NATURAL LANGUAGE PROCESSING AT THE XEROX RESEARCH CENTRE EUROPE

Jean-Pierre Chanod
Xerox Research Centre Europe
6, chemin de Maupertuis, 38240 Meylan, France
Tel.: +33 (0)4 76 61 50 75
e-mail: Chanod@xrce.xerox.com

ABSTRACT

The Xerox Research Centre Europe (XRCE, see http://www.rxce.xerox.com for more information) pursues a vision of document technology where language, physical location and medium - electronic, paper or other - impose no barrier to effective use.
Our primary activity is research. Our second activity is a Program of Advanced Technology Development, to create new document services based on our own research and that of the wider Xerox community. We also participate actively in exchange programs with European partners.
Language issues cover important aspects in the production and use of documents. As such, language is a central theme of our research activities. More particularly, our Centre focuses on multilingual aspects of Natural Language Processing (NLP). Our current developments cover more than fifteen European languages and some non-European languages such as Arabic, Turkish or Chinese. Some of these developments are conducted through direct collaboration with academic institutions all over Europe.
The present article is an introduction to our basic linguistic components and to some of their multilingual applications.

1 LINGUISTIC COMPONENTS

The MLTT (Multilingual Theory and Technology) team creates basic components for linguistic analysis, e.g. morphological analysers, taggers, parsing and generation platforms. These components are used to develop descriptions of various languages and the relation between them. They are later integrated into higher level applications, such as terminology extraction, information retrieval or translation aid. The Xerox Linguistic Development Architecture (XeLDA) developed by the Advanced Technology Systems group incorporates the MLTT language technology.
Finite-state technology is the fundamental technology on which Xerox language R&D is based. It encompasses both work on the basic calculus and on linguistic tools, in particular in the domain of morphology and syntax.

1.1 Finite-state calculus

The basic calculus is built on a central library that implements the fundamental operations on finite-state networks. It is based on long-term Xerox research, originated at PARC in the early 1980s. The most recent development in the finite-state calculus is the introduction of the replace operator [12]. The replacement operation is defined in a very general way, allowing replacement to be constrained by input and output contexts, as in two-level rules but without the restriction of only single-symbol replacements. Replacements can be combined with other kinds of operations, such as composition and union, to form complex expressions.
The finite-state calculus is widely used in our linguistic development, to create tokenisers, morphological analysers, noun phrase extractors, shallow parsers and other language-specific linguistic components.
The XRCE web site provides tutorials on finite-state technology as well as interactive tests.

1.2 Morphology

The MLTT work on morphology is based on the fundamental insight that word formation and morphological or orthographic alternation can be solved with the help of finite automata [11,13]:

1. the allowed combinations of morphemes can be encoded as a finite-state network;
2. the rules that determine the form of each morpheme can be implemented as finite-state transducers;
3. the lexicon network and the rule transducers can be composed into a single automaton, a lexical transducer, that contains all the morphological information about the language including derivation, inflection, and compounding.

Lexical transducers have many advantages. They are bi-directional (the same network for both analysis and generation), fast (thousands of words per second), and compact.
We have created comprehensive morphological analysers for many languages including English, German, Dutch, French, Italian, Spanish, and Portuguese. More recent developments include Czech, Hungarian, Polish, Russian, Scandinavian languages and Arabic.

1.3 Part-of-speech tagging

The general purpose of a part-of-speech tagger [4,7,14] is to associate each word in a text with its morphosyntactic category (represented by a tag), as in the following example:

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

The process of tagging consists in three steps:

1. tokenisation: break a text into tokens
2. lexical lookup: provide all potential tags for each token
3. disambiguation: assign to each token a single tag

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

- The tokenisation step uses a finite-state transducer to insert token boundaries around simple words (or multi-word expressions), punctuation, numbers, etc.
- Lexical lookup requires a morphological analyser to associate each token with one or more readings. A guesser which provides potential part-of-speech categories based on affix patterns handles unknown words.
- Disambiguation is done with statistical methods (Hidden Markov Model), although we also experiment with rule-based methods.

