse of the language community. Two examples of this can be found in the Raidió Teilifís Éireann site <http://www.rte.ie/av.html> where RealAudio™ news magazines and news reports in Irish can be found, and the BBC's own site <http://www.bbc.co.uk/cymru/live/news.ram>, where a RealPlayer™ shows the TV news in Welsh.

It is arguable, then, that the most valuable technologies to smaller languages at the moment are not "Language Technologies" as such at all, but broadcast and networking technologies which are largely neutral as to language, but need some language in which to transmit. They can project the informational and cultural outpourings of smaller language groups across the world in an economical and targeted way.

This is not to forget the presumed multiplier effect of the business, educational and government uses of what is more specifically language technology. Making money and supporting a way of life while still using a favoured language, such as Basque or Welsh, provide a healthy environment for the language community to grow and flourish; and in the modern world the language technologies will be required in any business or research establishment. It is precisely this effect which is responsible for the strange new dispersion of Welsh into the urban heartland of Wales which was noted at the outset.

Fundamental Work for Smaller Languages

So much then for the use which is currently being made of language technology in smaller language communities. A completely different approach is to ask about the work on the foundations of language technology that is being carried out focused on these languages, which may ultimately result in more sophisticated language technology in applications.

The most fundamental work of all, at least for text processing, is to ensure that there is a **coding standard** which can represent either the traditional writing system for the language, or (if there is not a written tradition) an adequate orthography which makes all the necessary distinctions. In the latter case, the decision has often been made to opt for a system of graphs which does not go beyond the ASCII characters, basically the unaccented letters of the Roman alphabet, lower and upper case. This is more or less good enough for English, and as it happens also for Dutch, Cornish, Basque and Latin, although none of these except the last is a perfect fit. Otherwise, though, such a restricted set is adequate only for the more recently written languages in sub-Saharan Africa, the Pacific islands, and Australia. Almost everywhere else either some extension to the ASCII alphabet is required, or a completely new alphabet[4]. And in every such case, a coding standard must be defined, agreed and propagated (something else for which web sites can be useful, as witness the Hawai'ian site <http://www.olelo.hawaii.edu>). Such codings almost inevitably require use of the eighth bit in the byte, and hence are not reliably transmitted directly by electronic transfer.

The advent of a 32-bit Unicode standard may ultimately obviate the need to establish these local standards (though woe betide the alphabets which get left out). But that eighth-bit problem may hang around. And in general it does seem that standardization problems are particularly hard to solve in small communities: as witness the three standards for spelling Cornish, which persist among no more than 500 speakers.

On the speech side, the fundamental task is to accumulate **databases of speech examples**,

[2] CELIAC, Avenida Ejercito Mexicano 1107, Colonia Ampliacion Dolores, Oaxaca, Oaxaca 68020 Mexico, or by phone at +52-951-59725 fax -59729; e-mail celiac@infosel.net.mx Information in English: Russ Bernard at: voice +1-904-376-4544; fax +1-904-376-8617; e-mail ufruss@nersp.nerdc.ufl.edu.

[3] The Akha Heritage Foundation, 386/3 Sailom Joi Rd, Maesai, Chiangrai, 57130 Thailand <http://www.akha.com; e-mail akha@loxinfo.co.th>

[4] One such development, which required modification of an existing, but totally non-Roman script, the Cree syllabary, is described in the context of Naskapi history by Bill Jancewicz: see Endangered Languages - What Role for the Specialist? (Proc. Second FEL Conference, 1998) available from the Foundation for Endangered Languages.

since large numbers of examples have turned out to be essential for training practical speech processing systems. Within Europe, this is increasingly being undertaken for the principal non-national languages: in Spain, for example, the telephone utility Telefónica I&D have recently followed up their 1993 VESTEL speech database of 15,000 Spanish speakers with VOCATEL and VOGATEL (approximately 7,000 speakers each in Catalan and Galician respectively), and are looking to build a new database for Basque.

This is a national project, but minority languages are also being included in European Union initiatives. The second instalment of ELRA's SPEECHDAT project, besides including Slovenian, Norwegian and some dialects of Swiss French and German along with the official EU languages, has brought Welsh into its collection of speech databases, with 2,000 speakers.

Although such speech databases are widely agreed to be fundamental for pursuing speech processing, the exact type of processing will indicate a given scale and structure: even for Welsh, there is room for a second speech corpus (being compiled at the University of Edinburgh) where only six speakers read magazine articles. This corpus will be much more deeply annotated than SPEECHDAT, and is above all focused on phonetic analysis of the Welsh language, rather than to provide a basis for speech recognition.

Text corpora are also being built up, where possible in Europe within the same structures and parameters as laid down for the official languages.

Thus the 1996-98 LE-PAROLE project included a 20 million-word Irish corpus along with those for the national languages of Europe. The genesis of this corpus was instructive and throws a little light on the predicament of smaller languages: the general requirement was for to newspaper text to constitute between 58% and 72% of a language's corpus, but there is no daily newspaper in Irish (as would be the usual situation for a minority language), and one of the weekly newspapers closed down quite soon after the project began. As a result, newspaper text in this corpus is closer to 45%. On the other hand, the language market is small enough for a huge percentage of Irish-language typesetting to be done by a single company; since that company (and its authors) were willing to co-operate, many of the problems of collection and standardization could be solved at a stroke. The resulting corpus will be used as a resource material for an Irish spell-checker, the first to be compatible with standard PC word-processing programmes[5].

