Multilingual Tools at the Xerox Research Centre

Jean-Pierre Chanod

Introduction

The Xerox Research Centre Europe (http://www.xrce.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 ten European languages and some non-European languages such as Arabic. 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 tools for linguistic analysis, e.g. morphological analysers, taggers, parsing and generation platforms. These tools 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.

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. 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.

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:

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.

Part-of-speech tagging

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

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

The process of tagging consists in three steps:

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

Each step is performed by an application program 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. Unknown words are handled by a guesser which provides potential part-of-speech categories based on affix patterns.
- Disambiguation is done with statistical methods (Hidden Markov Model), although we also experiment with fully rule-based methods.

Noun Phrase Extraction

For the purpose of terminology extraction from technical documents we designed a tool which applies finite-state techniques to mark potential terms, especially noun phrases corresponding to given regular patterns. The noun-phrase extraction tool consists of several modules: language independent programs (tokeniser, part-of-speech disambiguator, and noun phrase mark-up) and language dependant data (finite-state transducers and transition probabilities). This modular architecture allows rapid extension

to different languages. Currently, implementations for 8 languages (Dutch, English, French, German, Hungarian, Italian, Portuguese, Spanish) exist; more languages (e.g. Czech, Polish, Russian) are in preparation.

Noun phrase (NP) mark-up applies finite-state automata describing noun phrase patterns. These patterns rely on the simple (non-ambiguous) tagger output format, i.e. they consist of regular expressions on sequences of tokens and tags.

A very simple noun phrase description for a given language (e.g. French) may consist in a (possibly empty) sequence of adjectives followed by a noun and another sequence of adjectives. The automata which describe noun phrases are compiled into the final NP-mark-up. The compilation script uses the directed replace operation for the longest match and inserts brackets around maximal NPs (according to the NP patterns). The final NP-mark-up transducers are non-ambiguous, i.e. for every input they provide a single output containing non-recursive bracketing for NPs.

The following examples from the current realisations for French, Dutch and Spanish illustrate the application of the complete chain of tokenising, part-of-speech disambiguation and noun phrase mark-up:

> *Lorsqu'on tourne le commutateur de démarrage sur la position auxiliaire, l'aiguille retourne alors à zéro.*
>
> Lorsqu'/CONN on/PRON tourne/VERBP3SG le/DETSG
> **NP{commutateur/NOUNSG de/PREPDE démarrage/NOUNSG}**
> sur/PREP la/DETSG
> **NP{position/NOUNSG auxiliaire/ADJSG}**
> ,/CM l'/DETSG
> **NP{aiguille/NOUNSG}**
> retourne/VERBP3SG alors/ADV à/PREPA
> **NP{zéro/NOUNSG}**
>
> *De reparatie- en afstelprocedures zijn bedoeld ter ondersteuning voor zowel de volledig gediplomeerde monteur als de monteur met, minder ervaring.*
>
> De/ART **NP{reparatie-/CMPDPART en/CON afstelprocedures/NOUN}** zijn/VAFIN bedoeld/VVPP ter/PREP **NP{ondersteuning/NOUN}** voor/PREP zowel/CON de/ART **NP{volledig/ADJA gediplomeerde/ADJA monteur/NOUN}** als/PREP de/ART **NP{monteur/NOUN}** met/PREP minder/INDDET **NP{ervaring/NOUN}**
>
> *Para asegurar el funcionamiento óptimo de los vehículos, así como la seguridad personal del técnico, es imprescindible seguir los métodos apropiados de trabajo y los procedimientos correctos de reparación.*
>
> Para/PREP asegurar/VINF el/DETSG **NP{funcionamiento/NOUNSG óptimo/ADJSG de/PREP los/DETPL vehículos/NOUNPL}**,/COMA así~como/CONJ la/DETSG **NP{seguridad/NOUNSG personal/ADJSG del/PREPDET técnico/NOUNSG}** ,/COMA es/AUX imprescindible/ADJSG seguir/VINF los/DETPL **NP{métodos/NOUNPL apropiados/VPASTPARTPL de/PREP trabajo/NOUNSG}** y/CONJ los/DETPL

NP{procedimientos/NOUNPL correctos/ADJPL de/PREP reparación/NOUNSG}

Naturally, in a terminology management application, noun phrase extraction leads only to the selection of candidate terms. This automatic selection remains to be validated by human terminologists.

Additionally, by combining monolingual NP extraction as described above with alignment techniques based on statistical methods, one may extend the application to bilingual terminology extraction. Candidate terms are first extracted independently for language A and B. Aligned terms are then spotted by evaluating how often a given bilingual pair of terms (T_a, T_b) appears within aligned sentences. Again, in terminology management, bilingual extraction as well as alignment needs to be further validated by human specialists.