1.4 Incremental finite-state parsing

Finite-State Parsing is an extension of finite state technology to the level of phrases and sentences.
Our work [1,2] concentrates on shallow parsing of unrestricted texts, e.g. technical documentation, newspaper articles, web pages. We compute syntactic structures, without fully analysing linguistic phenomena that require deep semantic or pragmatic knowledge. For instance, PP-attachment, co-ordinated or elliptic structures are not always fully analysed. The annotation scheme remains underspecified with respect to yet unresolved issues. On the other hand, such unresolved phenomena do not cause parse failures, even on complex sentences.
Syntactic information is added at the sentence level in an incremental way, depending on the contextual information available at a given stage. The implementation relies on a sequence of networks built with the replace operator. The current system has been implemented for French and is being expanded to new languages. The parsing process is incremental in the sense that the linguistic description attached to a given transducer in the sequence relies on the preceding sequence of transducers, covers only some occurrences of a given linguistic phenomenon and can be revised at a later stage.
The parser output can be used for further processing such as extraction of dependency relations over unrestricted corpora. In tests on French corpora (technical manuals, newspaper), precision is around 90-97% for subjects (84-88% for objects) and recall around 86-92% for subjects (80-90% for objects).

1.5 The LFG PARGRAM project

The LFG PARGRAM project [8,9,10,15] is a collaborative effort involving researchers from Xerox PARC in Palo Alto, the Xerox Research Centre in Grenoble, France, and the University of Stuttgart in Stuttgart, Germany. The aim of the project is to produce wide coverage LFG grammars for English, French, and German which are written collaboratively, based on a common set of linguistic principles and with a commonly agreed upon set of grammatical features.
The grammarians use a new platform, the Xerox Linguistic Environment, which is still under development; a unification-based generator is also under development.
The grammars consist of phrase-structure rules and abbreviatory rule macros; LFG allows the right-hand side of phrase structure rules to consist of regular expressions (including the Kleene Star notation) and arbitrary Boolean combinations of regular predicates, so the rules in the grammar actually abbreviate a large set of rules written in a more conventional framework. The lexicons used by the sites consist of entries for stems, template definitions, and lexical rules. The Xerox Linguistic Environment allows for an interface to an external finite-state morphological analyser, and so the lexicons include entries for the information about morphological inflection supplied by the analyser.

2 MULTILINGUALITY AND INTERNATIONAL COLLABORATION

When XRCE was created in 1993, we engaged in a systematic effort to expand Xerox language components to a large set of new languages.
Such developments were initially conducted in-house (English, French, German, Italian, Spanish, Portuguese, Dutch). However, it soon became clear that our endeavours had to be supported by international co-operation, to take advantage of the expertise and previous work done in the academic communities and sometimes in SMEs of the various countries where the languages are spoken.
This international effort started in 1995 in direction of Central and Eastern Europe (CEE).
Our developments for CEE languages strongly rely on collaborations, which include direct collaborations with academic or commercial partners as well as participation in European projects, such as TELRI, Elsnet-Goes-East or the Copernicus program. Our academic partners include, among others, the Universities of Prague and Warsaw, the Academies of Science in Russia, Romania and Bulgaria. Our collaboration also includes exchange programs such as visits and training sessions in Grenoble

or in CEE and participation in summer schools such as EUROLAN 97 in Tusnad.

Our strategy with respect to the development of CEE languages, as for any new language, is to adopt an incremental approach. We first want to build basic tools that can be integrated into existing platforms and applications. While collaboration develops on such grounds, we hope to establish long term relations with the research labs in CEE, and, in parallel, we expect to see new business opportunities emerge for language technologies.

As far as applications are concerned, our major field of interest is in authoring and translation tools and comprehension aid of foreign language texts. Another area of interest is that of digital libraries, a means to give quicker and broader access to large repositories of knowledge via the network.

At the moment, on-going collaborations include the extension of our basic linguistic technology (e.g. morphological analysis, part-of-speech disambiguation) to new languages such as Czech, Hungarian, Polish, Romanian and Russian. Morphological analysers for such languages are already available, as well as part-of-speech taggers.

We certainly want to develop new language resources that will create new opportunities for document management in multilingual environments. But there is more to this international collaboration. We benefit a lot from exchanges and shared expertise. It brings in new research ideas. The simultaneous study of different languages stimulates the development of new algorithms and of more powerful architectures. Actually, beyond developing basic language tools, we already collaborate with CEE partners on broader themes, such as machine learning (inductive logic programming) in collaboration with Jozef Stefan Institute, Ljubljana, or translation aids and multilingual lexical databases within various European projects.