Lexicon compilation and **morphological analysis** are also going on with co-operation from comparable work done in the past for larger and better funded languages.

An example of this, though not for very small languages, is the GRAMLEX project, funded by EU's COPERNICUS programme in 1996-98. Here the techniques of morphological analysis and lexicographic coding developed by the French ASSTRIL for their own language and Italian have been explored for adequacy to the very different, and much more highly inflected, vocabularies of Polish and Hungarian. Furthermore, Polish has an extremely complex system of morphophonology, and the agglutinative structure of Hungarian means that the set of well-formed forms is essentially unbounded. The results of the project included useful corpora for Polish and Hungarian, practical inflexional coding schemes for these languages, and some concrete analysis of the weaknesses in Koskenniemi's Two-level Morphology, as well as an observation that certain lexicon maintenance software (INTEX) had a utility who transcended particular languages[6].

There are also some attempts for smaller language technologists to get together, and attempt to learn mutually rather than from other bigger language work. An example of this is the MELIN project, involving Irish, Welsh, Catalan and Basque terminologists and lexicographers in pooling results and (presumably) comparing strengths and weaknesses <http://www.ite.ie>. Such work, though, is still very much in its beginnings.

In the USA, there are a couple of interesting initiatives to generalize the value of language processing undertaken for particular, well-subscribed, languages. The BOAS project, or "Linguist in a Box", looks to automate the process of transition from a native speaker's knowledge of a language into a machine translation system which will convert utterances into English[7]. And a company which provides multimedia language learning environments has offered to package its expertise so that such aids will be available for smaller languages too <http://www.transparent.com/endangered/index.htm>.

Some Overall Points

Although in many ways, then, smaller languages are well positioned to take part in this second wave of language technology development, it must be remembered that there are aspects of language technology which are particularly adverse for traditional societies. And in the main, smaller languages do tend to be spoken in more traditional communities.

One of these is the fact that electronic storage and recall is in many ways a substitute for traditional reliance on memory. Time-honoured resources of song and storytelling may be lost in an environment where technical aids loom larger.

Furthermore, communities that come together electronically, through broadcasts or over the Net, will not have the motivation to privilege particular times as festivals which are conducive to particularly rich displays of language: ceremonies and eisteddfodau may find it harder to survive.

And above all, the translation of some linguistic activities to a novel, electronic, medium will tend to give younger people a new and more important role in knowledge transmission: this may, at least at first, be quite disruptive of traditional patterns of communication among the generations, shaking the inter-generational transmission on which languages rely in order to survive.

Despite all these especial difficulties, smaller languages will be coming to terms with language technologies all over the world in the next decade. They will find that there are benefits to be enjoyed, even as there are insidious side-effects to be avoid. There may even be aspects of the technologies which are more useful to them than to speakers of larger and more unitary languages: keeping in touch despite a global diaspora seems a first example of such a hopeful aspect.

And we can expect too that some technical developments, undertaken somewhere to support some feature of the sheer diversity of the host of smaller languages, will themselves bring benefits to a wider world. Smaller languages present a challenge. We still await the response.

Nicholas Ostler
Foundation for Endangered Languages
172 Bailbrook Lane
Bath BA1 7AA -- ENGLAND
Tel: +44-1225-852865
Fax: +44-1225-859258
Email: nostler@chibcha.demon.co.uk
http://www.bris.ac.uk/Depts/Philosophy/CTLL/FEL/

[5] A wealth of information about these and other speech and text corpora for minority languages can be found in the proceedings of the Workshop for Language Resources for European Minority Languages, Granada 1998, reviewed by me at <http://www.cstr.ed.ac.uk/~briony/SALTMIL/review.html>, and obtainable from the editor, Briony Williams, at HCRC, Edinburgh University <briony@cstr.ed.ac.uk>.

[6] On GRAMLEX, for Hungarian, contact Gábor Prószéky<info@morphologic.hu> and see http://www.morphologic.hu ; and for Polish, Zygmunt Vetulani at Uniwersytet im. A. Mickiewicza, Poznan <vetulani@math.amu.edu.pl>.

[7] Sergei Nirenburg: Project Boas: 'a Linguist in a Box' as a Multi-Purpose Language Resource' in Proc. First International Conference on Language Resources and Evaluation. Granada, 1998. vol. II, pp. 739-746. ELRA

Results of ELRA 199
ELRA Commissioning Produ

1. Introduction and Purpose of the Call

The European Language Resources Association (ELRA) has completed the selection of proposals for the first of a series of calls for the (co-)production and packaging of Language Resources (LRs). ELRA, as a non-profit organization, has chosen to devote some of its funds, both from the Language Resources Production and Packaging project as well as from language resources sales, in order to commission the production, packaging and/or customization of LRs needed by the Language Engineering Community, and so invited applications for production and/or packaging/repackaging projects which could be eligible for funding from ELRA. The purpose of the call has been to ensure that necessary resources are developed in an acceptable framework (in terms of time and legal conditions) by the LE players. This call was targeted towards projects with short time scales (projects lasting up to one year but preferably shorter).