Incremental finite-state parsing

Finite State Parsing is an extension of finite state technology to the level of phrases and sentences.

Our work concentrates on shallow parsing of unrestricted texts. 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 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).

2. APPLICATIONS

2.1. 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. 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.

2.2. Multilingual Information Retrieval

Many of the linguistic tools being developed at our Centre are being used in applied research into multilingual information retrieval. 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 SMART[4]. 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.

2.3. 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.

---

[4] This software is available for research purposes at ftp://ftp.cs.cornell.edu/pub/smart.

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.

*2.4. Xerox Translation and Authoring Systems (XTRAS***)**

2.4.1. XTRAS Terminology Suite

2.4.1.1. 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.

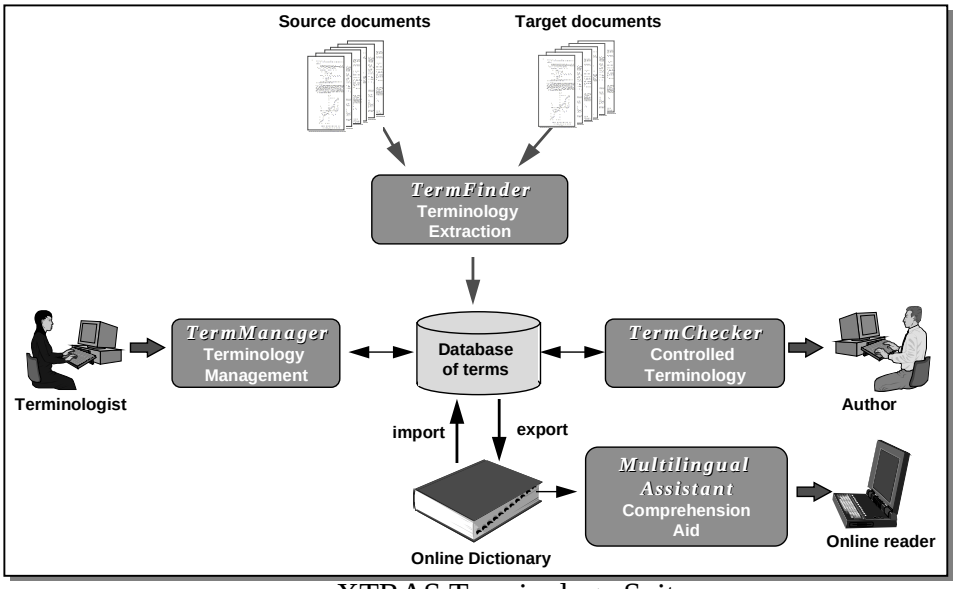
In addition, Danish, Swedish, Finnish, Norwegian, Czech, Hungarian, Russian, Romanian, Polish, Arabic, Japanese, Korean are under development.

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.

2.4.1.2. 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.



XTRAS Terminology Suite

2.4.1.3. 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.

2.4.1.4. Multilingual Assistant : Comprehension Aid Tool

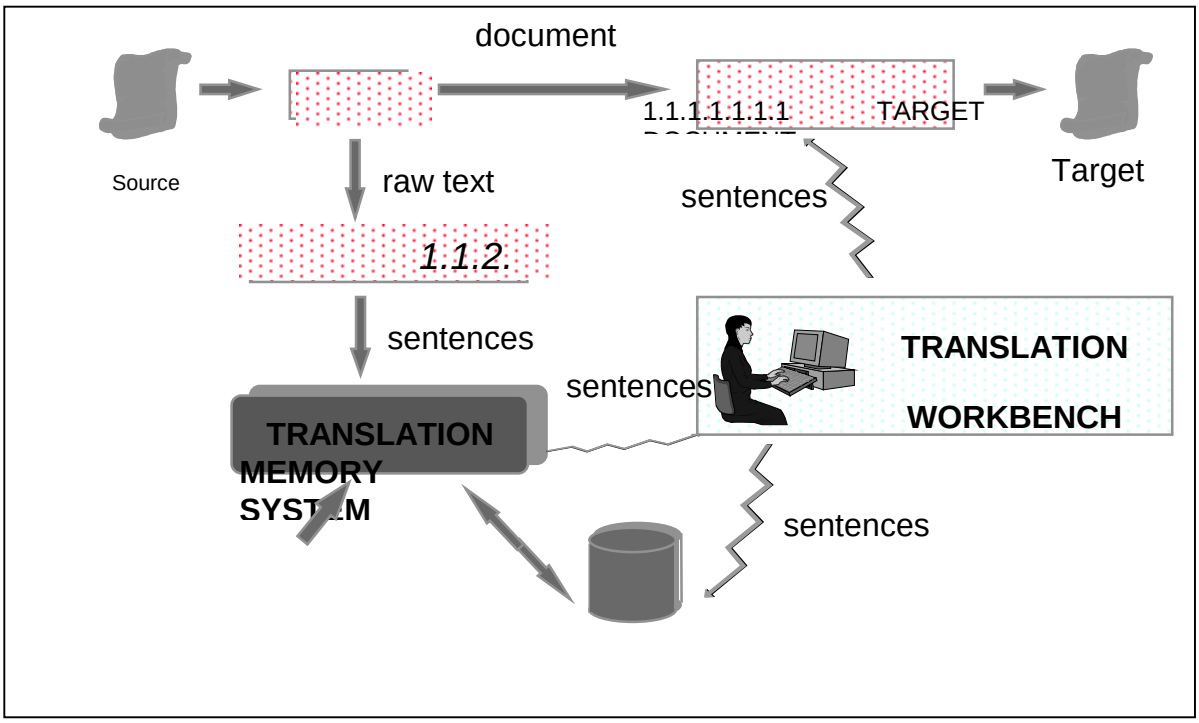
This Multilingual Assistant provides translation of words in context, using a general or specialised dictionary. It can differentiate between similar expressions that should be translated differently ("apply to" vs. "apply *something* to"). The Multilingual Assistant is based on the results of the Locolex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