2.1 European projects

The EU programs dealing with CEE are great enablers for our purpose. We participated in conferences and awareness seminars organised by Elsnet-Goes-East or by TELRI. Such events are extremely useful to us and most of our on-going collaborations were actually initiated through such events.

At the same time, we participate in European Copernicus projects, GLOSSER, which was successfully completed, STEEL and CONCEDE.

Such European projects address problems of effective communication, which are especially relevant in CEE, an emerging market with broad language diversity.

GLOSSER

GLOSSER aims at applying state-of-the-art linguistic technology to Computer-Assisted Language Learning. The project ran for two years, until April 1997 and built three software prototypes as a demonstration of concept. The GLOSSER prototypes are designed to help people who partially know a language but cannot read it quickly. The project vision is that speakers of Bulgarian, Estonian or Hungarian, with partial knowledge of English might read e.g. a software manual on the screen. Upon encountering an unknown or an unfamiliar use of known words the user can mouse it to invoke on-line help (follow a dynamic hyperlink). Help provides the following facilities:

- a morpho-syntactic analysis of the clicked words
- the entry of the corresponding stem in a bilingual dictionary
- similar examples in on-line bilingual corpora.

The project covers Bulgarian, Estonian and Hungarian as far as CEE languages are concerned. The project was co-ordinated by Alfa-informatica, Groningen University and involved the Bulgarian Academy of Science (Linguistic Modelling Laboratory assisted by the Institute for Bulgarian Language), the University of Tartu assisted by the Institute for Estonian Language (Tallinn), Morphologic from Budapest and the Xerox Research Centre Europe.

STEEL

The STEEL project will extend the functionality of existing translation aid tools to Czech and Polish with a special interest in providing translation assistance for technical and specialised documentation. This can be useful for comprehending documents such as textbooks or manuals that often contain special terminology, which is not part of a foreign reader's basic vocabulary.

The project will capitalise on experience and tools previously built as part of the Locolex/COMPASS EU project (comprehension assistant for English, French and German texts). In particular, the STEEL project will reuse and adapt the Locolex engine and interface.

Technically, the project will focus on the following aspects:

- updating existing bilingual resources (Czech or Polish as one language and English as the other one) and converting such resources to the format required by the Locolex engine;
- creating new bilingual terminological resources (including Czech or Polish as one language) in a semi-automatic fashion;
- adapting the Locolex engine and interface to the new languages and to new user needs in the specialised domain of terminology.

The project includes lexicographic work (dictionary revision and enhancement, validation of terminology extraction, validation of the resulting translation aid tool) and computational work (adaptation of existing software

to new languages, integration of newly developed linguistic resources in Czech and Polish for terminology extraction and alignment, integration of all the tools into the end user application).
Our STEEL partners include Prague, Warsaw, Tubingen and Lyon Universities, as well as two SMEs, Moravia Translations from Brno and Lexis from Wroclaw.

CONCEDE
XRCE recently joined a new European project, CONCEDE. The project will develop lexical databases for six CEE languages: Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovene. It will deliver medium-size TEI (Text Encoding Initiative) conformant lexicons, suitable for Language Engineering use. The CONCEDE lexical databases will be derived from and integrated with the MULTEXT-EAST parallel aligned corpus. The combined results of the two projects will constitute an integrated multilingual resource, which fully exploits the results of both projects.
Participating in this project is a great opportunity for Xerox. For one thing, it is a natural continuation of many of the partnerships we have already established in CEE, and a good sign that such collaborations fertilise R&D. Second, it addresses technical issues which are of strategic importance for our research, such as lexical databases and multilinguality, and we expect to gain a lot from joint work with the best experts in Europe at large.
Based on the positive lessons learnt from our collaboration in CEE, we will now develop international collaboration in a more systematic way, through the newly established Xerox Language Resources Group. This group will coordinate activities concerning the acquisition, development, maintenance and testing of language resources, such as mono- and multilingual dictionaries and corpora, part-of-speech taggers and shallow parsers. This will include developing a large community of users and experimenters for our basic tools, through research licenses, which will give all of us a much better view of the technical and human challenges involved by the multilingual information society.

