[*Translating and the Computer 27*, November 2005 [London: Aslib, 2005]

# Bringing Intelligence to Translation Memory Technology

Mickel Grönroos
Ari Becks
Master's Innovations Ltd
Tekniikantie 14
FIN-02150 Espoo, Finland

mickel.gronroos@masterin.com
ari.becks@masterin.com

## *Abstract*

In this paper we argue that Translation Intelligence is the next generation of translation memory technology that supersedes current state-of-the-art translation memory systems, as it is based on real self-learning and intelligent reuse of human translation knowledge, instead of simply attempting to recycle strings of characters, as traditional systems do.

We maintain that Translation Intelligence is the only cost-effective method for professional computer-aided translation that is usable by both professional translators and other professionals with frequent translation needs besides their main tasks. It has been shown that due to drawbacks in their techniques, current translation memories have only been able to reach a mere fragment of these wide customer groups.

## *1. Translation Market Today: Some Figures*

Estimates on the size of today's worldwide translation and localization services market vary from USD 4 billion to 15 billion[1], with the US and Europe as the main areas of business (approximately 40 % market share each)[2].

However, the market size of computer-aided translation (CAT) tools, especially translation memories, is noticeably lower. Estimates vary considerably, from USD 22 million[3] to 700 million[4], but nevertheless, the figures illustrate that the lion's share of translations is still done without any real computerized means.

As the translation need is expected to grow steadily, there is gigantic market potential for the provider of the *right* kind of translation tool that can be taken into use quickly, with very little tailoring or none at all.

Clearly, today's providers of traditional translation memory systems have not been able to meet this need that professional translators and other professionals translating beside their main work tasks have. These user groups are big, in Finland alone there

[1] According to the Localization Industry Primer (2003, 2nd Edition) by LISA, www.lisa.org
[2] According to Common Sense Advisory (2005), www.commonsenseadvisory.com
[3] According to MultiCorpora R&D Inc. (2002), www.multicorpora.com
[4] According to Globalization Insider XII/1.5 (2003), www.localization.org

are 4,000 professional translators, and the number of other professionals with translation needs can be counted in the tens of thousands.

## *2. The Translation Memory Pitfall*

Traditional translation memories are based on the presupposition that sentences recur from one text to another, either as such or with slight variation (in a mathematical, or fuzzy, sense). As this might be true for repetitive texts—such as new versions of previously translated documents—the statement does not hold for new, unrestricted texts where sentence repetition is in fact as low as 1 %[5].

Consequently, sentence repetition is the biggest obstacle for translation memories to reach new customer groups. It is quite frankly no surprise that only a fraction of the global, yearly translation volume is produced with the aid of translation memories. Clearly, a technology which actually only recycles sentences rules out all those customers who work with new texts of new fields from day to day or who produce so small translation volumes each year that the long-term benefit of a translation memory database is overshadowed by the cost of taking such a tool into use. But this user group—professionals translating beside their main work tasks—is enormous. It includes communication officers, secretaries, law firms, bankers, marketing experts, technical writers etc.

There is obviously a niche for a CAT tool that is easy and fast to take into use regardless of the text type to be translated.

## 3. *Moving From Translation Memories to Translation Intelligence*

With a view to overcome the drawbacks of traditional translation memories and create a translation tool suitable for both translation agencies and translators in general, the Finland-based translation technology company Master's Innovations Ltd invented a completely new method for computer-aided translation: **Translation Intelligence.**

In contrast to traditional translation memories, tools based on Translation Intelligence can be used for translating different types of both repetitive and less repetitive texts, and the time-to-market is up to ten times shorter, thanks to the unparalleled self-learning capability of the technology.

### 3.1. Flexible Segmenting vs. Static Segmenting

Translation Intelligence introduces *flexible segmenting* as opposed to the static sentence segmenting conducted by traditional translation memories.

By using artificial intelligence and previous human translation knowledge, a tool using Translation Intelligence will segment the source sentence at hand into smaller parts and translate these parts in turns instead of the full sentence in one go.

