Machine Translation, Translation Memories and the Phrasal Lexicon: The Localisation Perspective

Reinhard Schäler

Abstract

Software localisation is the ideal application area for translation automation because of *what* is being translated, *how* it is being translated and *who* is involved in the translation. But existing translation tools and, more specifically, machine translation and translation memory technology cannot effectively deal with *all* the translation requirements of the localisation industry. This article examines these needs by highlighting characteristics of software localisation which have a direct impact on its translation requirements. It shows how and why localisation companies have used machine translation (MT) and translation memory (TM) based technology to address issues arising out of time and budget constraints. The article concludes with an assessment of the inherent limitations of both the MT and the TM approach and reports on research which could eventually lead to an enhancement of the currently available technology through the use of a Phrasal Lexicon.

Reinhard Schäler

Reinhard Schäler has been working with Irish and overseas companies in the area of software localisation since 1986. Together with Irish universities and industrial partners, he has organised several international conferences and workshops. He has given lectures on the subject at Irish universities, presented papers at international conferences on Natural Language Processing and has published a number of articles on NLP and software localisation.

He has been a researcher in the Department of Computer Science, University College Dublin, for a number of years and is the manager of the Localisation Resources Centre (UCD). He is a founder member of the Software Localisation Interest Group (SLIG) and its current chairperson.

The Localisation Resources Centre

Ireland is the number one location in Europe, and probably world-wide, for the localisation of software and its documentation. To further develop this important industry, the Localisation Resources Centre was established in December 1995 at UCD with support from the Irish Government and the European Regional Development Fund (ERDF). The Centre offers the following services to the localisation industry: Localisation Tools Library; Evaluation of Localisation Tools; Education and Training; Research and Development of Localisation Tools; Consultancy; Industry Watch; Regular Publications.

Mr Reinhard Schäler
Localisation Resources Centre
Roebuck Castle, UCD, Belfield
Dublin 4, Ireland

Tel: +353-1-2830644, Fax: +353-1-2830669
E-mail: Reinhard.Schaler@ucd.ie
Url: http://lrc.ucd.ie

1. Software Localisation

Approximately ten years ago, large North American software publishers, realising that international markets offered enormous potential for growth, decided to establish their first manufacturing sites in Europe. They also set up development teams to adapt the original US-English product to the requirements of European users. These European manufacturing and development sites were soon supported by a growing service industry offering, initially, translation and DTP services and, later on, complete 'turn-key' solutions. The whole of this new industry has become known as the Localisation Industry.

The process known as *software localisation* covers a wide range of activities, including:

- adapting and re-engineering of software according to the requirements of European and world markets;
- software testing (QA) of localised products;
- translating of documentation;
- "porting" of multimedia products into other languages and cultures, involving, among others, actors for voice-overs and graphic artists for the adaptation of the visual contents;
- printing of documentation;
- duplication of diskettes and CD-ROMs.

Despite its image as a high-tech industry, a great number of tasks in the localisation process are still carried out manually and are, as a result, very labour intensive and costly.

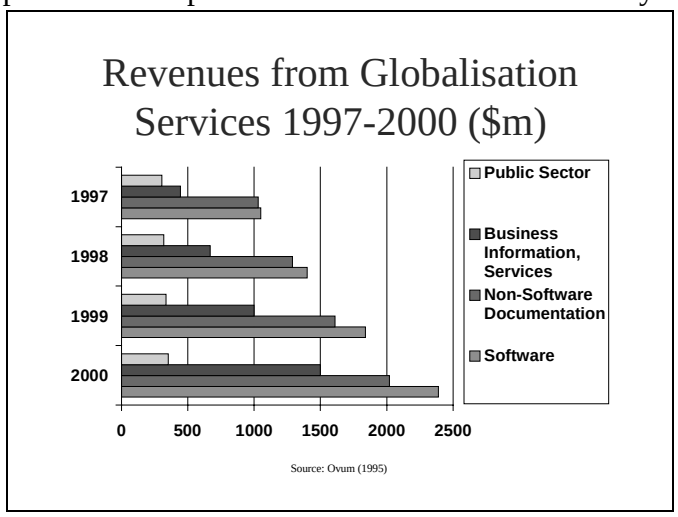
What makes software localisation different from other (even related) industries, e.g. in the translation or the I.T. sector, are its following characteristics:

1. It is an industry with phenomenal growth rates which will continue to expand into the next millennium.
2. Product life cycles are extremely short.
3. The source to be localised is never 'stable' (but always coming from the same restricted technical domain).
4. Localised products are supposed to become available at the same time as the base product.