2. Preference Lists for Proposals

From recent market monitoring, ELRA had identified several key speech and written resources. ELRA has categorized and prioritized this set of resources as indicated below:

SPEECH LANGUAGE RESOURCES (SLRs) - Preference list

1. SpeechDat like database
2. Speech database for embedded systems
3. Pronunciation lexica
4. Dialog corpus
5. Enrichment of existing SLRs within the ELRA catalogue
6. Multilingual speech synthesis database

WRITTEN LANGUAGE RESOURCES (WLRs) - Preference list

1. Large monolingual corpora
2. Parallel texts
3. Bi/multilingual computational lexica

MULTIMEDIA AND MULTIMODAL LANGUAGE RESOURCES - Preference list

1. Multimedia corpus
2. Multimodal corpus

3. Proposals received and evaluation process

Following the diffusion of the call for proposals on 8 February 1999, ELRA closed the call on 19 March 1999. All of the proposals were initially judged as acceptable to be evaluated by external experts and the ELRA review committee and were related to the preference lists indicated in the call. We received 29 proposals which cover the following areas (in some cases, there is overlap in 2 areas): 10 Speech; 1 Multi-Modal/Multi-Media; 11 Written Monolingual Corpora; 7 Written Multilingual Parallel Corpora; 3 Written lexica.

Each proposal was evaluated by a minimum of 3 external evaluators, chosen according to the areas in which the proposals were grouped.

The items listed below are among those considered as selection criteria for the evaluation of proposals. These categories were each clarified in detail by a number of questions (between 2 and 13 individual questions per category) that provided the reviewers with a standardized format of evaluating the main themes for each proposal.

1. Conformity with the scope and objectives of the call

As indicated, it is important to consider if each proposal falls within the scope and the specific objectives of the call and ELRA's mission behind it.

2. Industrial relevance

This category considers whether proposed projects are clearly related to existing or anticipated industry demand.

3. Objectives and results

Questions in this category aim at determining if proposed projects contribute to reasonable results of language engineering for the language(s) they address and an enrichment of the corpus material for this/those language(s).

4. The Proposers

Consideration is taken with regard to the roles, skills and experience of the team and whether or not they are sufficiently balanced for achieving the proposed objectives.

5. Work planning

It is important that projects provide a clear presentation of major activities (i.e., milestones and deliverables) with an associated calendar.

6. Budget

The proposed budget for each submitted project is evaluated with respect to the tasks to be accomplished and had to be realistic in view of the expected results.

7. Legal aspects

It is important that proposals clearly state that ELRA will be granted the distribution rights with regard to the results of work in the framework of this project. Also, it is important to note if the provider asks for royalties for the work that is to be accomplished.

99 Call for Proposals
uction of Language Resources

After the preliminary evaluation process of the 29 proposals with the assistance of 16 external experts, an initial list of 5 monolingual corpora, 5 bi-/multi-lingual corpora/lexica, 5 speech databases, and 1 multi-modal database was drawn up and presented for final review. All proposals were finally screened by a review committee that consisted of the ELRA Board members and a European Commission (DGXIII - Human Language Technologies sector) representative.

4. Short-list of proposals

On 3 May 1999, the ELRA Review committee selected a final short-list of candidate proposals. The Review committee issued several recommendations (e.g., merger of multiple proposals from the same team, ELRA only interested in a single module of multi-module projects, ELRA willing to co-finance a full project if complementary funding is obtained, budget reduction for some projects, etc). This short-list is therefore tentative and depends entirely on current negotiations with candidates in order to determine which projects will finally be funded. A summary of short-listed proposals is provided below.

A. Monolingual written (Catalan), monolingual written (Spanish) and parallel written (Catalan-Spanish) LSP text in two domains (law and economy)

The language resource (LR) to be built in this proposal includes 2 monolingual corpora and 1 bilingual corpus:

- Catalan monolingual corpus: 2 Million words (approximately 1 Million words for each domain)
- Spanish monolingual corpus: 2 Million words (approximately 1 Million words for each domain)
- Catalan and Spanish bilingual corpus: 1.200.000 words (300.000 words for each language and each domain)

All corpora in this project are ASCII text with SGML mark-up and Part-of-Speech tagging.

B. Scientific Corpus of Modern French

This project aims at building a corpus of contemporary written scientific French ; it will be a new resource. This would be a monolingual, mono-source corpus developed from the journal “La Recherche”, so as to obtain a multidisciplinary overview of scientific usage. The finished product would consist of some 450 articles from 1997 to 1998 covering 30 large themes for a total approximately 1.5 million words.

C. New corpus of written Business English

This project intends to create a new corpus of written Business English. This will form part of ongoing plans to create comparable and parallel text corpora in several domains and several languages. It will consist of an ASCII text corpus of 10 million words, with SGML markup, part-of-speech tags, and sentence and paragraph boundary markers.