3 APPLICATIONS

3.1 LOCOLEX: a Machine Aided Comprehension Dictionary

LOCOLEX is an on-line bilingual comprehension dictionary, which aids the understanding of electronic documents written in a foreign language [3,16,17]. It displays only the appropriate part of a dictionary entry when a user clicks on a word in a given context. The system disambiguates parts of speech and recognises multiword expressions such as compounds (e.g. heart attack), phrasal verbs (e.g. to nit pick), idiomatic expressions (e.g. to take the bull by the horns) and proverbs (e.g. birds of a feather flock together). In such cases LOCOLEX displays the translation of the whole phrase and not the translation of the word the user has clicked on.
For instance, someone may use a French/English dictionary to understand the following text written in French:

Lorsqu'on évoque devant les cadres la séparation négociée, les rumeurs fantaisistes vont apparemment toujours bon train.

When the user clicks on the word *cadres*, LOCOLEX identifies its POS and base form. It then displays the corresponding entry, here the noun *cadre*, with its different sense indicators and associated translations. In this particular context, the verb reading of *cadres* is ignored by LOCOLEX. Actually, in order to make the entry easier to use, only essential elements are displayed:

cadre I: nm
1: [constr,art] (of a picture, a window) frame
2: (scenery) setting
3: (milieu) surroundings
4: (structure, context) framework
5: (employee) executive
6: (of a bike, motorcycle) frame

The word *train* in the same example above is part of a verbal multiword expression *aller bon train*. In our example, the expression is inflected and two adverbs have been stuck in between the head verb and its complement. Still LOCOLEX retrieves only the equivalent expression in English to be *flying around* and not the entire entry for *train.*

train I: nm
5 : [rumeurs] aller bon train : to be flying round

LOCOLEX uses an SGML-tagged bilingual dictionary (the Oxford-Hachette French English Dictionary). To adapt this dictionary to LOCOLEX required the following:

- Revision of an SGML-tagged Dictionary to build a disambiguated active dictionary (DAD);
- Rewriting multi-word expressions as regular expressions using a special grammar;
- Building a finite state machine, which compactly associates index numbers with dictionary entries.

The lookup process itself may be represented as follows:

- split the sentence string into words (tokenisation);
- normalise each word to a standard form by changing cases and considering spelling variants;
- identify all possible morpho-syntatic usages (base form and morpho-syntactic tags) for each word in the sentence;
- disambiguate the POS;

- find relevant entries (including possible homographs or compounds) in the dictionary for the lexical form(s) chosen by the POS disambiguator;
- use the result of the morphological analysis and disambiguation to eliminate irrelevant sections;
- process the regular expressions to see if they match the word's actual context in order to identify special or idiomatic usages;
- display to the user only the most appropriate translation based on the part of speech and surrounding context.

Besides being an effective tool for understanding, LOCOLEX could also be useful in the framework of language learning. LOCOLEX also points out that existing on-line dictionaries, even when organised like a database rather than a set of type-setting instructions, are not necessarily suitable for NLP-applications. By adding grammar rules to the dictionary in order to describe the possible variations of multiword expressions we add a dynamic feature to this dictionary. SGML functions no longer point to text but to programs.

3.2 Text Mining and Multilingual Information Retrieval

Many of the linguistic tools being developed at our Centre are being used in applied research into text mining and multilingual information retrieval [5,6]. Multilingual information retrieval allows the interrogation of texts written in a target language B by users asking questions in source language A.
In order to perform this retrieval, the following linguistic processing steps are performed on the documents and the query:

- Automatically recognise language of the text.
- Perform the morphological analysis of the text using Xerox finite state analysers.
- Part-of-speech tag the words in the text using the preceding morphological analysis and the probability of finding part-of-speech tag paths in the text.
- Lemmatise, i.e. normalise or reduce to dictionary entry form, the words in the text using the part of speech tags.

This morphological analysis, tagging, and subsequent lemmatisation of analysed words has proved to be a useful improvement for information retrieval as any information retrieval specific stemming. To process a given query, an intermediate form of the query must be generated which he normalised language of the query to the indexed text of the documents. This intermediate form can be constructed by replacing each word with target language words through an on-line bilingual dictionary. The intermediate query, which is in the same language as the target documents, is passed along to a traditional information retrieval system, such as SMART1. This simple word-based method is the first approach we have been testing. Initial runs indicate that incorporating multi-word expression matching can significantly improve results. The multi-word expressions most interesting for information retrieval are terminological expressions, which most often appear as noun phrases in English.