1.1 A booming industry

The importance of a global perspective is no longer a disputed issue among the big North American software publishers. While the US internal market is stagnant and extremely competitive, Western Europe and especially the new markets in Eastern Europe and Asia are still 'underdeveloped'. Many publishers, therefore, are now investing considerable resources into their international business - an investment that is paying off: some of them already achieve more than half of their revenues from international markets. In this context, product internationalisation and localisation become more important than ever for company survival and growth.

In 1995, well known and respected British based marketing experts OVUM published a report on Globalisation in which they attempted to track the market for globalisation services and provide a forecast up to the year 2000. Among the four sectors surveyed were:

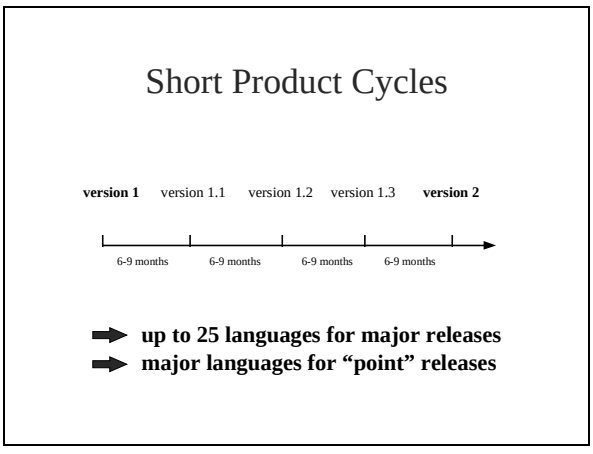


- Public Sector
- Business Information and Services
- Non-Software Documentation
- Software

According to their forecast, software localisation is one of the fastest growing industries world-wide and the most important sector within the globalisation-services market. The software sector will, at least over the coming 3 years, outperform all other sectors by more than doubling revenues between 1997 and 2000. In 1996, the world-wide translation products market was estimated to be worth around $300 million and growing by nearly 50% per year, reaching $1.5 billion in 2000.[1]

1.2 Short product cycles

Product life cycles have become shorter in recent years, with some software publishers only leaving 6-9 months between 'point releases' of their products. Not all releases of all products are always being localised for all languages. While big publishers, among them Microsoft, Lotus and Corel, localise major releases of the original product for approximately 20 languages, minor 'point releases' will only be published for major markets like Germany, France or Italy and, increasingly, some Asian markets including Japan.



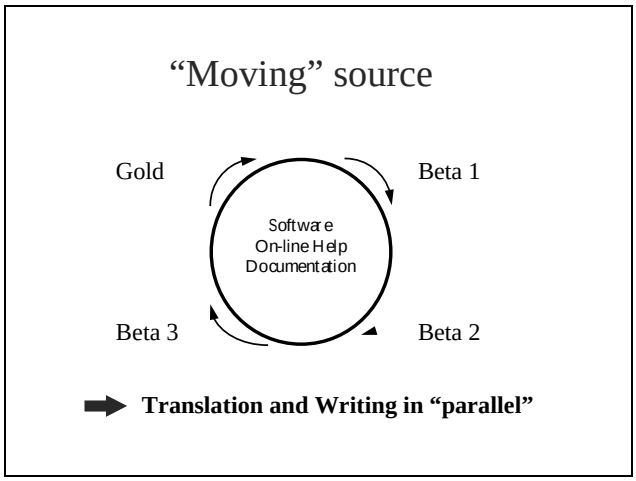
Obviously, the publishers would prefer to service all markets with localised versions of their current