D. German-French Parallel Corpus of 30 Million words

This German-French Parallel corpus is a 60 million word corpus (30 million for each language) for the purpose of developing, enhancing and improving translation aids (dictionaries, lexicons, platforms) for French-German and German-French translation.

E. Sets of bilingual LR dictionaries for English and Russian

The dictionary to package is an English-Russian LR-dictionary through reformatting of an existing source dictionary.

Automatic inversion of the preceding and manual editing will also be carried out to obtain a Russian-English LR-dictionary.

F. Crater 2 - Expanding Resources for Terminology Extraction.

The CRATER project went beyond the work of ET10/63, hand correcting the part of speech tagging in the corpora produced, and adding a third language, Spanish, to the corpus. The goal was to produce a 1,000,000 token corpus of Spanish, French and English, and to align these corpora with one another at the sentence level. CRATER produced these corpora on time and within budget. However, at the end of the project corpus data was still available which could not be incorporated within the delivered products in time. It is this corpus data which is the focus of this proposal. With additional modest funding, CRATER can expand significantly and useful data which is not currently in the public domain can be placed there.

G. An Italian Broadcast News Corpus

This project aims at collecting a multimedia corpus of radio broadcast news in Italian. The corpus will include audio signal, transcriptions, and documentation for the users. Broadcast news will be acquired from the digital archive of the Italian major broadcaster Radio RAI. Hence, the project aims at producing a new language resource (LR) starting from digital audio recordings.

H. Pronunciation lexicon of British English place-names, surnames and first names

The size of the projected database is currently estimated at circa 100-200 thousand main entries. Each word will be encoded with relevant information such as : thematic status (place-name, surname, first name), phonetic transcriptions ("main" and "secondary" phonetic variants), number of letters, number of syllables.

5. CONCLUSION

ELRA is currently negotiating contract conditions with short-listed candidates. A final list of the proposals to be funded will be provided in the next issue of the ELRA Newsletter.

Translation Memory as a linguistic resource in the Localisation Industry
A snapshot of the present and glance into the future

Sharon O'Brien, ALPNET Ireland

Translation Memory as a pre-requisite in localisation

The software localisation industry is characterised by frequent updates to source files, demands for short translation turn-around time and a constant downward pressure on price. For these reasons, the translation memory (TM) has become a valuable linguistic resource within localisation because it is perceived to help meet the demands of the industry.

The potential benefits of translation memory technology were quickly embraced by software localisation providers. Some were developing their own TM tools from the early 1980s. But, it is only since the mid-1990s that the use of translation memories has become a pre-requisite in localisation.

The initial drive for the use of TMs actually came from the supply side rather than the demand side. Localisation suppliers recognised that, just like the word processor before it, the translation memory tool would become a standard tool in the localisation process. The tools were evaluated and implemented and the concept was then sold to the demand side on the basis of increased quality, faster throughput and reduced costs.

Implementation Costs

For a localisation supplier, implementing translation memory technology was not, and still is not, a trivial task. First, a decision is taken on which tool will be used as a standard. This requires a good deal of market research, evaluation and testing to find the tool that suits a company's requirements. However, it is not only the localisation supplier's requirements which must be met, but also those of the company's wide-ranging customer base. Identifying one TM tool which fits all possible requirements has proved to be an impossible task for most, if not all, localisation providers. Consequently, localisation providers must have expertise and experience in many different TM tools, although each provider will generally favour one tool over another.

It could be argued that the requirement to provide services using more than one TM tool is not unrealistic. After all, a localisation provider must have expertise in different software development packages, programming languages, word processing tools, publishing technologies, platforms and operating systems. TM tools have simply been added to the list of software packages that the localisation provider has to have expertise and experience in.

However, the implications of this requirement are substantial. For each translation memory tool a company invests in, the following costs must be met:

- Licence Fees
- Training Fees
- Support Fees
- Internal Support Infrastructure
- Additional R&D costs

Experience with TM technology has shown to date that no one tool provides the solution to every problem. All software localisation providers have developed utilities around standard tools in order to integrate them with their own internal tools and to make the process of using them easier. The more tools you have to use, the more time and money you have to invest in these solutions.

Standardisation of TM tools

One could suggest that standardisation of TM tools would alleviate some of the costs alluded to above. If "standardisation" means standardising on one tool, then one is faced with the problem that competition between TM tool developers would be dissolved. This could easily lead to an inferior product due to lack of competition. Also, much of the effort spent by localisation providers on R&D to improve the performance of different commercial tools would be lost.

If, on the other hand, "standardisation" means developing a standard interchange format for the exchange of Translation Memories between different tools, then this would lead to a reduction in the problems mentioned.

Fortunately, the development and implementation of such a standard is well underway. This standard is called "TMX", which stands for "Translation Memory eXchange". It is being developed within the "OSCAR Special Interest Group" of the LISA organisation. For more information, readers should refer to http://www.lisa.org.