3.3 Callimaque: a collaborative project for virtual libraries

Digital libraries represent a new way of accessing information distributed all over the world, via the use of a computer connected to the Internet network. Whereas a physical library deals primarily with physical data, a digital library deals with electronic documents such as texts, pictures, sounds and video.
We expect more from a digital library than only the possibility of browsing its documents. A digital library front-end should provide users with a set of tools for querying and retrieving information, as well as annotating pages of a document, defining hyper-links between pages or helping to understand multilingual documents.
Callimaque is one of our projects dealing with such new functionalities for digital libraries. More precisely, Callimaque is a collaborative project between the Xerox Research Centre and research/academic institutions of the Grenoble area (IMAG, INRIA, CICG). The goal is to build a virtual library that reconstructs the early history of information technology in France. The project is based on a similar project, the Class project, which was started by the University of Cornell several years ago under the leadership of Stuart Lynn to preserve brittling old books. The Class project runs over conventional networks and all scanned material is in English.
The Callimaque project includes the following steps:

- Scanning and indexing around 1000 technical reports and 2000 theses written at the University of Grenoble, using Xerox XDOD, a system integrated with a scanner, a PC, a high-speed printer, software for dequeueing, indexing, storing, etc. Numerised documents can be reworked page by page and even restructured at the user's convenience. 30 Gbytes of memory are needed to store the images. Abstracts are OCRed to permit textual search.
- Documents are recorded on a relational database on a UNIX server. A number of identifiers (title, author, reference number, abstract, etc.) are associated with each document to facilitate the search
- Multilingual terminology derived from multilingual abstracts allows the system to process non-French queries.
- With a view to making these documents widely accessible, Xerox has developed software which authorises access to this database by any client using the http protocol used by the World Wide Web. The base is thus accessible via any PC, Macintosh, UNIX

station or even from a simple ASCII terminal (The web address is http://callimaque.grenet.fr).

- Print on demand facilities connected to the network allow the users to make copies of the scanned material. This connection will subsequently develop towards a high output ATM network.

3.4 Translation and authoring systems

Xerox linguistic tools are embedded in various products, including translation and authoring systems. They provide professional authors and translators with a suite of software support tools designed to facilitate their tasks and reduce their workload. The offer includes terminology management and translation memory. Such applications are based on XeLDA, the Xerox linguistic engine, which integrates the linguistic components, described above.

3.4.1 Terminology Suite

The Terminology Suite (see the diagram below) includes the following components, implemented for Windows 95 and Windows NT 4.0 and above:

TermFinder: Multilingual Terminology Extraction
TermFinder enables the user to semi-automatically create multilingual terminology, hence ensuring a huge productivity increase over manual terminology creation. TermFinder is based on the linguistic components described above, especially NP extraction tools and alignment. TermFinder supports Dutch, English, French, German, Italian, Spanish, and Portuguese. Any of these languages can be source or target.
In addition, Danish, Swedish, Finnish, Norwegian, Czech, Hungarian, Russian, Romanian, Polish, Arabic, Japanese, Korean are under development.

Built on top of Open Database Connectivity (ODBC), the database independent layer from Microsoft, TermFinder is independent from a specific database. TermFinder supports SGML, HTML, XML, iso-8859-1 and Rich Text Format documents.

TermManager : Terminology Database in Context
TermManager is the complement to TermFinder. It enables one to quickly manage the terminology that was created with TermFinder. One can modify it, add terms, remove others, and add specific information. The Term In Context view enables users to see all occurrences of a term in the context of the original sentences.
TermManager uses several views to display the terminology: Form View, to view all the information related to a term, Table View, to see information related to several terms, Dictionary view: to see terms that are related. One can define filters to see only a subset of the database. One can customise fonts, colours. One can create one's own fields to store user defined information.

TermChecker : Controlled Terminology Tool
The terminology that has been built using TermFinder can then be used by TermChecker to provide authors with interactive feedback, to help them increase the terminology consistency. This tool can be used both by the author for the source terminology and by the translator for the target terminology.
TermChecker is fully integrated with word processors. It provides the same look and feel than the standard spell checker function.