[1] Ireland is now the number one location in Europe, and probably world-wide, for the localisation of software and its documentation. Currently, 40-50% of the PC based software sold in Europe originates in Ireland. It is expected that this figure will rise to 60% over the coming years. In Ireland, the localisation industry accounts for around 2 billion pounds worth in exports. There are approximately 4,000 people employed directly in the translation, engineering and manufacturing of localised products. It is estimated that for every person directly employed two others are employed in dependent industries. (cf. Murphy 1994)

products - not least because users constantly deprived of up-to-date software from one publisher will soon turn to other publishers with a more progressive update policy. The high costs currently involved in updating products, however, make this impossible for less significant markets. Software publishers are, therefore, actively searching for methods and tools which would allow them to update their current localised products with each update of the original product at minimum expense.

1.3 "Moving source"

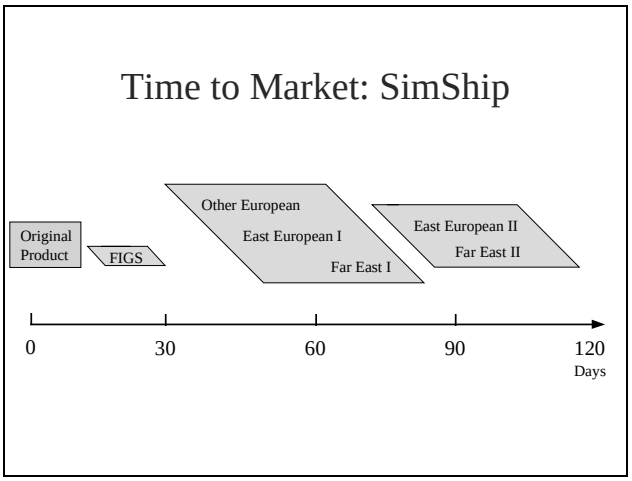
Only in exceptional circumstances can localisers work on a stable, final version of a system (including software, on-line help and printed documentation). Localisation teams are involved from the very beginning in a development project by providing feed-back on the implementation of international features into the base product and by identifying those areas which will need to be localised, i.e. text, templates, defaults etc.



The first translation pass is usually performed on late alpha or very early beta builds of products, at a stage when implementation problems regarding international features can still be addressed by the development team. Further builds are analysed as they come in and necessary modifications are implemented with each new beta build. This way the first localised versions of a base product are usually only very few days or weeks behind.

1.4 SimShip strategy

Simultaneous shipment ('SimShip') of the original and the localised version of an application became an issue in the early 1990s and was probably achieved for the first time with the German and French versions of 1-2-3 Release 2 by Lotus Development Ireland. The very costly exercise of simultaneous shipment was initially justified by the assumption that the company would loose revenue and valuable market share with every day that a product - although available in English - could not be sold in its localised version.



In 1996, 'SimShip' does not necessarily mean anymore that the localised product has to be shipped on exactly the same day as the original US version. Over the past few years, most companies have found that the enormous costs associated with the 'SimShip' strategy could not be justified. While they still aim for short 'deltas', i.e. the period between shipment of the original and the localised versions, 'SimShip' has now been redefined as shipping within the same financial quarter.

2. Translation and Software Localisation

The translation of the original software products, often into some 20 languages, is one of the key tasks in the localisation process (and still the single most expensive one). There are basically two types of 'texts' that have to be translated: the *software* itself (menu commands, dialogs, error messages etc.) and the *documentation* (on-line help and printed user manuals).