By operating on the sub-sentence level, where there is in all text types much more repetition than on the sentence level, Translation Intelligence is guaranteed to

[5] According to proprietary research conducted by Master's Innovations Ltd studying 10,000 Finnish newspaper sentences.

outperform traditional methods used in translation memories and lead to significant translation cost reductions.

## 3.2. Three Strategies Ensure Better Recall

Translation Intelligence features **three different strategies** when suggesting translations of the flexibly-sized segments. The effect is that a translation tool based on Translation Intelligence is able to provide the user with a translation suggestion in 99 % of the cases, even when faced with a completely new text.

The primary translation strategy is **translation recycling,** i.e. the flexibly-sized segments and their human-made translations are just reused. This is what a traditional translation memory system would also do.

> **Example on translation recycling:**
> If a human-made translation for "the issue is not discussed" already exists, and the same segment is to be translated again, the system will primarily reuse the existing translation.

The secondary strategy is **translation generation,** i.e. the system tries to reuse the flexibly-sized segments and their human-made translations as grammatical *translation patterns* whenever possible and generate translations based on such a pre-translated example that has a similar grammatical structure. If several equally suitable grammatical patterns exist, the system picks the best match, primarily using semantics and secondarily on the basis of use frequency or domain information. Translation Intelligence uses its built-in lexicon and its word-form generator to generate a correct translation suggestion in the target language.

> **Example on translation generation:**
> When aiming at translating "the house is not sold", the system will recognize that the grammatical pattern of this segment is similar to the pattern of the previously translated "the issue is not discussed" ("the" + "Noun-Nominative-Singular" + "is not" + "Verb-Past Participle").
> Therefore the system will use the human-made translation of "the issue is not discussed" to generate a translation for "the house is not sold", using the same target language pattern. It will also be able to translate correctly "the boy is not bullied", "the car is not stolen", and countless other similar phrases. For a traditional TM tool to reach the same level of coverage, each and every surface level clause would need to be stored separately.

Translation generation gives tremendous potential to a tool using Translation Intelligence; it does not merely reuse surface level strings, but actually learns logical or grammatical translation patterns from the user. Where a traditional translation memory would need to store every surface level segment with its translation separately in its database, a tool using Translation Intelligence needs only one translation pattern in its *Knowledge Base* to be able to translate innumerable similarly translatable segments! A rough estimate is that the same translation coverage can be achieved with a Knowledge Base of 50,000 translation patterns as with a conventional translation memory database of 1,000,000 translation units.

The last strategy applied is **translation heuristics,** during which a pure machine translation component takes over. This strategy is used when the memory-based strategies are unsuccessful, and it simply ensures wide translation coverage. The user always gets some translation suggestion even if it will require manual editing.

### 3.3. Initial Phase Is Cut Thanks to a Standard Knowledge Base Delivered to All

As Translation Intelligence uses flexible segmenting and handles translation units as grammatical translation patterns, the knowledge base used by the system is not as text-type dependent as a translation memory database is.

> **Example of text-type independence:**
> The translation pattern “turn” + “off”+ “the” + “Noun-Nominative-Singular” can be found in many different kinds of texts, but only in the manual of a kitchen appliance will you find a full sentence like “Turn off the dishwasher before opening the hatch”.

This means that a ready-made knowledge base can be produced, delivered to and used by all customers, largely regardless of what types of texts they translate. The gain here is that the annoying and labour-intensive initial phase of the tool is cut down to one tenth, as the user is provided with some high-frequent translation knowledge to start with.

### 3.4. Learning a Great Deal More

To sum up, a translation tool using Translation Intelligence learns from the human translator at an incredible pace as opposed to translation memories that merely recycle static sentences. This adaptiveness means that a tool using Translation Intelligence starts speeding up the translation process and cutting costs much faster than do traditional translation memory programs. We are talking months, not years!