Current usage

Although the TMX standard will go a long way to providing a solution to the problem of multiple translation memory tools and translation memory exchange, it will not make many of the other issues associated with the use of TMs disappear.

The localisation industry is now at a stage where some benefits are being reaped from Translation Memory. Overall, however, there is a general air of disappointment. The biggest failure of Translation Memory has been its inability to deliver on the expected cost and time reductions. On the other hand, most would agree that quality of translation has most definitely been aided through the use of translation memory tools.

There are many reasons why TM tools have failed to deliver on the expected benefits, only some of which are mentioned below.

A Translation Memory is a shared linguistic resource. The most benefit can be extracted if a group of translators share a translation memory over a network, thereby making use of each other's work in (almost) real-time. While many localisation providers have groups of translators in-house, freelancers are also used to a large extent. This means that TMs are not being shared over a network. It also introduces an additional task involving the management of different versions of the same TM. Different versions of a TM are "merged" at the end of a project. Although there are ways of controlling how the data is merged (through the use of "meta data" fields in the TM), the merging of translated data is somewhat haphazard and does not guarantee that the best translation of a sentence will not be overwritten by another, lower quality version.

Few companies (and by this I mean localisation providers and their customers) have a structured approach to the management of translation memories. TMs have been built up in a haphazard manner (usually project by project) without much thought for the possibilities of re-use across product lines even within the same company. This matter is further complicated by the fact that translation of software files is executed using different tools because the standard translation memory tools provide more support for documentation and help than for the translation of software files.

Frequently, there is little or no thought put into the labelling and management of data

within the TM. Maintenance of TMs is not something customers are willing to pay for, so it is simply not done in most cases.

While the points mentioned above have contributed significantly to the disappointment over translation memory, the highest contributory factor has to be the lack of integration with authoring tools and processes. For such a long time, translation has been seen in isolation from source content creation. Groups who manage translation for large corporations almost always work independently of the authoring or publishing group. So, when you try to explain that a change in the formatting of a sentence, or an unnecessary addition or deletion during the revision of a document can lead to a loss of an exact match in a translation memory and a subsequent increase in the cost of translation, you are frequently met with a complacent or even defensive attitude.

A related point is that most TM tools segment text on the sentence level. Authoring groups are, on the other hand, moving more towards SGML/XML and information management systems which deal with "chunks" of information (elements or entities) on a paragraph, topic, chapter or even book level. Frequently, the initiative to move in this direction focuses on the source files only. Management of translated data is left up to the translation group or localisation provider, thereby leading to quite different methods of storing and reusing information.

A Bleak Outlook for Translation Memory?

In the localisation industry we have completed the honeymoon period with translation memory and are settling into an uneasy period of realisation and admission that there are faults associated with this technology.

The future for translation memory is, however, not a bleak one. It has been established as a pre-requisite for all localisation providers. A standard for the exchange of different translation memory formats is well underway and the industry is waiting in anticipation for it.

The faults of TM technology have been identified. It is now possible to look into the future and predict what developments will come next.

New generations of TM tools will be Client/Server based which means that translators will be able to access large TMs over a network or over the Internet and the problems associated with the freelance nature of the business ought to be significantly diminished.

Translation Memories will have more powerful database technology at the back-end, enabling better control of data, easier maintenance and management.

Translation technology will be closely aligned with authoring technology. We will see integration with XML-based document management systems, where not only source information will be stored and maintained, but parallel chunks of multilingual information will be available and leveraged from there.

Segmentation algorithms within TM tools will be more flexible and customisable so as to be in tune with the segmentation in a document management system.

And, as is already happening, this technology will have add-ons such as controlled language checking utilities, access to powerful terminology databases and to multiple machine translation engines. The end-user will be able to choose what information to publish in which language and which medium to use.

The good news is that some of this technology is not too far away and the translation industry is beginning to buzz in anticipation of a new era which will go far beyond the translation memory system as we know it now.

Sharon O'Brien
ALPNET Ireland,
Ballymount House, Parkway Business Centre, Ballymount Cross,
Dublin 22 -- Ireland
Office Tel.: +353 1 456 97 60
Mobile: 087 239 24 28
Email: sharonob@alpnet.com

Machine Translation Certification

Eduard Hovy, USC Information Sciences Institute

Originally appeared in MT News International, #21 (Feb 1999) pp. 1-2. (reprinted with permission granted by the MTNI editor).

Ever since the IAMT regional associations were founded, there has been keen interest in giving some form of certification to commercial MT and related products that would serve as an educational and informative guide to potential users and others who follow this field. The panel at AMTA-98, "The AMTA Seal of Approval," chaired by AMTA President Eduard Hovy, not only sparked a lively debate at the conference but also fueled the long-needed impetus for action.

Starting with the panelists as a core, an ad hoc international committee was formed early this year under Hovy's leadership to propose a set of certification standards. The ad hoc committee will report to the IAMT Council, which is expected to eventually promulgate formal standards via an IAMT Committee on Certification Standards. The work being done is not for the benefit of any company or organization. It is not being paid. The certification initiative is an action being taken for the benefit and education of all who are concerned with machine translation. It is hoped to have a first version of the certification program complete by the beginning of 2000.