The translation of the *software* is technically quite challenging and presents translators with a number of issues that go far beyond those traditionally associated with translation. The nature of these issues depends largely on the design of the original software and varies according to the target language. Most software publishers have over the years developed their own proprietary in-house translation tools to help translators deal with these issues.[2]

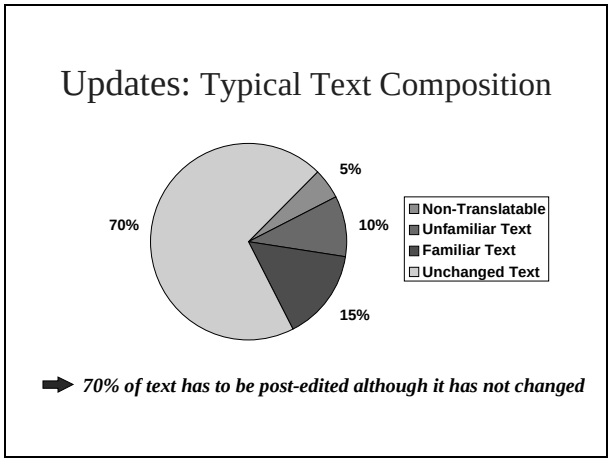
In terms of quantity, however, it is the *user documentation*, i.e. the on-line help system and the printed user manuals, which presents the biggest challenges. Projects of up to 1 million words cannot be managed anymore by individual translators. They require well managed and experienced teams able to guarantee a short turn around time while maintaining the required level of linguistic quality, the most important of which is consistency of style and terminology across:

- different areas of the documentation
- software and documentation
- different releases of the same product
- the same product family

The text to be translated is mainly composed of explanations and instructions outlining the features of an application and providing guidelines on how to use them,

[2] As Birch (1993) points out, software products should be designed from the outset with the multinational market in mind in order to achieve lower costs for localisation, faster time to market and higher customer satisfaction. He lists a number of potential issues associated with the translatability of user interfaces and provides examples for each of these to illustrate his points.

i.e. translators are generally dealing with a text of a consistent style and from a well defined and restricted domain. This requires, on the one hand, a certain period of familiarisation with the style and terminology - on the other hand it enables experienced translators who have become familiar with the product and the terminology to turn-around translations in sometimes amazing time frames.



As the software itself, the text to be translated very rarely is completely new. In fact, in very many cases it is updates of previous versions of the programme which are being translated and a high proportion of the text has already been translated and edited before. In a typical update situation up to 70% of the text has not changed at all, i.e. its original (or 'old') translation can be 'copied' straight into the new target text.[3] A further 15% is familiar text which has been modified but is not completely new, i.e. its translation has to be adapted to reflect the changes made in the source language.[4] Only 10% of the text is completely new or 'unfamiliar' while 5% is text which for various reasons should not be translated.

It is not unusual, therefore, that translators are remain with a product over a number of years and translate all of its regular updates.[5] Their translations almost become identified with the product - not unlike the identification of certain actor's voices with well-known American actors when dubbed.

Software localisation is one of the few industries which offers translators an acceptable level of income while allowing professionals to specialise exclusively in this area.[6] This, however, requires translators to continuously update their knowledge and renew their equipment as they are, in effect, always working on tomorrow's software products. Professional translators in the localisation business see the need to keep up with modern technology not as a burden but as a welcome, interesting ingredient adding to the diversity of an otherwise quite dull and often routine job - far removed from the intellectual challenges and creative liberties aspired to during their college education.

All the above make translation in software localisation an ideal application for automatic translation: machine translation (MT), translation memory (TM) based systems and other computer assisted translation (CAT) systems such as on-line terminology databases etc. Translators, agencies and publishers - because of the very nature of the business they are in - are all open to experimenting with new

[3] These are the *Exact Matches* in translation memory speak.
[4] These are the *Fuzzy Matches* in translation memory speak.
[5] All major software applications are being updated at regular intervals, often not exceeding 12 months.
[6] In countries like Ireland the software localisation industry is in fact the only industry which can provide professional translators who wish to specialise in one area of expertise with an acceptable level of income.

technology, they are already equipped with the necessary hardware to run it and - last but not least - they want to be seen by the industry as operating on the leading edge of technology, as innovators in their field.

It is not surprising, therefore, that in the past the software localisation industry has probably been the biggest single user of translation technology and that its innovative, dynamic and technology driven approach has already contributed to the development and opening-up of markets for new language technology applications.