Translation Intelligence currently supports translation between Finnish and English in both directions. A Finnish-Swedish-Finnish version is well under way, and Master's Innovations Ltd has the competence to develop translation support for the main European languages, if needs be.

## *4. Proving the Claim*

In this paper we have claimed that a CAT tool based on Translation Intelligence is faster to take into use than a traditional translation memory system, thanks to the preinstalled Knowledge Base and the unparalleled self-learning capabilities of the technology. We have argued that Translation Intelligence requires remarkably little domain-specific data to adapt itself to the customer’s use of language and way to translate. This suggests that Translation Intelligence is suitable not only for translation professionals, but also for other professionals with *occasional* translation needs.

### 4.1. Translating Recipes from Finnish to English

To prove our case, we conducted a translation test in which a translator used Master Translator Pro (MTP). Developed by Master’s Innovations Ltd, this interactive end-user CAT tool integrates with Microsoft Word and is based on the Translation Intelligence technology. The text to be translated from Finnish into English was a 10-

page compilation of recipes downloaded from various sites in the Internet, and consequently written by several people. The recipes more difficult to translate were placed towards the end of the document, and the simple ones at the beginning.

It is worth emphasizing that before the test, the MTP program had not been used to translate any food-related texts whatsoever. In other words, the system was neither tailored in advance, nor had it had the chance to adapt itself to translating cookery texts. All translation suggestions provided by the system were based on general translation knowledge available in the preinstalled Knowledge Base that is delivered to all customers.

The test data consisted of 10 pages that comprised 488 sentences or sentence-like units (e.g. headings and items in lists of ingredients such as “2 tbsp of brown sugar”). The total amount of Finnish words was 2,562 and the character count (spaces excluded) 17,122 characters. The mean length of a sentence or sentence-like unit was 5.25 words. The sentences were thus rather short. This is explained by the frequent occurrence of lists of ingredients, where each item in such a list would typically constitute a sentence-like unit. (Example 1 below shows the first recipe from the test text both in its original Finnish form and as the English translation done with MTP.)

In the test we wanted to study the following:

1. How well is the flexible segmenter of Translation Intelligence able to divide source sentences of previously unseen text into translatable units, i.e. to what extent are the suggested flexibly-sized segments such units of language that it would be meaningful to translate them as such without adjusting the segment length manually.
2. In what proportions and how well are the three different translation strategies—recycling, generation and heuristics—used when the program suggests translations of previously unseen text.
3. How many new translation patterns and domain-specific terms does the system learn by translating only 10 pages of text from a new field.

| **Ryppyperunat** | **Wrinkled potatoes** |
|---|---|
| 2 kg pieniä perunoita<br>4 ruokalusikallista karkeaa suolaa<br>vetta | 2 kg small potatoes<br>4 tablespoons coarse salt<br>Water |
| Pese perunat perusteellisesti mutta älä kuori niitä. Laita perunat kattilaan, lisää kylmää vettä, niin että perunat juuri ja juuri peittyvät. Lisää karkea suola. Keitä noin 20 min. Kaada suurin osa vedestä pois, jätä kattilaan noin sentin verran vettä, anna kiehua kunnes kaikki vesi on haihtunut ja perunat ovat kuivia. Ravistele kattilaa koko haihtumisen ajan. | Wash the potatoes thoroughly but do not peel them. Put the potatoes in a kettle, add cold water so that the potatoes are only just covered. Add the coarse salt. Let boil for approximately 20 min. Pour out most of the water, leave approximately one centimetre of water in the kettle, let it boil until all water has evaporated and the potatoes are dry. Keep shaking the kettle when boiling down the water. |

**Example 1:** An example of the potpourri of recipes translated during the test.

## 4.2. Flexible Segmenting Put to a Test

The average amount of suggested segments per page was 124 (total amount 1,236). On an average, each source language sentence or sentence-like unit was thus divided into 2.5 flexible segments and the typical length of a flexibly-sized segment was 2.1 words.