In the meantime, a number of draft proposals have been exchanged and discussed, and a rationalized consensus is emerging. The ad hoc committee is therefore considering issuing a formal statement that prefigures the certification. Basically, it is proposing that MT and related products be grouped into categories, and that each category have a "necessary and sufficient" set of criteria that describe it adequately and at the same time differentiate it from the other categories. Ultimately, a system will be certified in a given category because it meets the stated criteria.

There is agreement already that there should be two broad headings, (1) machine translation systems, and (2) translation support tools, and that each of these will be subdivided. It is proposed to categorize the MT systems by levels: basic, standard, and advanced. It is planned to break down the translation support tools into electronic dictionaries, terminology management systems, translation memory systems, etc. Each of the categories will be defined by a distinctive set of criteria, which are in the process of being developed.

This initial formal statement will be useful to help orient users and MT developers in the rapidly expanding world of machine translation systems and will allow the ad hoc committee to collect more accurate feedback.

Eduard Hovy
USC Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292-6695 -- USA
Tel: 310-822-1511 ext 731
Fax: 310-823-6714
Email: hovy@isi.edu
Project homepage: http://www.isi.edu/natural-language/nlp-at-isi.html

New Resources

ELRA-S0063 German SpeechDat(II) FDB-4000

The German SpeechDat(II) FDB-4000 consists of 4000 calls (1938 males, 2060 females, and 2 unknown-gender speakers) over the German fixed network, stored on 17 CD-ROMs in the final SpeechDat(II) database exchange 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.

The following items were recorded: 1 isolated digit; 1 sequence of 10 isolated digits; prompt sheet number >= 5; 9-11 digit telephone number (read); 15-16 digit credit card number (read, 150 different credit card numbers were found); 6 digit PIN code (read); 1 natural number (read); 1 money amount (read); 2 yes/no questions (spontaneous, not prompted); 3 dates (1 spontaneous, e.g. birthday; 1 prompted text form; 1 relative and general date form); 1 time of day (spontaneous); 1 time phrase (read); 3 application words; 1 word spotting phrase; 5 directory assistance names (1 spontaneous name (e.g. forename), 1 spontaneous city name, 1 read city name (from a list of 500 most frequent), 1 read company/agency name (from a list of 500 most frequent), 1 read proper name, fore- and surname (from a list of 150 names); 3 spellings (1 spontaneous, e.g. forename; 1 directory city name; 1 real/artificial word); 4 isolated words; 9 phonetically rich sentences (read).

The following age distribution has been obtained: 204 speakers are below 16 years old, 1685 speakers are between 16 and 30, 1166 speakers are between 31 and 45, 729 speakers are between 46 and 60, and 216 speakers are over 60.

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

Price for ELRA members:	for research use: € 35,000	for commercial use: € 45,000
Price for non members:	for research use: € 45,000	for commercial use: € 55,000

ELRA-S0069 Swedish SpeechDat(II) FDB-5000

The Swedish SpeechDat(II) FDB-5000 comprises 5000 Swedish speakers (2470 males, 2530 females) recorded over the Swedish fixed telephone network. The SpeechDat database has been collected and annotated by the Department of Speech, Music and Hearing, KTH. This database is partitioned into 25 CDs, each of which comprises 200 speakers sessions. The speech databases made within the SpeechDat(II) project were validated by SPEX, the Netherlands, to assess their compliance with the SpeechDat format and content specifications.

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

Each speaker uttered the following items: 3 application words; 1 sequence of 10 isolated digits; 4 numbers : 1 sheet number (5-10 digits), 1 telephone number (9-11 digits), 1 credit card number (16 digits), 1 PIN code (6 digits); 3 dates : 1 spontaneous (year of birth), 1 prompted date (word style), 1 relative and general date exp.; 1 word spotting phrase using an application word (embedded); 1 isolated digit; 3 spelled word : 1 spontaneous (own forename), 1 spelling of directory city name, 1 real word for coverage; 1 currency money amount; 1 natural number; 5 directory assistance : 1 spontaneous, own forename, 1 city of school at 7 years (spontaneous), 1 most frequent cities (set of 500), 1 most frequent company/agency (set of 500 names), 1 "forename surname" (set of 500 names); yes/no questions : 1 predominantly "yes" question, 1 predominantly "no" question; 9 phonetically rich sentences; 2 time phrases : 1 time of day (spontaneous), 1 time phrase (word style); 4 phonetically rich words.

The database also contains sentences uttered by all speakers for speaker verification purposes and dialectal studies. Each speaker uttered the same 8 sentences and connected digits strings (3-6 digits).

The following age distribution has been obtained: 315 speakers are below 16 years old, 2095 speakers are between 16 and 30, 1080 speakers are between 31 and 45, 1078 speakers are between 46 and 60, and 432 speakers are over 60.

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

Price for ELRA members:	for research use: € 35,000	for commercial use: € 50,000
Price for non members:	for research use: € 60,000	for commercial use: € 70,000

Introduction on the PAROLE project

LE-PAROLE project (MLAP/LE2-4017) aims to offer a large-scale harmonised set of "core" corpora and lexica for all European Union languages. Language corpora and lexica were built according to the same design and composition principles, in the period 1996-1998.