2.1 Machine Translation in Localisation

Issues related to increased competition, mounting pressure of bringing down the cost of translation (the single biggest cost in localisation), and the growing need to translate into more languages without a corresponding increase in the budget available, have prompted localisers since the early 90s to invest considerable amounts of money into either the implementation of MT systems or at least into long term evaluation and feasibility studies. By doing so, they have subscribed to the widely held view that MT had reached a level of maturity that would allow its successful operation in a commercial translation environment.

They were encouraged by reports from current and new users of MT technology within the localisation industry which had achieved significant savings in translation cost and a considerable increase in translation quality through the commercial use of available MT systems.

Success Stories

Daniel Grasmick from *SAP*, Europe's biggest indigenous software developer, reported at the MT Summit V in Luxembourg (July 1995) on his company's integration of MT into the translation process. He pointed out that at SAP

- MT is totally accepted by human translators;
- MT produces high quality translations at high speed;
- The operation has an excellent profitability rate;
- Its 13 employees (8 of whom are full-time) translated 3m words in 1994 and 1/2m words per month in 1995.

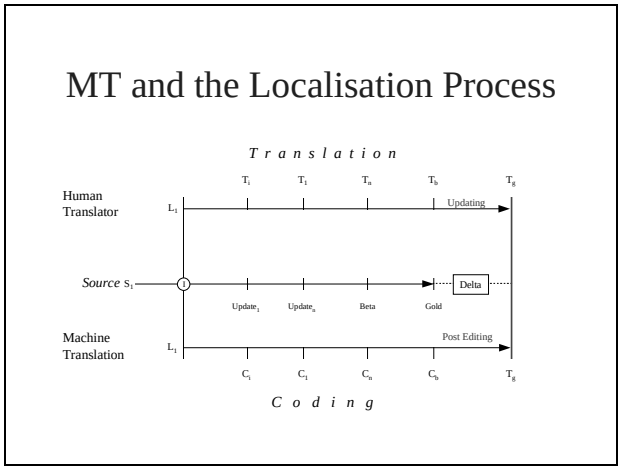
Gary Jaekel from *Ericsson Language Services* reported at the Localisation Industry Standards Association (LISA) Forum in Amsterdam (April 1995) that they were a user of MT since 1993. He pointed out that:

- By 1995, they had achieved a 50% productivity increase
- Experienced translators at Ericsson can produce 12 pages per day after working 3 months in production mode. Using MT on certain texts, the same work can now be done (on average) in an hour and a half.
- By 1995, they had achieved 100% increase in quality
  - permitted before (per 100 pages) : 16 minor, 1 major faults
  - achieved after (per 750 pages): no major, 4 minor faults

The problems which arose in connection with the introduction of MT at Ericsson were not, as originally expected, related to the quality of the translations but to the

(non-linguistic) limitations of the MT system and the format of documentation processed as well as the internal setup of the operation.

Among other localisation companies actively using or at present considering the use of MT technology are Berlitz, Corel, Gecap, Idoc, Lotus/IBM, and Oracle. To date, none of them have reported success stories similar to SAP and Ericsson. What they found was that if financial savings could, in fact, be achieved at all, they could probably only be expected after 2-3 years. Given the characteristics of the localisation process, an MT supported localisation project was also not likely to ship earlier than one run with human translators.



Their experience would indicate that, while *human* translators used the crucial period between the 'gold' date for the original source and the 'gold' date for the localised version to incorporate the final changes made to the last beta build in their translations, the translators working on the *MT* supported project spent that time post-editing the machine translation output based on the source gold build.

Other limitations of the (exclusive) MT approach encountered in the context of localisation are:

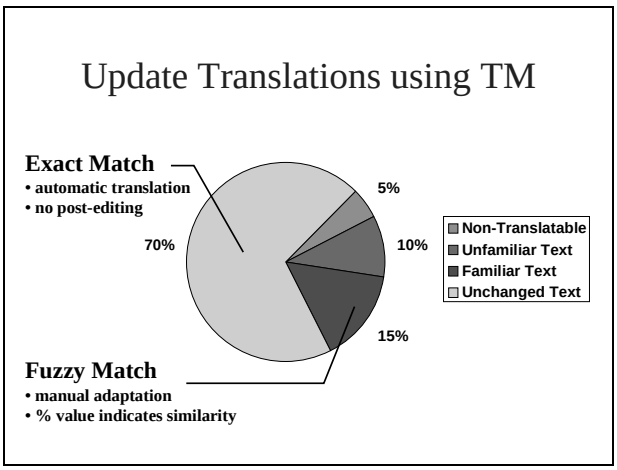
- *Translation Scenario*: Translations are often performed by freelance translators working from home in geographically diverse global locations without easy access to the central MT coding facilities.
- *Non-uniformity of material*: Many MT systems were not able to deal with the variety of text formats encountered in localisation. MT systems, for the most part, were also not able to re-use existing language resources (word lists etc.) and required an investment in development and maintenance that was difficult to justify.
- *Post-editing work is always lost*: While the performance of MT systems increases over time with the development of lexical and semantic dictionaries, output from current systems requires considerable post-editing to reach publishable standard. This necessary work is always lost - even if the text of the original has not changed. In other words, MT does not take advantage of the highly repetitive nature of the source text, it does not re-use the translation of text that has remained unchanged between the previous version and the current one.

2.2 Translation Memory Based Tools

Approximately 4 years ago, in 1992, the first translation memory (TM) based tools became commercially available. The basic idea behind these tools is simple: *Never translate or edit the same translation unit twice*. Translation memories are basically a

database of aligned sentence pairs, one being the translation of the other.[7] From a technical point of view they are straight-forward, easy to use and maintain, and provide a high degree of ergonomy and portability. Financially, they do not require a substantial investment: the tools themselves are affordable with a list price of around 2,000 dollars[8]; they do not require special hardware equipment and require only limited training of translators and CAT administrators.

Translation Memory (TM) based tools do not require coding as MT systems do. TMs are created either during the translation process by translators as they enter the translation or by 'aligning' previously translated files using so called alignment utilities. When a newer version of a previously translated file has to be translated, the system automatically offers or even inserts the translation of all the sentences[9] in the new file which were translated previously (and which have not changed), i.e. the system identifies 'exact matches'. If no exact matches can be found, the system attempts to find similar sentences in the TM to those in the new source file. If it finds similar sentences it displays them and the original source as so called 'fuzzy matches'. All the translator now has to do is to adapt the old translation to reflect the new source.[10]



In the context of localisation this means that in the typical update situation (be it between different beta builds or new releases of a product), around 70% of the text can be translated automatically using text segments previously translated and made available automatically through the translation memory.[11]

Although TM systems do not provide a practical solution for all translation problems in localisation they are widely seen as a very pragmatic solution - the famous *80/20 solution* - potentially saving time and money, and providing a higher degree of consistency while they are, at the same time, easy to use, PC based, integrated in popular word processing environments and, above all, affordable.

---

[7] Recent TM based systems allow the creation and use of multilingual translation memories.
[8] Substantial discounts are offered to users purchasing a number of licences at a time.
[9] Although the TM systems usually refer to 'segments' rather than 'sentences' to make clear that they do not just segment *text* into 'sentences' but also titles, headers etc., we will use the term 'sentence' for the purpose of this article.
[10] With recent systems, the degree of similarity between the old and the new source segment which make it an admissible 'fuzzy match' can be defined by the translator.
[11] Text formatting information preserved as TM systems offer special filters for the most widely used word processors and desk top publishing systems.

However, there is one basic short-coming which results from the design of the TM based systems and the approach chosen by their developers: The basic translation unit in TM based systems are sentences ('segments'). Phrases which have been translated before are not recognized as exact matches but, and even then only in a 'best-case' scenario, as fuzzy matches:

Consider the following example:

TM Entry I

[ENG] The bullets move to the new paragraph.
[GER] Die Blickfangpunkte rücken in den neuen Abschnitt.

TM Entry II

[ENG] The title moves to the center of the slide.
[GER] Der Titel rückt in die Mitte des Dias.

New sentence

The bullets move to the center of the slide.