Four out of five suggested segments (78.51%) were such that the translator could accept them as meaningfully translatable segments, i.e. segments that could be translated individually without compromising the high-quality translation of the full sentence (see figure 1). Even though this remarkably high figure can to some extent be explained by the relative simplicity and regularity of the clauses of this text type (e.g. consider the clauses “whisk the eggs lightly”, “add the sugar”, and “fry the onions until brown”), it is nevertheless clear that flexible segmenting based on previous segmentation knowledge is a powerful way to improve translation coverage when translating new texts as opposed to the static sentence segmentation of traditional TM systems.

So only one out of five suggested segments (21.49 %) was such that the translator needed to change the length of the segment before starting to translate it. Typically, the segment was extended at the end by including a few more words to get a meaningful unit to be translated, often a noun phrase or a verb phrase.

A slight increase in segmentation quality was also reported during the test (see figure 1). Whereas 22.17 % of the segments on the first five pages needed resizing, the percentage of resized segments on the last five pages of the test data were down to 20.82 %. This seems to be the trend, but the data set used is, however, too small for us
to draw any valid statistical conclusions.

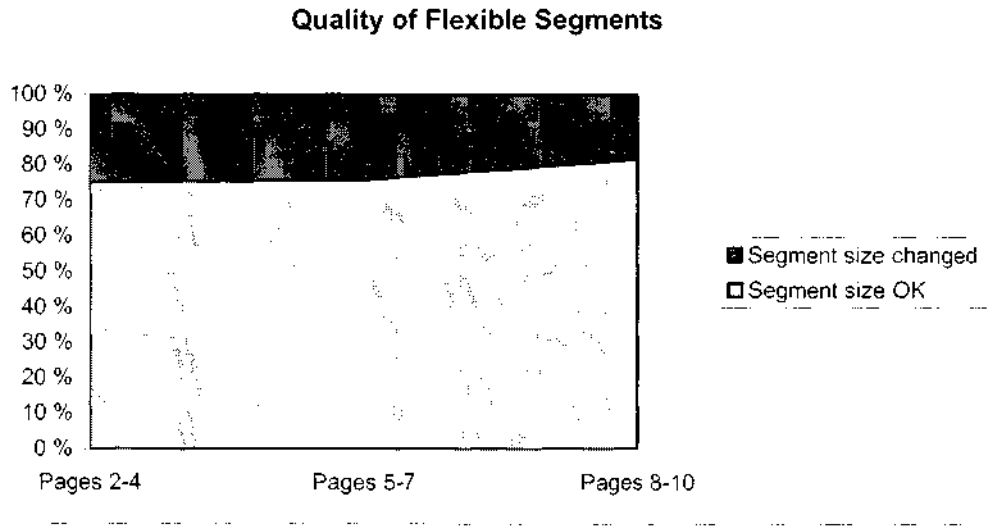


**Figure 1:** Roughly 45 of the suggested flexibly-sized segments were accepted by the translator. A slight increase in segmentation quality can be observed, i.e. the system learns to segment better during use.

## 4.3. The Three Translation Strategies Put to a Test

### 4.3.1. Distribution of Translation Strategies

Secondly, we wanted to study the distribution of translation strategies, i.e. in what proportion would Translation Intelligence through the Master Translator Pro program employ translation memory, translation generation and translation heuristics when suggesting translations for the flexibly-sized segments.

The calculations were made on the basis of the *final* flexible segmentation, i.e. after the translator had had a chance to resize the segments manually. This was done in order to discover the *real* distribution of translation strategies that a user is faced with when producing high-quality translations of previously unseen texts. It is, however, worth noting that if the length of a segment is adjusted manually, MTP will automatically shift to the translation heuristics mode to produce a translation suggestion of the resized segment. This means that the amount of machine-translated segments is correlated to the amount of resized segments. In other words, as approximately 20 % of the segments in the source text needed resizing (see figure 1 above), 20 % of the segments, at a very minimum, would also be translated using translation heuristics.