PAROLE Corpora

The harmonisation with respect to corpus composition (selection of corpus texts) was to be achieved by the obligatory application of common parameters for time of production and classification according to publication medium. No texts older than 1970 were allowed. The corpus had to include specific proportions of texts from the categories "Book", "Newspaper", "Periodical" and "Miscellaneous" within a settled range. With respect to the mark up of text structure and primary data, every single corpus text was to be encoded according to the PAROLE DTD, which is compatible with the DTD of the Text Encoding Initiative (TEI) and with that of the Corpus Encoding Standard (CES). As for linguistic corpus annotation, an equal proportion of the corpus texts (up to 250,000 running words) was to be morphosyntactically annotated according to a common core PAROLE tagset, extended with a set of language specific features. The checking of the tags was split in two: 50,000 words had to be checked for maximum granularity and 200,000 for part-of-speech (PoS) only.

Languages: Catalan, Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portuguese, Swedish, Belgian French, Irish, Norwegian.

PAROLE Lexica

The lexica (20,000 entries per language) were built conform to a model based on EAGLES guidelines and GENELEX results, underlying a common lexical tool adapted from the EUREKA-GENELEX project. This software tool was extended to support the PAROLE model and conversion and management processes of the resulting resources.

Languages: Catalan, Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portugese, Swedish, Spanish.

For more information on the LE-PAROLE project, visit the following Web sites:

http://www2.echo.lu/langeng/projects/le-parole/summary.html http://www.linglink.lu/le/projects/le-parole/index.html

ELRA-W0019 Dutch PAROLE Distributable Corpus

The Dutch PAROLE Distributable Corpus is a 3 million words selection from the 20 million words Dutch PAROLE Reference corpus. The Dutch corpus annotation and checking was made accordingly to the common core PAROLE tagset. The Dutch data were also checked for type.

The Dutch PAROLE Distributable Corpus contains the following texts:

BOOKS - Van Sterkenburg from:

- Wdlijst tot wdboek (1984, 65,640 words)
- Taal vt Journaal (1989, 57,101 words)
- WNT-portret (1992, 60,214 words)

NEWSPAPERS - Short Newspaper texts from:

- MN_Collection (1986-1988, 19,542 words)
- CVNP(S)-Collection (1983-1990, 179,220 words)

PERIODICAL - Short texts from:

- Local Papers (1985-1988, 47,019 words)
- Magazines (1985-1989, 164,589 words)

MISCELLANEOUS - Texts to be read out in TV-news broadcasts for:

- General audience (1992-1995, 1,285,824 words)
- Youth (1991-1995, 1,008,658 words)
- Short texts from Ephemera (1985-1986, 131,692 words)

Over 250,000 words of corpus texts (with TEI markup suppressed) have been PoS-tagged automatically. A total of 59,798 running words has been manually corrected and checked at least two times with respect to maximal granularity, according to a lexicographer's manual. The extra 9,000 words over the required 50,000 words compensate for the occurrence of ca. 5,300 'keywords' in the original texts. The fully corrected material has been subjected to an automated post-control operation, checking the pertinence relations between the various feature values, and instantiating default values in case a mismatch (indicating a correction error) was found. Ca. 200,000 words have been checked once for PoS and type. In addition to the required PoS, type was checked for reasons of quality. This material has been subjected to an automated correction procedure addressing the feature slots (positions) beyond the first two for PoS and type so as to solve discrepancies between the manually corrected PoS and type, and the possibly erroneous, automatically assigned values of the remaining slots.

Price for ELRA members:	for research academic use: € 300 for commercial use: € 2,000	for research use by a commercial company: € 1,000
Price for non members:	for research academic use: € 450 for commercial use: € 3,000	for research use by a commercial company: € 1,500

ELRA-L0031 Dutch PAROLE lexicon

The entry list of the lexicon consists of about 20,200 entries distributed over 13 parts of speech (POS). The entries have been described along the dimensions of morphosyntax and syntax. Morphosyntactic information consists of various lexical properties, like gender, number, case, person, inflection, etc. Syntactic descriptions consist of typical complementation patterns associated with the various lemmata.

The composition of the entry list of the lexicon is based on 3 corpora from the Instituut voor Nederlandse Lexicologie (INL) and 2 lexica. The corpora contain a total of about 54 million words and have been automatically annotated for part-of-speech and lemma. The lexica contain morphosyntactic information of various kinds. For verbs, nouns, adjectives and adverbs, lemmata that were covered by at least 2 corpora and the 2 lexica were selected on the basis of cumulative frequency, coverage (distribution over sources) and inflected forms. For the smaller parts of speech, these selection requirements appeared to be too strict. Entry selection for these parts of speech was based on ranked frequency.