Although the two phrases, *The bullets move* and *to the center of the slide*, had already been translated before TM systems cannot combine phrases from different entries in the translation memory to form a new phrase, nor can they identify the 'best match', in this case probably *The title moves to the center of the slide*, and substitute the changed segment, *The title moves*, with the correct new phrase, *The bullets move*, to produce the required translation:

Die Blickfangpunkte rücken in die Mitte des Dias.

To solve this problem, translation memory systems need to be 'linguistically enhanced' using techniques which are based on those developed in the context of Example Based Machine Translation (EBMT).

3. Extending MT and CAT: The Phrasal Lexicon

In a basic implementation of Example Based Machine Translation (EBMT) a bi- (or multi-) lingual corpus of aligned source and target segments (not necessarily sentences) provides examples which during the translation are matched against the new source (analysis). This is done using sophisticated algorithms which measure the distance or similarity between the examples and the new source. In a second step the examples identified as appropriate are combined (transfer) and the new target is generated (synthesis). One of the best known representatives of the EBMT approach, Nagao, defined the concept in 1984 as that of implementing the human learner's technique of using examples as a guide to translation.

> *Man does not translate a simple sentence by doing deep linguistic analysis, rather, man does translation, first, by properly decomposing an input sentence into certain fragmental phrases (...),*
>
> *then, by translating these fragmental phrases into other language phrases, and finally by properly composing these fragmental translations into one long sentence.*

Almost 10 years before Nagao, Joseph Becker proposed the concept of the *Phrasal Lexicon* at the 1975 TINLAP conference.

> *I suspect that we speak mostly by stitching together swatches of text that we have heard before; productive processes have the secondary role of adapting the old phrases to the new situation.*

The basic assumptions of the *Phrasal Lexicon* are:

- The use of language is based at least as much on ***memorisation*** as on any impromptu problem solving.
- The process of language production is ***compositional***.
- The phrasal lexicon provides the ***patterns*** that can provide major elements of "new" expressions.

There are, as far as we could establish, few implementations of this concept. Sato and Nagao developed an experimental system and the Dutch company DLT developed what became known as the Bilingual Knowledge Bank (BKB) which is an aligned corpus of equivalent texts in two languages, structurally analysed by the same type of parser into translation units. There is also a research project at the Localisation Resources Centre (University College Dublin) and the University of Manchester Institute of Science and Technology (UMIST) which is exploring the application of modern computational linguistic techniques to this approach.[12] This project has developed a basic research prototype based on Becker's idea about the phrasal lexicon. The aim of this project is to establish whether the approach could, if applied to the restricted domain of software localisation, overcome the inherent limitations of translation memories as outlined above.

The experiments with the phrasal lexicon carried out at the Localisation Resources Centre (University College Dublin) and the University of Manchester Institute of Science and Technology (UMIST) cover:

- The production of an aligned, bilingual *phrasal lexicon* (with linguistic analysis)
- The use of *one parser/grammar* for English and German to analyse source/target
- The "translation" of new source text by *combining/substituting* known phrases
- The use of a *dictionary* as an index into Phrasal Lexicon

These experiments will test two conceptually different strategies of re-using previously translated phrases for new translations:

(I) *Tree combining* (identify matches for phrases in new source in phrasal lexicon; combining known phrases to reflect the new source). It is expected that this approach will be computationally 'cheap' but that it will require a large phrasal lexicon to produce satisfactory results.

(ii) *Tree substitution* (identify 'best match' for new source in phrasal lexicon; substituting modified phrase(s) in the new sentence with phrases from the phrasal lexicon). Our expectations here are that this approach will be computationally

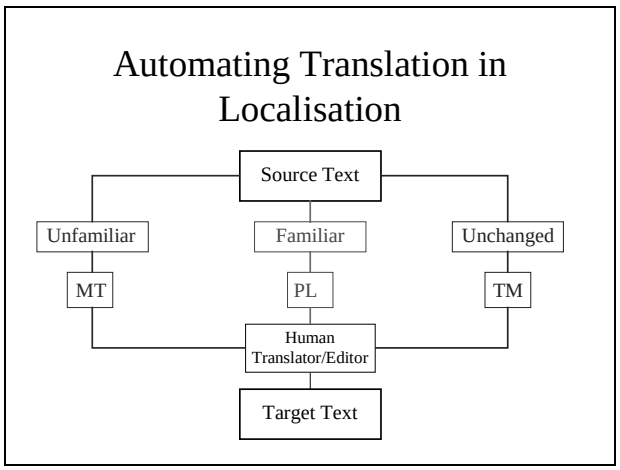
[12] This research was initiated by Prof. Alan Ramsay in the Department of Computer Science (UCD). Prof. Ramsay is now at UMIST.