The assumption was that the last translation strategy, translation heuristics, would be quite dominant at the beginning of the translation, but the translation memory and translation generation strategies would outweigh it in the long run, as the system gets accustomed to translating recipes.

On an average, 5.22 % of the translation suggestions of the test document segments (or 6.06 % for those on the last five pages) were made using the primary translation strategy, translation recycling. In other words, 1/20 of the segments were such that a translation was found in the Knowledge Base that had no prior cookery knowledge. The result clearly indicates that there is much more domain-independent repetition on the sub-sentence level than on the sentence level. Some 6 % may not seem like much, but compared to the percentage of full matches you would get with a traditional, sentence-based translation memory system when translating previously unseen text, the percentage is remarkably high. Traditional translation memories, with their static segmentation, would have rendered next to nothing straight from the translation memory, because the recycled segments, which could be used on the surface level, were considerably shorter than a whole sentence, only 1-4 words.

By far the most commonly used strategy was the secondary strategy, translation generation, which is the unique translation method employed by Translation Intelligence. On an average, 63.87 % of the translation suggestions were made using this method.

The last strategy, translation heuristics, was used to suggest translations of 30.91 % of the segments. Keeping in mind that all resized segments (about 20 % of all segments) fall under this category, only roughly 10 % of the segments that did not need resizing were translated using heuristics.

The statistics on the distribution of translation strategies for the first five pages as compared to the last five pages of the test data show some tendencies (figure 2). It

seems that the amount of recycled translations tends to grow slowly (from 4.30 % to 6.06 %), whereas the amount of machine-translated segments tends to decrease (from 31.79 % to 30.12 %). However, to actually prove this tendency as statistically significant, ten pages are not enough. Translation Intelligence needs more data to tailor itself properly for translating documents of a new text type.

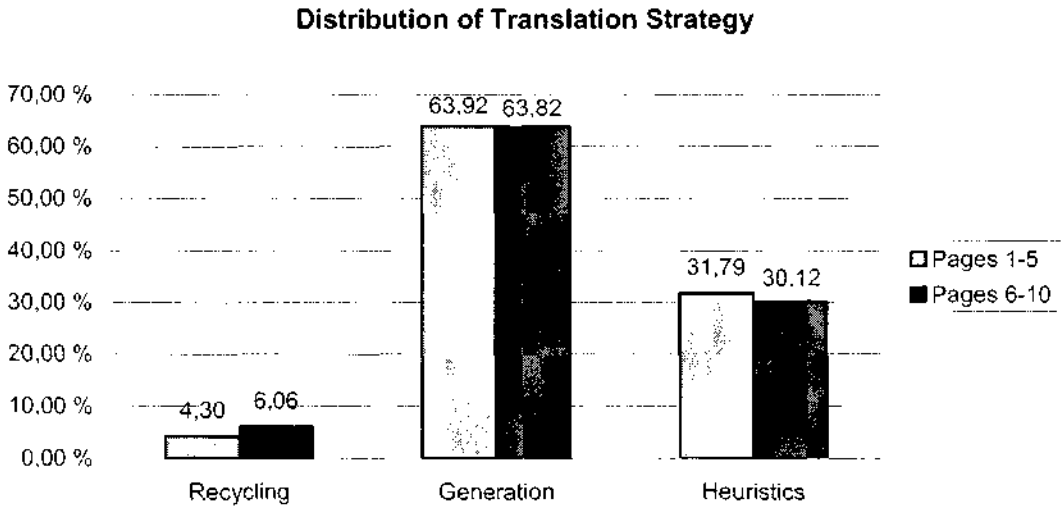


**Figure** 2: Translation generation is the most commonly used translation strategy when translating texts from a previously unseen domain. Translation recycling is the least used strategy, even though the method tends to be used more and more as Translation Intelligence gets accustomed to the domain at hand. Similarly, the amount of machine-translated segments (translation heuristics) tends to decrease over time.