The entries, uniquely defined by the combination of part of speech (e.g. noun) and subtype (e.g. common vs. proper noun), are provided with morphosyntactic information according to the Dutch set of PAROLE categories and features, and, where available, with syntactic

(ELRA-L0031 Dutch PAROLE lexicon continued)

information. Morphosyntactic information is automatically extracted from the INL lexica. Syntactic data have been collected manually, by inspection of corpus data and - where necessary - consultation of reference works. The corpus consulted consists of the newspaper component and the varied component of the 38 Million Words Corpus 1996.

The lexicon is set up as an SGML file (over 30 MB of plain ASCII). Its contents have been encoded in a distributed manner: all formative entities (like lemmata, syntactic phrases, feature bundles) are SGML entities, related by a pointer mechanism to other entities.

The lexicon contains the following categories: adjectives (3,298 entries), adpositions (80 entries), adverbs (554 entries), articles (3 entries), conjunctions (70 entries), determiners (59 entries), interjections (235 entries), nouns (12,279 entries), numerals (77 entries), pronouns (85 entries), residuals (186 entries), unique (1 entry), verb (3,274 entries).

Price for ELRA members:	for research academic use: € 400 for commercial use: € 9,000	for research use by a commercial company: € 2,000
Price for non members:	for research academic use: € 600 for commercial use: € 13,500	for research use by a commercial company: € 3,000

EUROWORDNET

Now available: Dutch wordnet; English wordnet (additional relations which are missing in WordNet1.5); Spanish wordnet.
Available soon: Czech wordnet; Estonian wordnet; French wordnet; German wordnet; Italian wordnet.

The multilingual EUROWORDNET database consists of the following modules:

A. LR(1) Common Components

- The Inter-Lingual-Index, which is a list of records (ILI-records), in the form of synsets mainly taken from WordNet1.5 or manually created. An ILI-record contains: synset (set of synonymous words or phrases, mostly from WordNet1.5); part-of-speech; one or more Top-Concept (Optional); one or more Domain labels (Options); a gloss in English (mostly from WordNet1.5); a unique ID linking the synset to its source (mostly WordNet1.5).
- Top-Ontology: an ontology of 63 basic semantic classes based on fundamental distinctions. By means of the Top-Ontology all the wordnets can be accessed using a single language-independent classification-scheme. Top-Concepts are only assigned to ILI-records.
- Domain-Ontology: an ontology of subject-domains optionally assigned to ILI-records
- A selection of ILI-records, the so-called Base-Concepts, which play a major role in the different wordnets. These Base-Concepts form the core of all the wordnets. All the Base-Concepts are classified in terms of the Top-Concepts that apply to them.
- WordNet1.5 (91,591 synsets; 168,217 meanings; 126,520 entry words) in EuroWordNet format.

B. LR(2) Language-Specific Components

The specific wordnets are language-internal structures, minimally containing : set of variants or synonyms making up the synset; part-of-speech. language-internal relations to other synsets; equivalence relations with ILI-records; a unique-id linking the synset to its source.

WordNet1.5 is itself also distributed as part of EuroWordNet and is as such free. WordNet1.5 is the property of Princeton University.

Each wordnet is distributed with LR1 and includes documentation on LR1 and the distributed wordnet. All the data are distributed as text-files in the EuroWordNet import format and as Polaris database files (see below LR3). The EuroWordNet viewer (Periscope, see below LR3) can be used to access the database version. Polaris has to be licensed to modify and extend the database version.

The wordnets are distributed without: glosses, usage labels, morpho-syntactic properties, examples, word-to-word translations.

C. LR(3) Software

The multilingual EUROWORDNET Database consists of three components:

- The actual wordnets in Flaim database format: an indexing and compression format of Novell.
- Polaris (Louw 1997): a wordnet editing tool for creating, editing and exporting wordnets.
- Periscope (Cuypers and Adriaens 1997): a graphical database viewer for viewing and exporting wordnets.

The Polaris tool can import new wordnets or wordnet fragments from ASCII files with the correct import format and it creates an indexed EUROWORDNET Database. Furthermore, it allows a user to edit and add relations in the wordnets and to formulate queries. The Polaris toolkit makes it possible to visualise the semantic relations as a tree-structure that can directly be edited. These trees can be expanded and shrunk by clicking on word-meanings and by specifying so-called TABs indicating the kind and depth of relations that need to be shown. Expanded trees or sub-trees can be stored as a set of synsets, which can be manipulated, saved or loaded. Additionally, it is possible to access the ILI or the ontologies, and to switch between the wordnets and ontologies via the ILI. Finally, it contains an interface to project sets of synsets across wordnets.

The Periscope program is a public viewer that can be used to look at wordnets created by the Polaris tool and to compare them in a graphical interface. Word meanings can be looked up and trees can be expanded. Individual meanings or complete branches can be projected on another wordnet or wordnet structures can be compared via the equivalence relations with the Inter-Lingual-Index. Selected trees can be exported to text files. The Periscope program cannot be used for importing or changing wordnets.

The prices are based on the number of synsets in each wordnet and differ for the kind of usage and ELRA-membership.

For prices, please contact ELRA (or see ELRA Web site).

Technical support may be provided by members of the consortium. It will be implemented through bilateral agreements between the user and the member of the consortium responsible for the data acquired.