'expensive' but will only require a smaller phrasal lexicon to produce satisfactory results.

The issues related to the development of the phrasal memory approach so far have been:

- Large development overheads (grammar, parser, lexicon)
- Large processing overheads (parsing)
- Large storage overheads (structural representation)

Although the initial experiments with the 'new' phrasal lexicon approach have not been concluded and, therefore, not yet produced results which could justify us to jump to premature conclusions, we feel that the phrasal lexicon could be the solution to the translation of text which with current TM based systems can - at best - only be identified as a fuzzy match.



The phrasal lexicon has the potential to offer translators all the advantages of the translation memory (speed, consistency, cost savings - no coding or post-editing as with MT systems) and more because it is not restricted to the matches found at sentence level but includes matches identified at the lower phrase level - a strategy which will yield a higher percentage of exact matches as it is dealing with smaller translation units.

4. Conclusion

Translation in localisation is without a doubt one of the most suitable application areas for MT and CAT. Whether translation technology products will be used more extensively in the localisation industry depends on its economic viability, in other words: does the use of automatic translation systems bring down the cost of translation? Although there are some case studies available, some of which we briefly discussed, the question has yet to be answered.

However, there seems to be a consensus in the industry which suggests that, under average circumstances, stand-alone MT systems are not commercially viable. According to this view, only when combined with translation memory technology which overcomes some of its limitations can MT become useful. There are already attempts by some developers to offer this solution which has been received with great interest, although the translation and localisation agencies currently prefer to work with TM systems only, as they assess the implications and opportunities arising from a possible future integration of MT systems.

While definitely not yet offering an alternative to already available technology in a commercial translation environment, research into the phrasal memory approach will continue with the aim of assessing the potential of this technology to further enhance currently used MT and TM technology.

5. References

[Becker, 1975] Joseph D. Becker, Bolt Beranek and Newman. The Phrasal Lexicon. *Proceedings of Theoretical Issues in Natural Language Processing*, pages 70-73. Cambridge, Massachusetts (10-13 June 1975).

[Birch, 1993] Richard E. Birch. Making user interfaces translatable. *Proceedings of the First Irish Conference on Language Technology*. Dublin (12 May 1993.)

[Grasmick, 1995] Daniel Grasmick. 2 MT Systems and still hungry. *Proceedings MT Summit V*. Luxembourg (10-13 July 1995).

[Jaekel, 1995] Gary Jaekel. Machine Translation and being business-like. *Proceedings LISA Forum (Annual Meeting).* Amsterdam (3-5 April 1995).

[Murphy 1994] Barry Murphy, National Software Director, Irish National Software Directorate, Forbairt (the Irish agency for science and industrial development), private communication included in: *Proposal for the Establishment of the Localisation Resources Centre*, Dublin 1994 (not published).

[Nagao, 1984], M. Nagao. A framework of a mechanical translation between Japanese and English by analogy principle. In: A. Elithorn and R. Banerji, editors, *Artificial and Human Intelligence*, pages 173-180. North Hollan, Amsterdam, 1984.

[Ovum 1995a] Ovum. *Globalisation - Creating New Markets with Translation Technology* (by Rose Lockwood, Jean Leston, Laurent Lachal). Ovum, London, 1995.

[Ovum 1995b] Ovum. *Translation Technology Products* (by Jane Mason and Adriane Rinsche). Ovum, London, 1995.

[Sadler, 1989] V. Sadler. (1989) *Working with analogical semantics: disambiguation techniques in DLT*. Distributed Language Translation, 5. Foris Pub., Dordrecht, 1989.