2.4.2. XTRAS 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.

How does it work?



XTRAS Translation Memory Overview

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. 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 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. The database for storage is computationally efficient and can maintain a large amount of information using a minimum of resources. The database holds pairs of sentences, (source and target) containing the following history: the source of the sentence, the source and target languages of the sentence, the number of times the sentence occurs, when the sentence was written and by whom and the last time the sentence was accessed and by whom. The sentences will also carry their original format

As the storage facility can show these details, the project manager will have no trouble in editing and cleaning up texts.

- Import: Various data sources (text files) can be fed into the translation memory, including other translation memory systems for example Trados, IBM TM/2, bilingual dictionaries (by extracting translations).
- Export: The data from the translation memory can be moved to a file of text which contains aligned sentences and to documents using other translation memory systems.

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; secondary words (adverbs and adjectives) are different but the main words (verbs) are the same
- changes in the order of phrases
- changes in the order of words

Generation of Translation: If there is a difference between the match and the searched sentence, the aim is to find the closest possible target sentence and so minimise the work

of the translator. Translation memory can generate such a modified match if the difference is small, for example relating to punctuation, case or number.

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.

3. Selected references

Aït-Mokhtar, Salah and Chanod, Jean-Pierre (1997a): "Incremental finite-state parsing", in *Proceedings of Applied Natural Language Processing 1997*, Washington, DC.

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

Bauer, D., Segond, F. and Zaenen, A. (1995): "LOCOLEX: the translation rolls off your tongue." in *Proceedings of the ACH-ALLC conference*, Santa Barbara, pp. 6-8.

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

Grefenstette, Gregory (1994): *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Press, Boston.

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

Hladka, Barbara and Hajic, Jan (1997): "Probabilistic and Rule-based Tagger of an Inflective Language" In *Proceedings of Applied Natural Language Processing 1997* Washington, DC.

Kaplan, Ronald M. and Kay, Martin (1994): "Regular Models of Phonological Rule Systems". *Computational Linguistics*, 20:3 331-378.

Karttunen, Lauri (1994): "Constructing Lexical Transducers". In *Proceedings of the 15th International Conference on Computational Linguistics*, Coling, Kyoto, Japan.

Karttunen, Lauri (1995): "The Replace Operator". *In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics* (ACL-95) 16-23.

Koskenniemi, Kimmo (1983): "A General Computational Model for Word-Form Recognition and Production. Department of General Linguistics". University of Helsinki.

Kupiec, Julian and Wilkens, Mike (1994): *The dds tagger guide version 1.1*. Technical report, Xerox Palo Alto Research Center.

Maxwell, III, John T. and Kaplan, Ronald M. (1991): “A method for disjunctive constraint satisfaction.” In Tomita, Masaru (ed.), *Current Issues in Parsing Technology*. Kluwer Academic Publishers, Dordrecht, pp.173-190.

Nerbonne, John, Karttunen, Lauri, Paskaleva, Elena, Proszeky, Gabor and Roosmaa, Tiit (1997): “Reading more into Foreign Languages”. In *Proceedings of Applied Natural Language Processing* 1997 Washington, DC.

Schiller, Ann (1996): “Multilingual Finite-State Noun Phrase Extraction.” In: ECAI '96 workshop on "Extended finite state models of language", Budapest.

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