### 4.3.2 Quality of Translation Strategies

The proportions of the different translation strategies show only one side of the coin. What we were most interested in finding out was, how *accurate* would the different strategies be, i.e. what is the quality of the suggested translations (figure 3). In other words, to what extent could the translator accept the recycled, generated or machine-translated translation suggestions without correcting them? This is possibly the most important criterion when considering usability and efficiency—if a strategy mostly makes worthless translation suggestions, it can be argued that the strategy should not be applied at all.

Of the recycled translation suggestions (the least used strategy with unfamiliar text), as many as 96.88 % were acceptable without modification. Many of these were simply one-word translations (such as the coordinating conjunction “ja” that existed in the Knowledge Base with the translation “and”). Nevertheless this observation shows that translations of segments on the sub-sentence level can at least to some extent be used in a largely text-type-independent fashion.

Of the generated translation suggestions (the most commonly used strategy with unfamiliar text) as many as 72.80 % were such that the translator could accept as suitable translations for the segments at hand. This verifies the claim that largely text-type-independent semantic and grammatical translation patterns do exist and can

viably be used to generate translations of segments from a previously unseen domain, when utilizing previous human translation knowledge learned from another domain.

Of the machine-translated suggestions (used in 1/3 of the cases when faced with unfamiliar text), only 17.68 % were such that the translator could accept without modifications. The rest of them required modification. However, these raw translations were used as a basis for providing an appropriate translation. In other words, even though the machine-translated suggestion was syntactically incorrect in many cases, it was a great help in getting the terminology right.

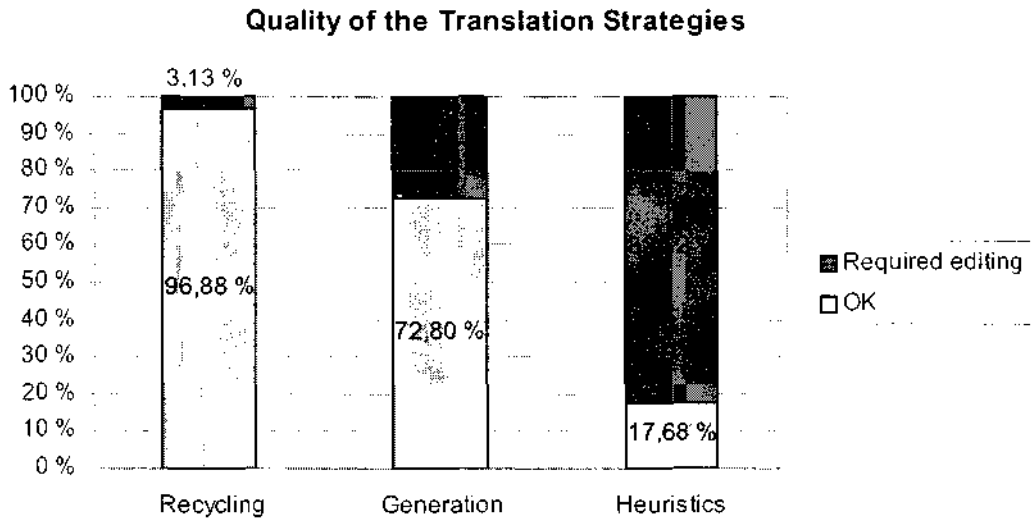


**Figure 3:** Recycled translations tend to be usable without modification (leftmost column), whereas machine-translated suggestions need manual editing in more than four cases out of five (rightmost column). Over two thirds of the generated translation suggestions—the unique feature used in Translation Intelligence—could be accepted as such (middle column).

To conclude, if we disregard the translation strategy applied, 57.01 % of the translation suggestions were such that the translator could accept them without modifications. In other words, using Translation Intelligence through the Master Translator Pro program that had not received any cookery-related tailoring, we were able to translate over half of the ten pages of recipes correctly!

## 4.4. What Did the System Learn?

The key to success of Translation Intelligence is the technology's ability to learn from the user's way to translate and thus adapt to the type of texts that the user translates. The system learns constantly during interactive translation, and on several levels. It learns terminology, conventional translation units and translation patterns. It learns to segment the text better, and it learns spelling rules for both source and target languages.