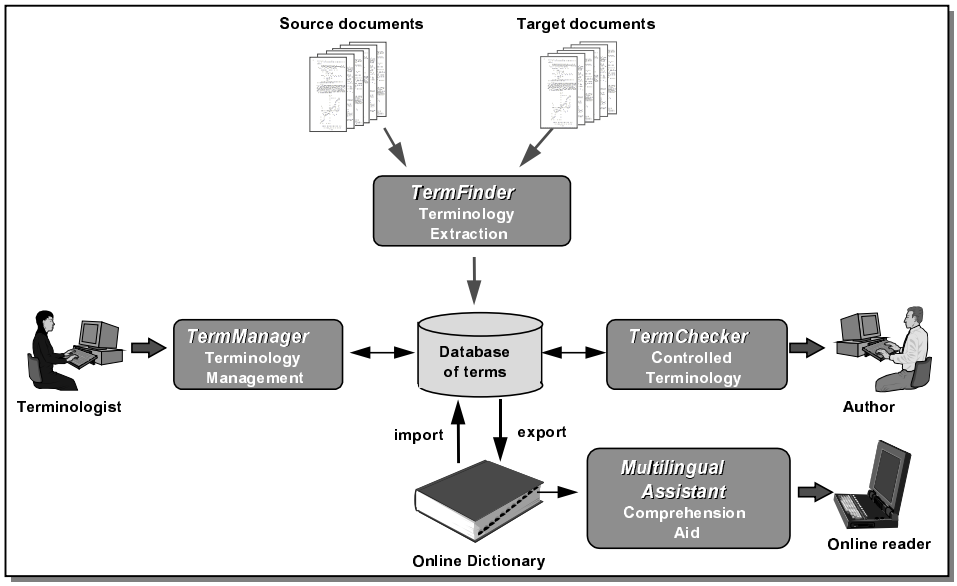

Figure 1: Terminology Suite

Multilingual Assistant : Comprehension Aid Tool
This Multilingual Assistant provides translation of words in context, using a general or specialised dictionary. It can differentiate between similar expressions that should be translated differently ("apply to" vs. "apply something to"). The Multilingual Assistant is based on the results of the Locolex project described above.

3.4.2 Translation memory

Translation memory helps the translation process by recognising previously translated texts: the system "keeps" sentences that have been previously translated, with their corresponding translation. When a new document has to be translated, or an updated version of an existing document, the translation memory can rapidly find identical or similar sentences and retrieve them for the translator to view. This will save any unnecessary duplication of work for the translator whilst increasing consistency and quality of translations. By cutting down on the repetitious and routine work, Translation Memory frees up the translator to focus on new texts and thereby reduce the overall time and cost of translation.

The Filter: A filter receives the source document to be translated which it parses, extracting information about the structure, such as titles, styles, paragraph marks etc. The process simultaneously extracts the text itself, plus some additional formatting, such as character style, (bold, italic, underlined…) in order to store as much data as possible to reduce the efforts of the human translator. This format information is stored independently from the format of the input document and so can relate to parts of text as well as the whole text. Additional data can be added such as page numbers, document identification etc.
The filter can read the most well known document formats (RTF SGML HTML MIF Interleaf) and in this way is word processor independent. The filter reads character codes in English, French, German, Italian, Spanish, Portuguese and Dutch for the source documents. An indefinite number of target languages can be supported when written in Unicode characters.

Segmentation: The input text is split up into units of translation which are to be stored in the translation memory database, normally consisting of whole sentences and their formatting. This formatting is copied to the output sentences without any modifications. However, other pieces of text may be considered as translation units, such as titles, lists, figures, captions etc. and stored accordingly. A list of abbreviations is maintained to enable proper recognition by the user, for example to avoid interpreting every occurrence of a period as the end of a sentence. This list can be extended and modified

Translation Memory Core System: it performs several functions:

- Manages the translation memory database (storage, administration, import/export)
- Processes the source sentences by retrieving them from the translation memory and/or by retrieving similar sentences
- Retrieves the translation which has been stored for matching sentences (perfect matching) and, in the case of non-identical sentences (fuzzy matching or no match), generating a close translation.

Storage and Administration: Documents to be translated are grouped together to form projects and assigned a manager who will define the characteristics of that project, by domain, customer, source language and target language for example. The manager can add/remove texts to/from the project, delete them, file them and merge two translation memories if required.