During the translation of the ten pages of recipes. 562 new translation units of various sizes were added to the Knowledge Base and marked with Cookery as their domain. Most of these units were also analysed by the program and stored as translation patterns to be used in translation generation. Moreover, the system learned 157 new

cookery terms in Finnish and English, immediately ready for reuse. The Finnish spell checker increased by 64 new correct spellings and the English grew by 8 .

### 4.5. Drawbacks of the Test

Due to the limited amount of test data (less than 500 sentences), the recipe test cannot be used to reliably predict how *fast* Translation Intelligence will adapt to a new text domain such as cookery. Some tendencies were reported, e.g. that the amount of recycled translations is inclined to grow over time, whereas the amount of machine-translated segments seems to decrease. But in order to get statistically valid data on how fast Translation Intelligence adapts to a new text type, more test data is needed, probably at least 5,000 sentences.

Rather, the recipe test has given us valuable information on how well Translation Intelligence will manage when faced with a new domain, given that no tailoring is done in advance. This is something that the new users of Master Translator Pro and also of other traditional translation memory systems are faced with—the initial phase, when the system is brought up to speed. The recipe test gives a fairly good overview of the capabilities inherent in Translation Intelligence and in the end-user tool Master Translator Pro when taken into use for the first time. As 57 % of the ten test pages could be translated automatically, it seems reasonable to suggest that MTP, using Translation Intelligence, is faster to take into use than a traditional TM system. Several customers using MTP in everyday translations confirm this.

## *5 Customer* Case: *Innovative Business Oy*

### 5.1. About the Customer

Innovative Business Oy. a Helsinki-based company that provides consulting services and software needed an effective solution for the translation of psychological reports from English into Finnish. The company regularly produces fairly large amounts of psychological reports, and schedules are often tight. Having made comparisons between the translation programs on the market, the company started cooperation with Master's Innovations Ltd.

Translations were first made to clear the client’s translation backlog and to add special terminology into the translation program. A professional translator using the Master Translator Pro program did the translation work and tailoring, which took about one man-month. After that, the client’s personnel started using the translation program on their own. According to the client's estimate, they soon benefited as much as 80% in efficiency gains.

### 5.2. An Interview with the Customer

*What was it like to start using Master Translator Pro?*
“I think that the program is really easy to learn, and the user interface is clear,” says Mr Jukka Väisänen from Innovative Business.

*According to your estimate, how fast will the program pay itself back in your use?*
'”It paid itself back during this first project, if we don’t put a full price on the work that we did ourselves. I have a feeling that it has been amortized during this first client

project, and from now on, using the program will clearly save money on every profile required in Finnish (and also one or two days per case)."

*To what extent have you benefited from the program in your translation work?*
"At first it took us relatively much time and resources to work with the program, but after 20 or 30 reports it started getting notably faster. We now use the program to translate an average of ten reports (one or two A4 pages each) in an hour, file saving and other procedures included. A translation agency would spend from half an hour to an hour per report. So, roughly, we accomplish in an hour what a translation agency does in a day."

"Master Translator Pro is for us the only sensible and cost-efficient solution to produce high-quality translations within tight schedules," declares Mr Väisänen.

## *6. References*


Lommel, Arle & Ray, Rebecca ed. (2004). *LISA 2004 Translation Memory Survey. Translation Memory and Translation Memory Standards.*

Wheatley, Alan (2003). "eCoLoRe (eContent Localization Resources for Translator Training). A Major Breakthrough for Translator Training". *Globalization Insider XII/2.4.*

Beninatto, Renato & DePalma, Donald A. (2005). *Ranking of Top 20 Translation Companies.* Common Sense Advisory (http://www.commonsenseadvisory.com/pdf/05070 l_QT_top_20.pdf)

- (2002). *The Full-Text Multilingual Corpus: Breaking the Translation Memory Bottleneck.* MultiCorpora R&D Inc.