Search and Retrieval: Input for translation memory consists of sentences with some formatting information. Searches for these sentences can take place in more than one translation memory and can be defined and prioritised by the user, to obtain the best matches first. Any differences between the input sentence and the matching sentence are taken into account by the system and include:

- formatting differences; some characters do not have the same style
- case differences
- punctuation differences
- words are substituted; changes in proper nouns, acronyms, numbers
- linguistic differences; one word has the same base form but not the same surface form - number, tense, gender
- insertion or deletion of one or more words; modifiers (adverbs and adjectives) are different but head words (nouns, verbs) are the same
- changes in the order of phrases
- changes in the order of words

The Translator's Workbench: The workbench is the store for sentences and their matches. It allows the translator to translate sentences that have not been found and to verify matches (perfect and fuzzy) that have been found in the translation memory. The workbench can take information from several translators and merge information from several documents.
It provides a graphical interface, which displays as much information as possible to help the translator work quickly and efficiently.

4 CONCLUSION

This article cannot cover all our research and development activities in great details. As we briefly

indicated, new areas of interest cover lexical databases, building of ontologies, sense disambiguation or machine learning (inductive logic programming).

If some tools are already integrated in real life products and services, such as translation memory, mono- and bilingual terminology extraction or multilingual assistant, we do prepare for the next generation of linguistic tools and services, with higher accuracy, broader coverage and finer grained analysis.

We may foresee that, in the near future, language technology will be more systematically embedded in multiple environments, ranging from intelligent networked copiers to web applications, providing summarisation, translation and authoring aids or efficient knowledge management.

This will also serve the goal of preserving language and cultural diversity in the information society.

5 REFERENCES

[1] Salah Aït-Mokhtar, Jean-Pierre Chanod, "Incremental finite-state parsing", in Proc. of Applied Natural Language Processing 1997, Washington, DC. April 97

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

[3] D. Bauer, F. Segond, A. Zaenen. "LOCOLEX: the translation rolls off your tongue." in Proc. of the ACH-ALLC conference, Santa Barbara, pp. 6-8, 1995.

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

[5] Gregory Grefenstette. Explorations in Automatic Thesaurus Discovery. Kluwer Academic Press, Boston, 1994.

[6] Gregory Grefenstette, Ulrich Heid and Thierry Fontenelle. "The DECIDE project: Multilingual Collocation Extraction." Seventh Euralex International Congress, University of Gothenburg, Sweden, Aug 13-18, 1996.

[7] Barbora Hladka and Jan Hajič. "Probabilistic and Rule-based Tagger of an Inflective Language". In Proc. of Applied Natural Language Processing 1997 Washington, DC. April 97

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

[9] Kaplan, Ronald M. and Joan Bresnan. 1982. Lexical-Functional Grammar: A formal system for grammatical representation. In Joan Bresnan, editor, The Mental Representation of Grammatical Relations. The MIT Press, Cambridge, MA, pages 173--281.

[10] Kaplan, Ronald M. and John T. Maxwell. 1996. LFG grammar writer's workbench. Technical report, Xerox PARC.

[11] Lauri Karttunen. "Constructing Lexical Transducers". In Proc. of the 15th International Conference on Computational Linguistics, Coling, Kyoto, Japan, 1994.

[12] Lauri Karttunen. "The Replace Operator. In Proc. of the 33rd Annual Meeting of the Association for Computational Linguistics, ACL-95} 16-23, Boston, 1995.

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

[14] Julian Kupiec and Mike Wilkens. The DDS tagger guide version 1.1. Technical report, Xerox Palo Alto Research Center, 1994.

[15] [Maxwell, III, John T. and Ronald M. Kaplan. 1991. A method for disjunctive constraint satisfaction. In Masaru Tomita, editor, Current Issues in Parsing Technology. Kluwer Academic Publishers, Dordrecht, pages 173--190.

[16] John Nerbonne, Lauri Karttunen, Elena Paskaleva, Gabor Proszeky and Tiit Roosmaa. "Reading more into Foreign Languages". In Proc. of Applied Natural Language Processing 1997. Washington, DC. April 97

[17] F. Segond and P. Tapanainen. Using a finite-state based formalism to identify and generate multiword expressions. Technical Report MLTT-019, Xerox Research Centre, Grenoble, 1995.